Integrating Pre-Trained Language Models into Novel Query Expansion Pipelines

Iain Mackie

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

March 2025

Abstract

In our increasingly digital society, proficiency in finding valuable and useful information is crucial within everyday personal and professional life. Information Retrieval (IR) is the academic field that focuses on discovering useful information (documents) that fulfil a user's information need (query). In particular, search systems process a user query and return a ranked list of documents determined by the query-specific relevance calculated by a retrieval model. Retrieval models have traditionally been based on query-document term overlap; however, dense embedding models are becoming increasingly prevalent with the emergence of Pre-Trained Language Models (PLMs).

Lexical mismatch is a classic problem within information retrieval, whereby a user query fails to capture their complete information need, leading to retrieval models failing to find relevant documents. A common approach for this issue is query expansion, involving the augmentation of the query with supplementary information to enhance the retrieval of relevant documents. This process is usually done automatically through pseudo-relevance feedback (PRF), where a set of documents from a first-pass retrieval algorithm are assumed relevant, and are used to expand the query with useful context. This approach has proven beneficial for both sparse and dense retrieval models. In this thesis, I hypothesise that integrating PLMs into multi-stage query expansion pipelines can improve performance over current sparse and dense expansion methods. Leveraging the capabilities of PLMs to generate and rank relevant content to build better expansion models, which should particularly help queries that require reasoning or contextualisation.

This thesis examines existing retrieval and expansion models, identifying their shortcomings to focus the contributions. This includes constructing a formal definition of complex queries and constructing new datasets, such as CODEC, designed to evaluate the effectiveness of our proposed retrieval models on this query type. I first show that by simply using PLMs to re-rank our first-pass candidate set of documents before query expansion improves sparse and dense retrieval by 5–8%. This motivates the development of a new fine-grained expansion model, Latent Entity Expansion (LEE), that achieves a further 2-8% gain in NDCG by explicitly modelling knowledge using terms and entities. Furthermore, I introduce a novel expansion pipeline, termed "adaptive expansion", that iterates between retrieving new batches of documents and updating the expansion model through PLM re-ranking. Adaptive expansion leads to state-of-the-art effectiveness gains without requiring any additional re-ranking computation.

This thesis's second central research thread explores not using pseudo-relevance feedback at all; instead, leveraging the generative capabilities of PLMs to build our query expansion models directly. I introduce Generative Relevance Feedback (GRF), which shows that sparse expansion using PLM-generated content improves MAP between 5-19% over traditional PRF. I also show that GRF is highly effective when combined with dense and learned sparse retrieval. Furthermore, I increase the retrieval effectiveness of these pipelines by incorporating Generative Relevance Modelling (GRM) to mitigate hallucination by scaling the weight of generated documents in our expansion model. We propose Relevance-Aware Sample Estimation (RASE) to ground the generated documents to the target corpus and use PLMs to estimate relevance.

Overall, this body of work demonstrates the potential of query expansion when combined with novel pipelines that leverage the capabilities of PLMs. This paradigm shift provides a foundation for future advancements, promising more efficient and effective search systems.

Contents

List of Tables

List of Figures

Acknowledgements

This PhD journey has been an incredible and sometimes challenging experience, and I am deeply grateful to the many individuals who have supported, guided, and inspired me along the way.

First and foremost, I would like to extend my heartfelt thank you to my supervisors, Dr. Jeffrey Dalton and Dr. Richard McCreadie. Your invaluable guidance, support, and belief in me have been instrumental in shaping this work. I am sincerely grateful for your mentorship and the opportunities you have provided throughout my PhD.

I feel very fortunate to have been part of the GRILL Lab, where collaboration and camaraderie made each day both enjoyable and rewarding. Carlos Gemmell, Federico Rossetto, Paul Owoicho, Alessandro Speggiorin, Sophie Fischer, and Dr. Shubham Chatterjee, thank you for the countless discussions, idea exchanges, and moments of laughter. Personal highlights include research trips to California, conferences from Madrid to Taiwan, and post-work drinks on Ashton Lane.

My research journey has also been enriched by collaborations with multiple experts in the field of information retrieval. In addition to my supervisors and members of the GRILL Lab, I would like to thank Dr. Sean MacAvaney, Dr. Andrew Yates, Thong Nguyen, Dr. Ivan Sekulic, Prof. Fabio Crestani, and Elena Soare for their invaluable contributions and insights.

I would also like to acknowledge the Alexa Prize team members, whose creativity and dedication provided a unique and unforgettable experience during my PhD. This project served as a springboard for my research and my next professional chapter, founding Malted AI.

On a personal note, I owe a profound debt of gratitude to my family. To my parents, Yvonne and Kenny Mackie, and my sister, Fiona Mackie, thank you for your unwavering support, encouragement, and belief in me from the beginning. Your help and motivation throughout this journey have been invaluable.

Lastly, to my beautiful wife, Louise Mackie — thank you for encouraging me to pursue this PhD and for believing in me every step of the way. Your love, patience, and unwavering support have been my essential through the challenges and triumphs. I could not have done this without you.

Declaration

With the exception of chapters 1, 2 and 3, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

1.1 Background

This thesis advances state-of-the-art query expansion by integrating pre-trained language models in expressive retrieval pipelines. I concentrate on the task of document retrieval, with a particular focus on complex queries where current search systems struggle. Before outlining my objectives and contributions, I provide an overview of the task, terminology, and foundational elements that underpin this work.

1.1.1 Document Retrieval

Document retrieval is a core task within the field of Information Retrieval, where a retrieval algorithm needs to return a relevance-ranked list of documents from a collection [15, 134]. The ubiquity of search systems is evident across many facets of our lives, encompassing web search engines such as Google, Bing, and Baidu, domain-specific enterprise search like Bloomberg and LexisNexis, and even the search for personal data across our devices. Consequently, the development of effective and robust retrieval models is crucial for facilitating valuable information access.

More generally, a search system takes a user "query" as input, which is a representation of the user's information needs, and returns a ranked list of documents from a target document collection. This collection can contain millions or even billions of unique documents; thus, the document retrieval model needs efficiency to be executable over the entire collection. For example, conventional document retrieval models, such as TF-IDF [187] or BM25 [172], are term-based models that leverage term distributions within the query, documents, and overall corpus to approximate query-document relevance.

1.1.2 Pre-Trained Language Models

The emergence of pre-trained language models (PLMs), built upon the Transformer architecture [195], has revolutionised the field of natural language processing (NLP). Early models such as GPT [167] and BERT [47] demonstrate that training these large deep learning models on extensive datasets yields significant improvements in understanding and generating coherent text. Over subsequent years, both the models and datasets have scaled even further, with notable successors like GPT4 [2], Claude 2, LLama 2 [192] and Mistral [86] gaining widespread adoption across public and professional domains owing to their advanced capabilities.

Similar to other fields within natural language processing, pre-trained language models have transformed state-of-the-art performance across information retrieval tasks [217]. For example, they excel at document re-ranking [104, 149, 150], even being integrated into commercial web search engines to enhance ranking and question answering (QA),[1] I have also seen the advent of vector-based search systems [92, 112, 213], where both queries and documents are encoded in a low-dimensional (dense) space, which has been shown to excel on factoid web queries [147]. This is an initial phase, with the newfound capabilities of PLMs poised to continue to reshape the landscape of information retrieval for years to come.

1.1.3 Query Expansion

Lexical mismatch [18] is a problem for search systems where a user's query contains terms that substantially differ from those present in relevant documents, posing a challenge for retrieval models to operate effectively. To address lexical mismatch, query expansion techniques [173] augment the query by expanding with terms closer to the underlying information need. An automatic approach commonly employed is *pseudo-relevance feedback* (PRF), where the top-k documents from an initial retrieval set are assumed relevant, and the query is updated. Noteworthy classical methods [1, 136, 173, 221] are grounded in term-based probabilities derived from the query and assumed relevant documents. The formulation of query expansion has evolved to encompass additional features [136] and structured information extracted from knowledge bases [42].

Figure 1.1 shows how query expansion can enrich the original query about the social changes brought about by the black death with useful concepts, i.e., [pandemic], [wage], [revolt], and [population]. These terms reduce the lexical mismatch between the query and relevant documents, and the second-pass query improves retrieval effectiveness.

Moreover, the advent of using PLMs to encode text into dense vectors has led to the application of vector-based PRF. This involves enhancing query vectors by incorporating feedback document vectors [106, 145, 201, 219]. Figure 1.2 shows how the query vector is updated based on feedback vectors to improve retrieval effectiveness.

[1] https://blog.google/products/search/search-language-understanding-bert/

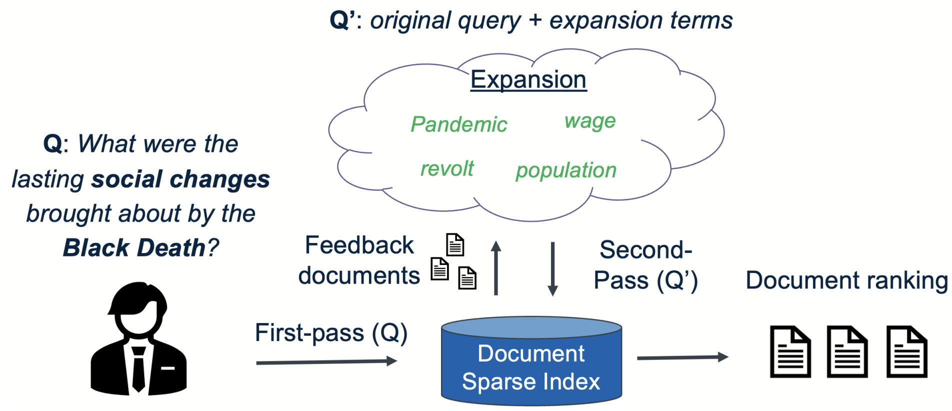


Figure 1.1: An example of sparse query expansion

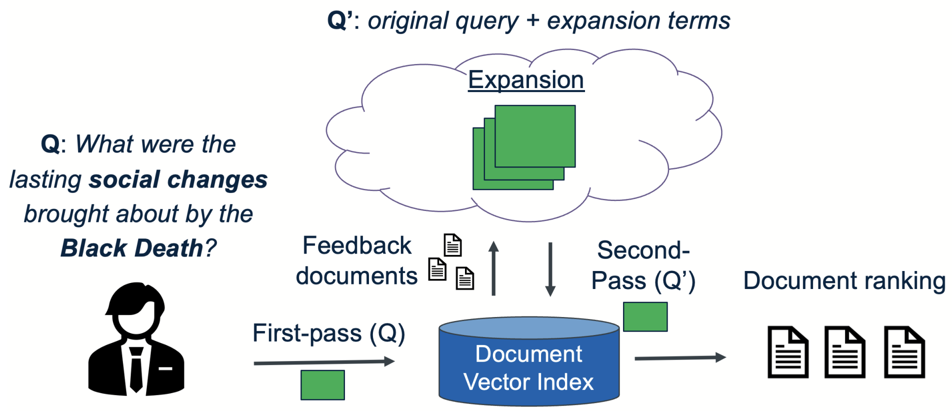
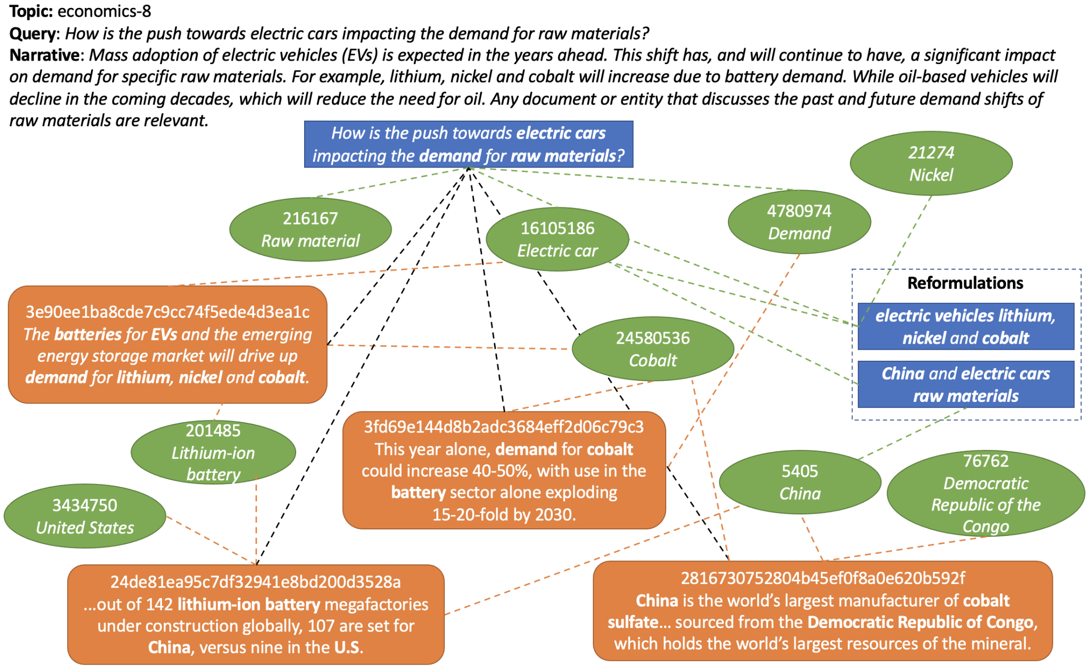

Q': original query + expansion terms
Expansion
Q: What were the lasting social changes brought about by the Black Death?
Feedback documents
Second-Pass (Q')
Document ranking
First-pass (Q)
Document Vector Index

Figure 1.2: Dense Query Expansion

1.1.4 Query Types

With vast quantities of annotated query-level data [147] and advances in deep learning models [92, 104, 112, 149, 150], certain categories of queries have extremely high retrieval effectiveness [33, 34]. However, this thesis intentionally shifts its focus away from conventional "easy" factoid-focused queries. Instead, I focus on more challenging queries, where existing state-of-the-art sparse and dense retrieval methods still have significant headroom for improvement [20, 130, 131].

Figure 1.3 depicts an example of a "complex" queries. Complex queries are defined as being multifaceted, concerning multiple entities and concepts, and requiring significant amount of knowledge and comprehension to contextualise the topic. For instance, "How is the push towards electric cars impacting demand for raw materials?". This information need requires knowledge from various documents, e.g., raw materials that constitute electric cars, supply chain dynamics, and sales growth. While simultaneously understanding critical entities and relationships (e.g., [Electric car], [Raw Materials], [Cobalt], [Lithium-ion battery], and [Demand]) becomes vital to grasping the intricacies of the topic.



Topic: economics-8
Query: *How is the push towards electric cars impacting the demand for raw materials?*
Narrative: *Mass adoption of electric vehicles (EVs) is expected in the years ahead. This shift has, and will continue to have, a significant impact on demand for specific raw materials. For example, lithium, nickel and cobalt will increase due to battery demand. While oil-based vehicles will decline in the coming decades, which will reduce the need for oil. Any document or entity that discusses the past and future demand shifts of raw materials are relevant.*

How is the push towards electric cars impacting the demand for raw materials?
21274 Nickel
216167 Raw material
16105186 Electric car
4780974 Demand
Reformulations
electric vehicles lithium, nickel and cobalt
China and electric cars raw materials
3e90ee1ba8cde7c9cc74f5ede4d3ea1c The batteries for EVs and the emerging energy storage market will drive up demand for lithium, nickel and cobalt.
24580536 Cobalt
201485 Lithium-ion battery
3fd69e144d8b2adc3684eff2d06c79c3 This year alone, demand for cobalt could increase 40-50%, with use in the battery sector alone exploding 15-20-fold by 2030.
5405 China
76762 Democratic Republic of the Congo
3434750 United States
24de81ea95c7df32941e8bd200d3528a ...out of 142 lithium-ion battery megafactories under construction globally, 107 are set for China, versus nine in the U.S.
2816730752804b45ef0f8a0e620b592f China is the world's largest manufacturer of cobalt sulfate... sourced from the Democratic Republic of Congo, which holds the world's largest resources of the mineral.

Figure 1.3: An example of a complex query: *How is the push towards electric cars impacting the demand for raw materials?* This topic requires contextualisation across multiple documents and entities.

1.2 Thesis Statement

I hypothesise that integrating pre-trained language models into multi-stage query expansion pipelines will improve document retrieval effectiveness over traditional sparse and dense models. My hypothesis is rooted in the notion that PLMs offer the capability to both rank and generate relevant content, serving as the foundational elements to build novel query expansion models. These advancements will prove especially advantageous for complex queries requiring additional reasoning or contextualisation over multiple documents and entities.

1.3 Motivation

Pre-trained language models have shown substantial capabilities in language comprehension and generation of fluent text [2, 47, 86, 167, 192]. These models' strong effectiveness extends to various information retrieval tasks [217], including document re-ranking models [149, 150] and dense retrieval systems [92, 213]. In particular, current neural ranking systems are performing very well on specific classes of topics [33, 34]. For example, information needs that are factoid, require basic lookups, or demand only a short answer to satisfy the request [130].

In particular, I focus on the task of document retrieval, with an emphasis on the development of novel query expansion models that harness PLMs to enhance system effectiveness. The history of query expansion is deeply rooted in term models [1, 136, 137, 173, 221] and integration of entity-based information [42, 135, 208, 214]. More recently, there has been a surge in the adoption of dense vector-based PRF models [106, 145, 201, 219], and even the exploration of feedback within the learned sparse paradigm [96].

However, all these expansion models struggle when first-pass retrieval fails. For example, Figure 1.4 shows a real topic from CODEC where poor feedback causes the second-pass query to drift off-topic and become worse than the original query. The query is expanded with terms or vectors that represent non-relevant concepts with respect to social changes brought about by the Black Death, i.e., [police violence] and [modern social unrest]. These types of expansion failures are much more likely for complex topics needs that require critical information spanning multiple documents and entities [131].

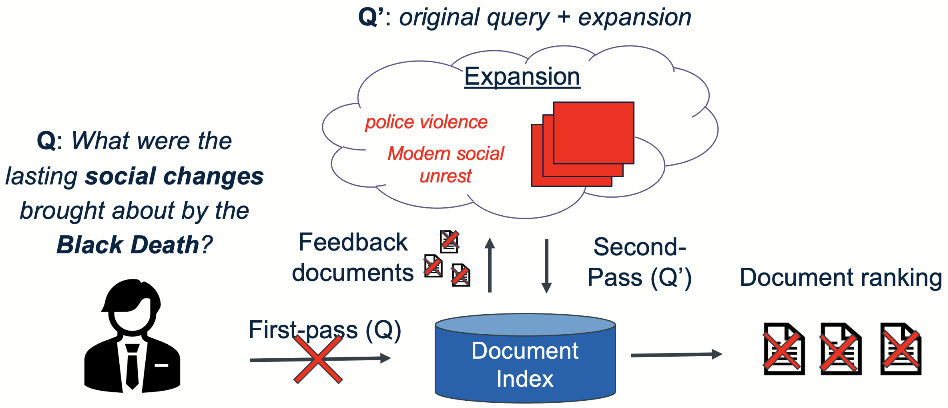


Figure 1.4: Query expansion with poor feedback

In this thesis, I believe that re-designing query expansion pipelines and models, anchored in the capabilities of PLMs (Figure 1.5), holds the potential for substantial effectiveness improvements over traditional sparse and dense models. The underlying hypothesis stems from the belief that PLMs can both rank and generate relevant content, laying the groundwork for innovative multi-stage query expansion models. I believe these next-generation query expansion models will be especially advantageous for complex queries requiring extensive reasoning or contextualisation.

1.4 Contributions

This thesis makes substantial contributions to the field of information retrieval with a specific emphasis on query expansion. In particular, my work showcases significant performance improvements by developing new datasets, conducting comprehensive analyses on complex queries and model behaviour, and introducing novel retrieval models. In the following section, I outline my major contributions contained within each chapter.

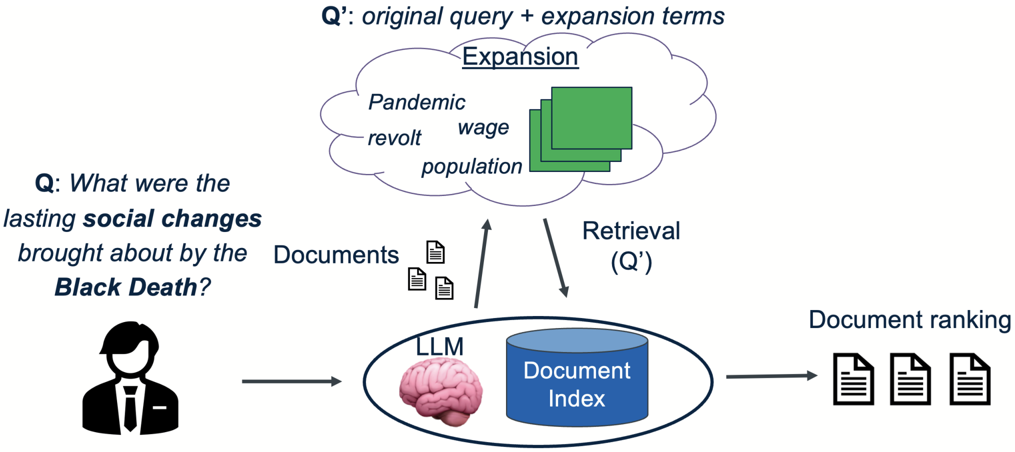

Q': original query + expansion terms
Expansion
Pandemic
revolt
wage
population
Q: What were the lasting social changes brought about by the Black Death?
Documents
Retrieval (Q')
LLM
Document Index
Document ranking

Figure 1.5: New query expansion pipelines using PLMs

1.4.1 Chapter 4: Datasets for Complex Queries

Complex Query Criteria

I conduct analysis of challenging queries on prior data, and define a process to categorise complex queries manually. This complex criteria is based on comprehensive query and model analysis of TREC Deep Learning [130]. Complex queries are multifaceted, concern multiple entities, and require deep comprehension and knowledge. I develop an annotation process to identify the query types that can be applied across datasets.

Analysis of Prior Datasets

I show that standard datasets, such as TREC Deep Learning, contain primarily non-complex queries. Nonetheless, I find that a large proportion of the Robust04 queries are complex based on my developed manual identification methods. I also conduct detailed system analysis depicting where state-of-the-art retrieval and re-ranking systems struggle and opportunities for improvement. This analysis supports that TREC Robust04 is a reasonable dataset for evaluating complex information needs, while motivating the development of new datasets.

Construction of the CODEC Dataset

My analysis of queries that challenge existing ranking models motivates the development of a new dataset allowing fundamental research on complex topics [131]. This focuses on social science domains (history, finance, politics) and provides complex essay-style queries, golden "narratives" capturing the information need, query facets, and dense annotations of relevant document and entities. The novel resource provides complex queries to support the development and evaluation of new retrieval models. I also conduct analysis on state-of-the-art retrieval and re-ranking systems to identify opportunities for improvement in expansion models. In particular, I find that even basic query expansion methods using entities and query facets improve ranking effectiveness. These explorations act as motivation for future work.

1.4.2 Chapter 5: Query Expansion with PLM Ranking

PLM Re-Ranking for Expansion Feedback

A known limitation of pseudo-relevance feedback approaches is that their effectiveness hinges on high precision within the document feedback set. To address this, I rethink query expansion pipelines by harnessing the ranking capabilities of pre-trained language models [111] to construct a more focused expansion model [128]. This paradigm shift increases performance across sparse and dense retrieval by 5-8%. Additionally, I introduce the concept of “adaptive expansion”, inspired by the iterative refinement of queries by humans. This involves employing a neural re-ranker to update the expansion model as more documents are scored iteratively, thus avoiding re-ranking additional documents. This work makes significant effectiveness gains on complex queries.

Fine-Grained Query Expansion with Words and Entities

I investigate novel query expansion models utilising feedback from re-ranking from PLMs. Notably, my investigation reveals that employing a re-ranking PLM to construct relevance models based on passage spans proves more effective than utilising entire documents. In this context, I introduce an enhanced expansion model called Latent Entity Expansion (LEE), which applies fine-grained word and entity-based relevance modelling incorporating localised features [128], and further increasing NDCG by 2-8%. Additionally, LEE improves recall of the hardest 5% of queries by 0.6 and shows significant gains on entity-centric queries due to a strong entity representation.

1.4.3 Chapter 6: Query Expansion with PLM Generation

PLM Generative Content for Query Expansion

Rather than relying on pseudo-relevance feedback for query expansion based on documents from the target corpus, I propose rethinking query expansion pipelines to use pre-trained language to generate synthetic documents. My contribution lies in the first published work [127] to show that PLMs can effectively generate text for probabilistic query expansion, which I call Generative Relevance Feedback (GRF). I show that long-form text generation, tailored to the style of the target corpus, emerges as the most effective standalone generation technique for query expansion. However, aggregating text across all generation techniques yields a more effective query model than any individual generation technique in isolation. This work shows that GRF improves MAP between 5-19% over RM3, a comparable sparse PRF model. Furthermore, I show that GRF generalises to dense and learned sparse retrieval, improving over comparable PRF techniques by around 10% on both precision and recall-oriented measures. This work shows that improved contextualisation using PLMs improves effectiveness on complex queries.

Reducing Hallucinations with Generative Relevance Modelling

My analysis of GRF reveals that hallucinations, which I categorise as PLMs generating non-relevant text content, result in specific queries going off-topic. To address this issue, I propose several solutions, including a straightforward PRF fusion method and a more intricate pipeline approach. In particular, I propose Relevance-Aware Sample Estimation (RASE) to enhance the term weighting within a Generative Relevance Model (GRM) by estimating the relevance of each generated document. In particular, identifying real documents similar to each generated document from the target corpus and employ a neural re-ranker to estimate their relevance. This pipeline approach improves MAP by 6-9% and R@1k by 2-4%, surpassing previous GRF methods, enabling the incorporation of diverse subtopics required to contextualise complex queries.

1.5 Originals Of Material

In this section, I catalogue the academic papers that were produced from the research as part of this dissertation. These papers were published at prominent peer-reviewed Information Retrieval conferences.

1. **Iain Mackie**, Jeffrey Dalton, and Andrew Yates. 2021. ***How Deep is your Learning: the DL-HARD Annotated Deep Learning Dataset***. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21).* Association for Computing Machinery, New York, NY, USA.

   The full resource paper shows that current ranking models are effective only on certain query categories (i.e., easier factoid queries), and there is significant headroom for improvement on other categories (i.e., queries with multiple facets or requires a long-form answer). These learnings motivate the development of CODEC. This work is detailed in Chapter 4.

2. **Iain Mackie**, Paul Owoicho, Carlos Gemmell, Sophie Fischer, Sean MacAvaney, Jeffrey Dalton. 2022. ***CODEC: Complex Document and Entity Collection***. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22).* Association for Computing Machinery, New York, NY, USA.

   This full resource paper creates a dataset that allows fundamental research on complex topics, showing that query expansion and entity-centric methods offer opportunities to improve ranking effectiveness. This work is detailed in Chapter 4.

3. **Iain Mackie**, Shubham Chatterjee, Sean MacAvaney, and Jeffrey Dalton. 2024. ***Adaptive Latent Entity Expansion for Document Retrieval***. In *ECIR Workshop on Knowledge-Enhanced Information Retrieval (KEIR).*

   This full workshop paper proposes "adaptive expansion" that leverages PLM-based re-rankers to create fine-grained relevance models. Additionally, I introduce a new expansion model, Latent Entity Expansion (LEE), that uses both terms and entities, and incorporate localised features. This work is detailed in Chapter 5.

4. **Iain Mackie**, Shubham Chatterjee, and Jeffrey Dalton. 2023. ***Generative Relevance Feedback with Large Language Models***. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23).* Association for Computing Machinery, New York, NY, USA.

   This short paper is the first published work to show that PLMs can effectively generate text content for probabilistic query expansion models. I show these achieve substantial gains over traditional PRF approaches. This work is detailed in Chapter 6

5. **Iain Mackie**, Shubham Chatterjee, and Jeffrey Dalton. 2023. ***Generative and Pseudo-Relevant Feedback for Sparse, Dense and Learned Sparse Retrieval***. In *CIKM Workshop on Large Language Models' Interpretation and Trustworthiness (PLMIT).*

   This full workshop paper demonstrates the PLM-generated content effectively contextualises queries across dense and learned sparse retrieval, showing the generalisability of GRF. Furthermore, I show that generative content can cause "hallucinations" and propose fusion methods to improve recall further. This work is detailed in Chapter 6.

6. **Iain Mackie**, Ivan Sekulic, Shubham Chatterjee, Jeffrey Dalton, and Fabio Crestani. 2023. ***Relevance-Aware Sample Estimation for Generation Relevance Modeling***. In *Arxiv*.

   This preprint shows that subtopic-based generated documents helps to contexulaise complex multifaceted topics. I effectively weighted documents within an expansion model based on the relevance of similar documents from the target collection. This work is detailed in Chapter 6.

1.5.1 Other Research Outcomes

There are also a number of research papers that were outputs of my PhD but are directly not included in this thesis due to being off-topic, shared task participation, or lack of sole contribution. These include:

7. Shubham Chatterjee, **Iain Mackie**, and Jeff Dalton. 2023. ***DREQ: Document Re-ranking***

Using Entity-Based Query Understanding. In *European Conference on Information Retrieval.*

8. Thong Nguyen, Shubham Chatterjee, Sean MacAvaney, **Iain Mackie**, Jeff Dalton, and Andrew Yates. 2024. ***Improved Learned Sparse Retrieval with Entity Vocabulary.*** In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 12 –16, Miami, Florida

9. **Iain Mackie** and Jeffrey Dalton. 2022. ***Query-Specific Knowledge Graphs for Complex Finance Topics***. In *AKBC Workshop on Knowledge Graphs for Finance and Economics.*

10. Carlos Gemmell, Sophie Fischer, **Iain Mackie**, Paul Owoicho, Federico Rossetto, and Jeff Dalton. 2022. ***GRILLBot: A flexible conversational agent for solving complex real-world tasks***. In *1st Proceedings of the Alexa Prize Taskbot.*

11. Sophie Fischer, Niklas Tecklenburg, Philip Zubel, Eva Kupcova, Ekaterina Terzieva, Daniel Armstrong, Carlos Gemmell, **Iain Mackie**, Federico Rossetto, and Jeff Dalton. 2023. ***GRILLBot-v2: Generative Models for Multi-Modal Task-Oriented Assistance.*** In *2nd Proceedings of the Alexa Prize Taskbot Challenge (2023).*

12. Sophie Fischer, Carlos Gemmell, Niklas Tecklenburg, **Iain Mackie**, Federico Rossetto, and Jeffrey Dalton. 2024. ***GRILLBot in Practice: Lessons and Tradeoffs Deploying Large Language Models for Adaptable Conversational Task Assistants.*** In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4951-4961.

13. Sophie Fischer, Carlos Gemmell, **Iain Mackie**, Jeff Dalton. 2022. ***VILT: Video instructions linking for complex tasks.*** In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval.*

14. Elena Soare, **Iain Mackie**, and Jeffrey Dalton. 2022. ***DocuT5: Seq2seq SQL Generation with Table Documentation.*** In *Arxiv*.

1.6 Thesis Outline

The remainder of this dissertation is organised as follows. Chapter 2 surveys the background literature and current state-of-the-art methods in topics relevant to this research. In Chapter 3, I outline the methodology and experimental setups of this work, including defining complex queries based on system analysis. In Chapter 4, I develop new datasets focused on complex topics and produce insights that form the basis of my novel methods in Chapters 5 and 6. Chapter 5 focuses on developing new query expansion models based on PLM re-ranking feedback, incorporating both words and entities. In Chapter 6, I extend my investigation to leverage the generative capabilities of pre-trained language models directly for query expansion, introducing methods to mitigate harmful hallucinations. Lastly, Chapter 7 serves as the conclusion, summarising this dissertation's key findings, contributions, and outlining future work.

Chapter 2

Related Work

In this literature review, I cover the extensive body of research that my thesis builds upon and extends. I begin by providing an overview of pre-trained language models since the introduction of the Transformer architecture [195]. This includes exploration of the different categories of PLMs: encoder-only [29, 47, 117], decoder-only [22, 168, 192], and encoder-decoder models [101, 169]. I discuss these models' unique capabilities when combined with large-scale training data and computational resources. This review underscores the potential of applying PLMs to this thesis to rank and generate relevant content for query expansion.

I discuss the literature on query complexity. While structured queries such as Boolean [6, 157, 215], SQL [51, 87, 178], SPARQL [9] fall outside the direct scope of this thesis, there are synergies with my work on entity-based retrieval. I also examine and contrast factoid [197] and non-factoid queries [30]. Factoid queries are seen as generally easier for modern systems [20], particularly those leveraging pre-trained language models [20], as demonstrated on factoid benchmarks like TREC Deep Learning [33, 34] and SQuAD [170]. In contrast, non-factoid queries, are widely regarded as more challenging [20]. However, there is a notable scarcity of datasets for non-factoid queries [20], especially when comparing document retrieval and question-answering communities. Addressing this gap, I developed criteria to define complex queries and focus this thesis on these query types. I identify existing datasets, such as TREC Robust 2004 [198], and introduce new datasets, CODEC [131], to benchmark and advance retrieval models tailored to these complex information needs.

The primary focus of this thesis is on document retrieval, which I provide a history ranging from Boolean retrieval [36] to statistical models [172, 187] to probabilistic language modelling [156, 164]. I also discuss the emergence of leveraging pre-trained language retrieval for initial retrieval, including dense retrieval [91, 212], ColBERT [92], and learned sparse retrieval [55, 123]. Furthermore, I highlight how statistical and probabilistic retrieval systems remain robust baselines, participially in out-off domain settings [191] and for entity-centric queries [17], where they can outperform more complex PLM models. In this thesis, I benchmark against and extend multiple search paradigms to show that my proposed PLM pipelines are effective and robust.

I also provide a comprehensive literature review of document re-ranking, a technique that has gained popularity with the emergence of deep learning and, more recently, pre-trained language models. This includes encoder-only [104,126,149] and encoder-decoder models [58,150], trained to rank relevant documents to improve precision. Later in this thesis, I will leverage these models' ranking capabilities to build more accurate and fine-grained query expansion models. Additionally, I extend adaptive re-ranking [125], where the candidate pool is dynamically updated during re-ranking based on the document clustering. Building upon this framework, I introduce adaptive expansion, where the candidate pool is refined using a query expansion model based on PLM feedback, with the aim of improving retrieval effectiveness without increasing computational demands.

New query expansion models and pipelines are the primary contributions of this thesis. I place my work in the context of the literature, covering term-based expansion models [1, 136, 221], entity expansion models [42, 135, 180], and dense expansion [105, 145, 219]. In this thesis, I explore whether using PLMs to rank relevant content significantly improves retrieval effectiveness across these search paradigms. Furthermore, I introduce a new expansion model, LEE, that combines text and entity representations into a fine-grained probabilistic model based on PLM-ranked content. Additionally, I cover the current literature that leverages content from pre-trained language models to improve relatival effectiveness [21, 59, 121, 151]. Building on this, I introduce multiple novel expansion pipelines that leverage generated content, aiming to improve retrieval effectiveness.

2.1 Pre-Trained Language Models

Machine learning techniques have been prevalent in natural language processing to learn useful representations to understand and process human language. For instance, pre-trained text embedding methods such as Word2Vec [139] and GloVe [158] can encode basic semantic relationships and contextual information between words. Furthermore, deep convolutional (CNNs) and recurrent neural networks (RNNs) have been widely used for years but struggled in scaling to effectively learn contextualized word representations required for complex language tasks.

In 2017, the Transformer architecture [195] enabled the emergence of pre-trained language models (PLMs). These large models leverage self-attention to learn deep semantics between words through extensive training on web-scale text datasets based on language tasks such as next-sentence prediction and masked language modelling. PLMs' capabilities to understand and generate coherent text have led to significant advancements in natural language processing (NLP). Figure 2.1 shows the Transformer architecture, with the encoder on the left and the decoder on the right; see the original paper for a complete explanation.

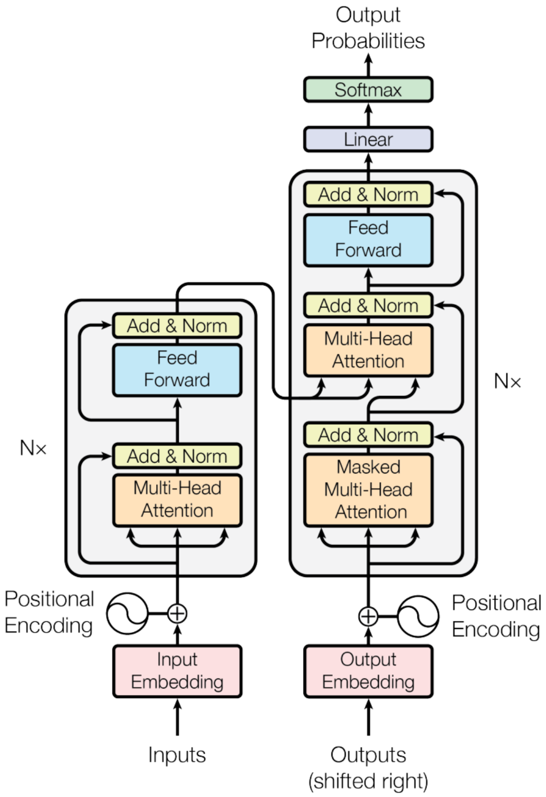


Figure 2.1: The Transformer architecture [195]

2.1.1 Encoder-Only PLMs

Encoder-only pre-trained language models contain only transformer encoder layers and create fixed-size embeddings based on inputted text. PLMs build context-aware representations for each token and more effectively capture semantic context than static word embeddings, such as Word2Vec [139] and Glove [158], where word representations do not vary based on context. PLMs can differ given the number of transformer layers, the token vocabulary, and the pre-training objectives. For example, the most common encoder-only PLMs is BERT [47], employing masked language modelling (MLM) and next sentence predication pre-training techniques to learn the deep language semantics present within the large training corpus. Other popular encode-only models include, RoBERTa [117], ELECTRA [29], and XLNet [216]. These models are useful for classification NLP tasks, like search relevance [91, 92, 104, 126, 150, 212], sentiment analysis [5, 70, 183, 186], and entity linking [206, 206], but not the generative tasks more community associated with the public perception of PLMs. I employ several encoder-only models for dense retrieval and document re-ranking baselines, while improving dense retrieval effectiveness via query expansion leveraging PLM's generation and ranking capabilities.

2.1.2 Decoder-Only PLMs

Decoder-only pre-trained language models comprise a decoder module, i.e., multiple multi-head attention and feed-forward transformer decoder layers solely. The most common family of PLMs is the Generative Pre-trained Transformer (GPT) models [167], which, unlike encoder-only models, generate coherent and contextually relevant sequences, demonstrating proficiency in text completion and generation. OpenAI shows through scaling data, model size, and computation, from GPT-1 [167] to GPT-2 [168] to GPT3 [22]. Decoder-only PLM's text generation capabilities continue to improve, ultimately leading to public, business, and fasciation of these models with the release of ChatGPT. There have also been open-source releases of decoder-only PLMs, such as Llama [192] from Meta and BLOOM [98] from the research community. In this thesis, I experiment with large decoder-only PLMs [22] to generate relevant text to improve query expansion models.

2.1.3 Encoder-Decoder PLMs

Encoder-decoder PLMs models sometimes called "Seq2Seq" models, use the encoder to process the input sequence to create a context vector, and the decoder utilises this context vector to generate an output sequence. For example, both T5 [169] and BART [101] are pre-training on diverse tasks and data to effectively handle various sequence-to-sequence tasks, such as machine translation [93, 189, 189] or code generation [3, 202, 203]. I primarily use encoder-decoder PLMs [150] for PLM ranking capabilities within my expansion pipelines.

In recent years, both the models and datasets have scaled even further, with notable successors like GPT4 [2], Claude 2, LLama 2 [192] and Mistral [86] gaining widespread adoption across public and professional domains owing to their advanced capabilities. For this reason, I focus on incorporating PLMs into novel retrieval pipelines leveraging their next-generation capabilities to rank and generate relevant context.

2.2 Complex Information Needs

Within the field of information retrieval, Christopher Manning describes "an information need is a topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need" [134]. There is a wide variety of information needs, considering the diversity of user requirements for information access.

In this thesis, I focus on complex queries, which I define as multifaceted, knowledge-heavy queries that require comprehension and thorough research. In this section, I will review the literature and prior studies on different aspects of query complexity, specifically structured queries, factoid and non-factoid queries, and hard queries.

2.2.1 Structured Queries

Although this thesis does not directly address structured queries, it is important to provide an overview of this literature to contextualize this research. This approach contrasts with the search engines that process free-text queries and documents. Structured queries are formalised expressions used to retrieve specific information from a well-defined structure [134], such as relational databases, knowledge graphs, or search engines using Boolean logic.

As already discussed, Boolean retrieval is a search technique that employs Boolean operators (AND, OR, NOT, etc.) to combine keywords or logical relationships for retrieving relevant information from a document collection. These operators enable the construction of complex, structured queries applicable to various domains, including high-recall search across patents [94, 157], legal proceedings [153, 215], and healthcare research [6, 71, 90].

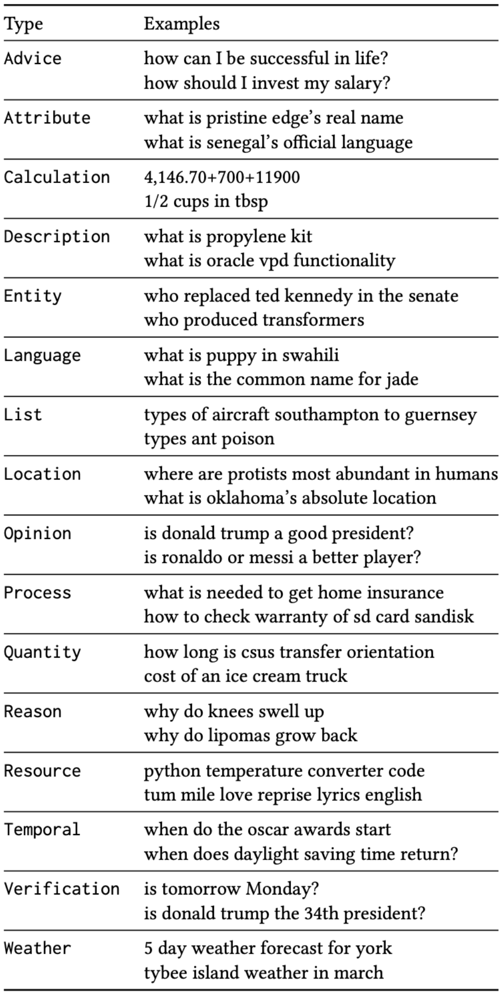
Structured queries extend to databases through SQL (Structured Query Language) [51], a standard query language for accessing and manipulating relational databases. SQL provides capabilities for creating complex manipulations of data using commands like SELECT, WHERE, and JOIN. There has been significant research on the related tasks of SQL-based question answering [87]. In recent years, these systems have been able to generate more complex SQL queries through leveraging PLMs [113, 178, 200], external context [184], and multi-stage PLM pipelines [60].

Lastly, and most relevant to this thesis, structured queries play a critical role in knowledge graph-based question-answering systems [45,81,84]. For example, using SPARQL [9] to extract precise answers from interconnected entities and relationships stored in large knowledge graphs, such as Wikipedia. For example, QUEST [132] is a retrieval dataset of entity-seeking queries with implicit set operations. While datasets like QALD [160], DBpedia–Entity [66] and TREC CAR [49] are specifically for the task of ranking entities.

In contrast, my work builds a joint probabilistic representation of documents (or passages) and entities, extending prior literature leveraging free-text knowledge graphs [48, 129, 222]. Specifically, I leverage the capabilities of pre-trained language models to identify and rank relevant passages and entities within query expansion pipelines, bridging structured representations and unstructured text to enhance retrieval performance.

2.2.2 Factoid Queries

Factoid queries are defined as concise, factual questions designed to retrieve specific information unstructured data [197], for example, “what is Senegal’s official language?”. The development of question-answering taxonomies has been an ongoing research area for decades [73, 107, 190], with roots tracing back to the 1970s [100]. Over time, these taxonomies have been extended to encompass a more granular variety of questions, particularly with a shifting focus to open-domain QA [24]. More recent work has emphasised understanding different types of output required to satisfy various information needs and the query itself [73, 107, 190]. Additionally, research has explored the impact of domain-specific factors on question-answering systems, including benchmarks such as TREC CAST [44].



Type	Examples
Advice	how can I be successful in life? how should I invest my salary?
Attribute	what is pristine edge's real name what is senegal's official language
Calculation	4,146.70+700+11900 1/2 cups in tbsp
Description	what is propylene kit what is oracle vpd functionality
Entity	who replaced ted kennedy in the senate who produced transformers
Language	what is puppy in swahili what is the common name for jade
List	types of aircraft southampton to guernsey types ant poison
Location	where are protists most abundant in humans what is oklahoma's absolute location
Opinion	is donald trump a good president? is ronaldo or messi a better player?
Process	what is needed to get home insurance how to check warranty of sd card sandisk
Quantity	how long is csus transfer orientation cost of an ice cream truck
Reason	why do knees swell up why do lipomas grow back
Resource	python temperature converter code tum mile love reprise lyrics english
Temporal	when do the oscar awards start when does daylight saving time return?
Verification	is tomorrow Monday? is donald trump the 34th president?
Weather	5 day weather forecast for york tybee island weather in march

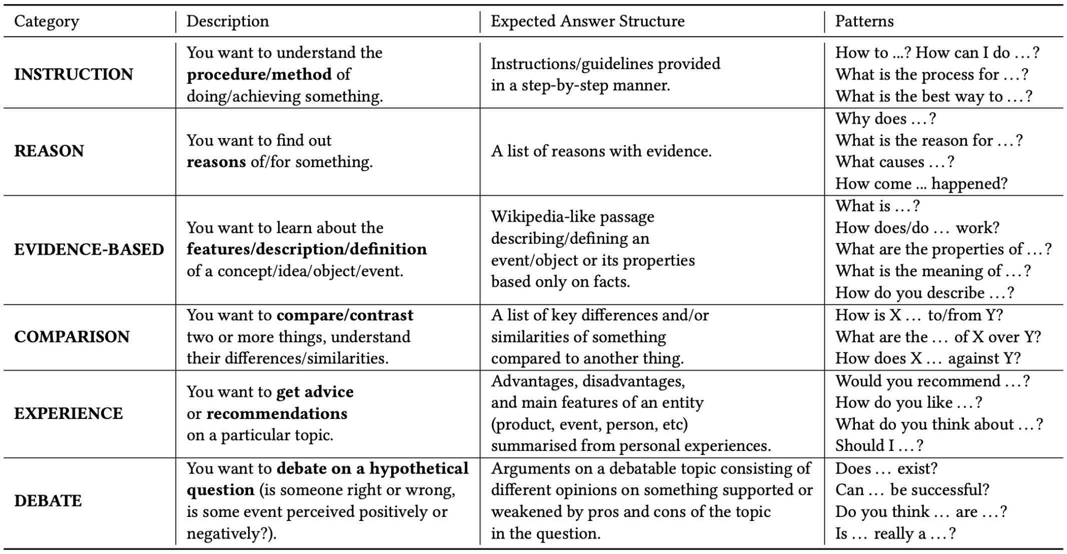
Figure 2.2: Question Intent Taxonomy [25]

A widely used benchmark focusing on the type of information need is the MS MARCO collection [148], which consists of web queries sourced from Bing's web QA system. Figure 2.2 depicts the different query intents taxonomy defined by Cambazoglu et al. [25], where information needs range from accessing lists of aircraft models, facts on specific countries, and calculation for cooking measurements. My findings support MS MARCO collection [148], and the associated TREC Deep Learning tracks [33, 34], largely focus on QA or basic lookups question intents, which I re-confirm through my analysis and annotations.

Nonetheless, factoid question answering has become increasingly challenging to benchmark system improvement, especially with the rise of highly effective pre-trained language models [149, 150]. For example, on TREC Deep Learning [33, 34], the median mean reciprocal rank (MAP) exceeds 0.8 for both document and passage ranking tasks. Similarly, the most effective systems on SQuAD [170], a leading benchmark for factoid questions over Wikipedia, achieve F1 scores over 93%. Thus, this thesis shifts focus away from factoid questions and instead develops a taxonomy for complex queries. Identifying and building datasets with greater headroom for model improvement allows developing and evaluating new state-of-the-art query expansion pipelines leveraging PLMs.

2.2.3 Non-Factoid Queries

Non-factoid queries required longer answers such as opinions or explanations to satisfy the information need [20], in contrast to the concise answers or facts typical of factoid queries. This category of queries has traditionally received less attention in research, with only a few studies and taxonomies addressing them [28, 62, 142, 196]. The most comprehensive recent work on categorising non-factoid question answering comes from Bolotova et al. [20], who presented a detailed taxonomy of non-factoid question categories. As shown in Figure 2.3, they conclude that non-factoid questions are more challenging than factoid questions and highlighted the need new datasets.



Category	Description	Expected Answer Structure	Patterns
INSTRUCTION	You want to understand the **procedure/method** of doing/achieving something.	Instructions/guidelines provided in a step-by-step manner.	How to ...? How can I do …? What is the process for …? What is the best way to …?
REASON	You want to find out **reasons** of/for something.	A list of reasons with evidence.	Why does …? What is the reason for …? What causes …? How come ... happened?
EVIDENCE-BASED	You want to learn about the **features/description/definition** of a concept/idea/object/event.	Wikipedia-like passage describing/defining an event/object or its properties based only on facts.	What is …? How does/do … work? What are the properties of …? What is the meaning of …? How do you describe …?
COMPARISON	You want to **compare/contrast** two or more things, understand their differences/similarities.	A list of key differences and/or similarities of something compared to another thing.	How is X … to/from Y? What are the … of X over Y? How does X … against Y?
EXPERIENCE	You want to **get advice** or **recommendations** on a particular topic.	Advantages, disadvantages, and main features of an entity (product, event, person, etc) summarised from personal experiences.	Would you recommend …? How do you like …? What do you think about …? Should I …?
DEBATE	You want to **debate on a hypothetical question** (is someone right or wrong, is some event perceived positively or negatively?).	Arguments on a debatable topic consisting of different opinions on something supported or weakened by pros and cons of the topic in the question.	Does … exist? Can … be successful? Do you think … are …? Is … really a …?

Figure 2.3: Non Factoid Question Taxonomy [20]

Several datasets have been created to focus on non-factoid queries, primarily within the context of question answering. For example, NFL6 [30] is derived from Yahoo's Webscope L6 collection, while ANTIQUE [65] consists of open-domain non-factoid questions, also sourced from Yahoo Answers. NLQuAD [185] focuses on long-form non-factoid question answering, requiring document-level language understanding over 13,000 BBC articles. Additionally, the ELI5 dataset [52] contains 270,000 diverse questions that demand explanatory, multi-sentence answers evidencing web search results.

In this thesis, I identify a gap in the information retrieval literature. Specifically, I developed criteria to focus on complex non-factoid queries based on the analysis of current datasets. These queries are multifaceted, knowledge-heavy, and require expensive reasoning and research to understand the topic. Minimal document retrieval datasets meet these criteria, with TREC Robust 2004 [198] being one of the few exceptions. Unlike typical QA-style queries, Robust04 focuses on more complex queries and the retrieval of long documents, such as “Commercial harvesting of marine vegetation such as algae, seaweed and kelp for food and drug purposes” or “Aside from the United States, which country offers the best living conditions and quality of life for a U.S. retiree?”. To address this gap, I built a new dataset to support research on advanced query expansion methods on complex information needs.

2.2.4 Hard Queries

A related, but not explicitly the same as query complexity, is the question “What makes queries challenging?”. For instance, TREC Robust 2024 [198] was developed to address challenging information needs by identifying queries where existing systems struggle. For example, Robust04 contains detailed “narratives” describing the information need, around 100-150 words, highlighting depth of the information needs, which are untimely hard to capture in short keywords queries and lead to lexical mismatch [18].

The literature has explored various factors that contribute to retrieval effectiveness. For example, longer queries have traditionally been more challenging to handle than short keyword queries [78]. However, this trend is shifting, as pre-trained language models have proven more effective at longer descriptive queries due to their natural language understanding capabilities [104, 150]. Meanwhile, factoid queries, often those that begin with ‘wh-’ words, are generally considered easier than more open-ended long queries [20, 63]. Furthermore, queries containing or concerning multiple entities have also shown to be challenging for search systems [17], especially neural models [179]. Additionally, Culpepper et al. [38] emphasise that users frequently reformulate queries to meet their information needs better, revealing significant performance variability across these query variants and underscoring the importance of query representation in retrieval systems.

Many of these factors are formally studied in the task of query performance prediction (QPP) [69, 224], which aims to predict the effectiveness of a query without relying on relevance judgments that are often unavailable in real-world scenarios. Generally, QPP can be categorised into pre-retrieval [68] and post-retrieval methods [11]. For example, incorporating features like query length [69], query term statistics [67], Clarity Score [37], Weighted Information Gain (WIG) [224], and modern neural-based QPP models [10, 13, 46].

In this thesis, I focus on complex queries that contain many of the characteristics of challenging queries. These include rich and multi-faceted information needs, entity-centric topics, long queries, and open-ended questions. However, I intentionally exclude challenging but non-complex queries from the scope of this work. This deliberate decision allows for a focused exploration of novel PLM-enhanced expansion techniques that are specifically tailored to this category of complex queries.

2.3 Document Retrieval

Document retrieval involves retrieving a ranked list of relevant documents from a document corpus given a user query [15, 134]. Performing this task necessitates a search system that can effectively run over document corpora containing millions or billions of documents [155]. Therefore, involving the application of a highly efficient algorithm to estimate query-document relevance over the entire corpus to produce an initial pool of candidate documents.

This section will provide an overview of the current literature to contextualise my advancements. This includes boolean retrieval, term-based retrieval leveraging inverted indexes, and probabilistic language models. I will also cover recent document retrieval techniques leveraging pre-trained language models, such as dense retrievals, ColBERT, and learned sparse retrieval.

2.3.1 Boolean Retrieval

Boolean retrieval is a search technique that uses Boolean operators to combine keyword logic to retrieve relevant information from a collection of documents. With this technique, Boolean query expressions are logical combinations of terms and operators AND, OR, and NOT. For example, "[generate] AND [rank]" would return all documents from the corpus containing words [generate] and [rank]. Similarly, using "[generate] OR [rank]" would return documents containing either term, while "[generate] AND NOT [rank]" excludes documents that include the term [rank].

This approach is closely associated with direct pattern matching in textual data and is often referred to as "grepping", after the Unix command grep. Boolean retrieval is particularly advantageous for certain types of information-seeking tasks that prioritise high-recall search. For example, patent search [94, 157], e-discovery [153, 215], and systematic review for healthcare [6, 71, 90] . Nonetheless, boolean retrieval has inherent limitations, such as the inability to capture the contextual relevance or the relative importance of documents based on the frequency or proximity of terms. Boolean queries operate on rigid matching rules and are the opposite of the search system required for complex queries.

2.3.2 Term-based models

Term-based retrieval models leverage term distributions within the query, documents, and overall corpus statistics to approximate query-document relevance for effective ad hoc search. These models rely on inverted indexes [36] as a core data structure that facilitates efficient search operations. Inverted indexes enable offline processing of term occurrences at an individual document and document corpus level, significantly reducing the computational overhead at query time.

The vector space model [175] represents documents and queries as vectors in a multidimensional space, where each dimension corresponds to a term in the corpus. The similarity between a document and a query is measured using cosine similarity, enabling the ranking of documents over the corpus. This model overcomes the binary matching limitations of Boolean retrieval by allowing partial matches and ranking results. This is extended with the Term Frequency Inverse Document Frequency (TF-IDF) [187] as a statistical measure that reflects the importance of a term within a document based on the number of times a term appears in the document but is offset by the frequency of the word in the corpus.

Furthermore, Okapi BM25 [172] is a term-based retrieval model used to estimate the relevance of documents to a given search query, considering term frequency, document length, and the overall corpus statistics. BM25 is an extension of TF-IDF [187], which was developed to address some practical limitations, including incorporating a saturation function for term frequency, logarithmic IDF computation, and document length normalisation. Overall, term-based retrieval models, although simple, are still highly effective and strong baseline systems, participially in out-of-domain situations [191]. I show this during the thesis, where multiple innovation query expansion models are built upon these term-based foundations.

2.3.3 Language Models

A language model is a probabilistic function assigning a probability to strings composed of terms from a given vocabulary [134]. Specifically, this probabilistic framework represents and ranks documents based on their likelihood of generating a given query. This approach underpins the Query Likelihood (QL) Model [164] by constructing a language model from each document to assess the highest probability of generating a query given that language model. These probabilities are typically estimated using maximum likelihood estimation (MLE) [156], with techniques such as Jelinek-Mercer and Dirichlet smoothing to handle issues of data sparsity and unseen terms effectively. This probabilistic framework can be easily extended to incorporate other types of feedback and relevance signals. In this thesis, I make several advancements in language modelling by building new models based on ranked and generated content from PLM.

2.3.4 Dense Retrieval

Contextualised embeddings from pre-trained language models have brought the “dense retrieval” wave within the information revival community. This search paradigm allows initial retrieval without relying purely on sparse signals, offering a potential solution to the problems of word mismatch and polysemy. Dense retrieval models primarily leverage encoder-only PLMs, often referred to as “bi-encoders,” which take two inputs (query and document) and generate two representations. The model’s objective during training time is to learn to map query and document embeddings spatially close to each other; however, notably, they are encoded independently and not reliant on each other. Generally, document embeddings are pre-computed and stored in a vector database, such as Faiss [88]. At search time, a query is first embedded to create a "query vector," and methods such as approximate nearest neighbour (ANN) [80] are used to retrieve semantic similarity "document vectors" efficiently. Most current research literature leverages BERT [47] models and large datasets, such as MS MARCO [148], to train them suitably.

The Dense Passage Retriever (DPR) [91] employs distinct encoders for processing queries and textual content extracted from the corpus. Both encoders utilize the [CLS] representation derived from BERT as their respective output representations. In contrast, ANCE [212] adopts the representation from RoBERTa and employs a unified encoder for both the query and the document. For long documents where text content does not fit within BERT's 512-token context window, "max-passage" aggregation approaches are often applied. Here, documents are sharded into passages, and the passage with the maximum score represents overall document relevance. ANCE improved the effectiveness of DPR by constructing harder negative samples using the Approximate Nearest Neighbor (ANN) index.

Recent extensions have looked to improve the effectiveness of dense retrieval through improving new data and training methodologies [108] and including entity representations [193]. Furthermore, recent work has shows decoder-only models can be effectively leveraged for dense retrieval [220]. In this thesis, I compare dense retrieval compared to sparse and entity-centric retrieval methods on complex queries. Furthermore, I experiment using PLMs as part of expansion pipelines aiming to improves dense retrieval effectiveness.

2.3.5 ColBERT

In contrast to single-representation dense models such as ANCE and DPR, ColBERT [92] employs multiple representations for dense retrieval, where each query and document are encoded into a dense representations. They propose "late interaction" that allows interactions between terms in the query and terms in the documents from the corpus in a manner that is compatible with existing nearest neighbour search techniques. ColBERT uses ANN search as initial retrieval, followed by a late interaction mechanism for re-ranking to produce a final ranking.

ColBERT-TCT [112] manages to distill the knowledge from ColBERT's expressive late interaction functions for computing relevance scores into a simple dot product to enable single-step ANN search. This thesis does not focus much on ColBERT-style models, but I leverage ColBERT-TCT as my default dense retrieval component and baselines.

2.3.6 Learned Sparse Retrieval

Learned Sparse Retrieval (LSR) [55, 123] refers to an emerging family of retrieval models that combine the interpretability and efficiency of traditional sparse retrieval methods with the contexualisation of modern neural models. LSR models do not rely on dense vector representations;

instead, they create weightings over a vocabulary to represent a query and documents that are compatible with an inverted index. Thus, it provides both efficiency and explainability compared to dense retrieval.

For example, EPIC (Expansion via Prediction of Importance with Contextualization) [123] creates expanded dense document representations using a bi-encoder model and expanding with the WordPiece vocabulary. SPLADE [55] uses BERT MLM head and sparse regularisation to learn query and document sparse weightings and expansions, and other methods include uniCOIL [109], DeepCT [40], and DeepImpact [133]. In this thesis, compare learned sparse retrieval to traditional sparse and entity-centric retrieval methods on complex queries. I also incorporate pre-trained language models as part of expansion pipelines with the aim of improving LSR effectiveness.

2.4 Document Re-Ranking

With the advancement in deep learning and pre-trained language models, modern search systems are typically split into two distinct phases. First is the initial retrieval stage, where an efficient search algorithm retrieves a candidate pool of documents over the entire document corpus. Second, a more effective "re-ranking" algorithm, which may not be practical to run over the entire corpus, re-ranks the candidate pool to produce a final ranking shown to the user.

This has become the de-facto state-of-the-art document ranking setup, which I use case baselines and extend in this thesis. I cover literature across deep learning, encoder-only and encoder-decoder PLMs, and adaptive expansion.

2.4.1 Deep Learning

In the early-to-mid 2010s, the deep learning movement showed that neural networks could learn useful representations to understand and process human language. For example, pre-trained text embedding methods such as Word2Vec [139] or GloVe [158] can encode basic semantic relationships and contextual information between words.

Within information retrieval, there are several query-document representations trained using deep learning models or "representation-based models". For example, The Deep Structured Semantic Model (DSSM) [74] uses a Siamese network to map query and document pairs into a shared semantic space, enabling semantic similarity measurement. Furthermore, Convolutional Deep Structured Semantic Model (CDSSM) [182] extends DSSM to incorporate convolutional neural networks (CNNs) over terms in the document when building representations.

Deep learning models can also compute the similarity function between a query and a document to directly calculate a relevance score ("interaction-based models"). These interactions are typically operationalised using a similarity matrix with rows corresponding to query terms and columns corresponding to document terms. Unigram models like Deep Relevance Matching Model (DRMM) [61], and Kernel-based Neural Ranking Model (KNRM) [211] aggregate the similarities between each query term and each document term. While position-aware models like ConvKNRM [41], Position-Aware Convolutional-Recurrent Relevance Matching (PACRR) [75] and Co-PACRR [76] use additional architectural components to identify matches between sequences of query and document terms. The final relevance score is then calculated from the features extracted from the models and processed to produce a query–document relevance score. Moreover, representation-based and interaction-based approaches are not mutually exclusive. The DUET model [140, 141] is a well-known hybrid method that combines the representation-learning component with an interaction-based component responsible for identifying exact term matches. This thesis does not directly cover these older models due to the significant effectiveness gains since the emergence of pre-trained language models.

2.4.2 Encoder-Only PLMs

In contrast to bi-encoders, which independently encode queries and documents for dense retrieval, "cross-encoders" employ an encoder-only PLM to process both the query and candidate text during inference. These models directly capture contextual relationships and dependencies between the query and document, which enhances retrieval effectiveness. However, due to computational constraints, cross-encoders cannot be applied to the entire corpus at search time, necessitating a re-ranking step in the pipeline.

The first application of BERT for text ranking, by Nogueira and Cho [149], is where a query and a passage are BERT models with a classification layer to predict a relevance score. Using PLMs for document reranking brought about significant improvements in ranking precision, even being integrated into commercial web search engines to enhance ranking and question answering.[1]

Despite these performance improvements, PLMs cannot easily ingest the full text when ranking long documents due to the input constraints of these models. Various strategies deal with this problem; for example, BERT-maxp [39, 218] shard long documents into passages the model can score individually as a proxy for overall document relevance.

Other approaches include CEDR (Contextualized Embeddings for Document Ranking) [126] that use contextual embeddings from BERT for text ranking by incorporating them into pre-BERT interaction-based neural ranking models (DRMM [61], KNRM [211], and PACRR [75]). This approach addresses BERT's input length limitation by performing chunk-by-chunk inference over the document before combining relevance signals from each chunk.

Meanwhile, PARADE [104] deals with long documents by aggregating the representations across multiple passages. More precisely, PARADE chunks long text into passages and aggregates representation on each passage's [CLS] representation. The authors report results across

[1] https://blog.google/products/search/search-language-understanding-bert/

several pulling methods, such as average and max pooling, using a CNN [99], and using attention layers [195] over the passage embeddings.

Overall, cross encoders for re-ranking have significantly improved effectiveness, and I adopt CEDR and PARADE as state-of-the-art full-ranking baselines to measure gains that can be achieved by leveraging PLMs in query expansion pipelines on complex topics. For my PLM ranking capabilities, as part of my query expansion pipelines, I leverage encoder-decoder re-rankers [150], but these results should generalise to encoder-only or decoder-only re-rankers.

2.4.3 Encoder-Decoder PLMs

The use of pre-trained language models (PLMs) for re-ranking has been extended to include encoder-decoder architectures like T5 [150]. In this approach, the model is fine-tuned to generate the tokens “true” or “false”, indicating whether a given document is relevant to the query, and a softmax is applied exclusively to the logits of the prediction tokens during decoding to extract a score. Leveraging T5 for passage ranking has demonstrated superior effectiveness compared to BERT-based models, offering improved ranking effectiveness and greater efficiency [111]. Extensions to this include RankT5 [226], where a T5 model is fine-tuned with pairwise or listwise ranking losses to optimise ranking performances.

Nonetheless, similar to BERT-based ranking models, encoder-decoder models still face challenges when processing long documents that exceed context window limits. To address this, techniques like max passage aggregation have proven effective when combined with T5, demonstrating strong performance in document ranking [150]. Additionally, MORES [58] presents a modular re-ranker framework that leverages BART [102] for query-to-document cross-attention, offering a robust solution for long-document ranking leveraging encoder-decoder models.

In this thesis, I leverage encoder-decoder model’s [150] ranking capabilities as a core component in my novel query expansion pipelines. Specifically, I use these models in a number of ways to incorporate relevant text and knowledge into expansion models.

2.4.4 Adaptive Re-Ranking

Within information retrieval, there has been substantial work on iterative processes to improve the effectiveness or efficiency of retrieval systems [35, 120, 144]. For example, using query performance prediction [69,224], heuristics or models to vary the depth of pooling for annotation processes [57] or reduce the information finding effort of search users [14,16,118]. Additionally, Lv and Zhai [120] use heuristics to vary the per-query weighting of the original query and feedback information within Relevance Modeling, compared to pre-set across all queries.

A recent advancement in re-ranking pipelines is “adaptive re-ranking,” which enhances passage ranking effectiveness with minimal impact on efficiency [124]. Unlike traditional re-ranking pipelines that are limited by the recall of the initial candidate pool retrieved by the

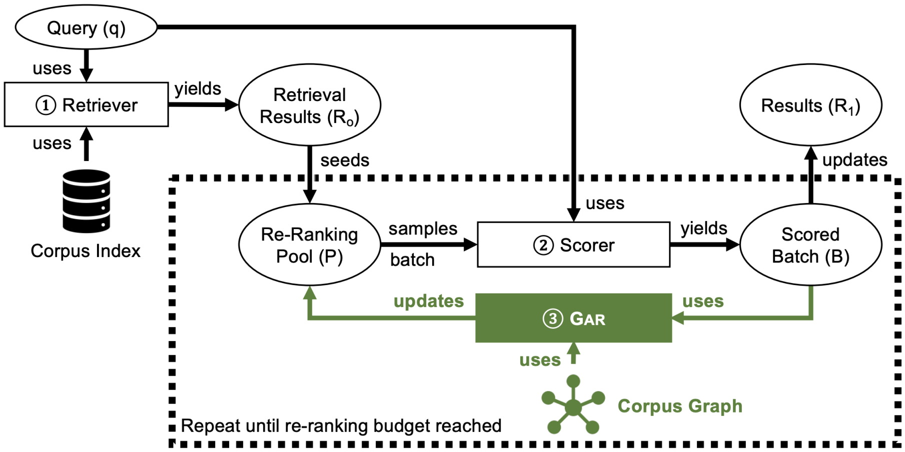


Figure 2.4: Traditional re-ranking exclusively scores results seeded by the retriever. GAR (Graph-based Adaptive Re-ranking) [124] adapts the re-ranking pool after each batch based on the computed scores and a pre-computed graph of the corpus.

first-stage search system and cannot be expanded during re-ranking, adaptive re-ranking enables dynamic updates to the candidate pool. An iterative process prioritises documents that are semantically similar to those already ranked highly, leveraging the Cluster Hypothesis [82], which suggests that relevant documents tend to be semantically related and form clusters.

Figure 2.4 illustrates this process in practice. Starting with a query q, an initial retrieval set R_0 is retrieved. The first batch of size b is added to the Re-ranking Pool P, and a PLM re-ranker (Scorer) scores this batch to produce B, and updates the ranked results R_1. The currently highest-scoring document from R_1 guides the selection of a second batch of documents using Graph-Based Adaptive Re-ranking (GAR) document similarity. These documents are added to P, PLM scored, and R_1 is updated. This iterative process, alternating between initial retrieval and the GAR process until R_1 contains the target number of document, balancing exploration and exploitation to improve relevance.

I build on this work by leveraging PLM re-rankers to dynamically update novel query expansion models, called "adaptive expansion". By extending adaptive re-ranking to document retrieval on complex queries, I explore the benefits of combining adaptive expansion and new advanced query expansion models that use PLMs.

2.5 Query Expansion

When users express their information needs through queries, these queries are often underspecified or suffer from lexical mismatches [18]. While numerous approaches have been developed to address these challenges, query expansion remains a widely used and effective solution. This technique enriches a query with additional context to help contextualise and support the retrieval of more relevant documents [36]. Initially, query expansion relied on static thesaurus lookups, but it has since evolved into sophisticated probabilistic and neural models.

Relevance feedback [134] is a query expansion and refinement technique first proposed in the 1960s. It involves user interaction to identify relevant documents from an initial ranking and query, before automatically reformulating the query by adding and re-weighting terms, and a new ranking is generated using this modified query. This concept was extended to *pseudo-relevance feedback*, where the system assumes that the top-ranked documents from the initial query are relevant, bypassing the need for explicit user feedback.

In this section, I cover the background literature on query expansion. This includes covering probabilistic term-based expansion models, entity based models, dense pseudo-relevance feedback models, and recent approaches expanding with generative content.

2.5.1 Term-Based Expansion

As discussed, inverted indexes allow efficient and effective retrieval, including vector space models, statical methods like TF-IDF and BM25, and probabilistic methods like Query Likelihood. Based on these foundations, multiple query expansion techniques have been proposed. For example, the Rocchio algorithm [173] is a relevance feedback method based on the vector space model. It refines the initial query vector by adding the mean vector of relevant documents while subtracting the mean vector non-relevant documents.

Furthermore, Bo1 [8] is a pseudo-relevance feedback method based on the Divergence From Randomness to identify terms in the relevant documents with differing distribution compared to the overall collection. Additionally, KL expansion [221], based on Kullback-Leibler divergence, is an asymmetric divergence measure originating in information theory.

Language modelling can be easily extended for query expansion; for example, the Relevance Model of Lavrenko and Croft [221] is an instance of a likelihood model that incorporates pseudo-relevance feedback and achieves strong empirical results. The probabilistic model captures term relationships and expands the query context by refining the query representation using an initial set of assumed relevant documents. Furthermore, RM3 [1] is an extension of the Relevance Model that adds query interpolation and is a common baseline for initial retrieval. Lastly, LCE [136], another extension of the Relevance Model, uses Markov Random Fields to incorporate term dependencies during expansion and uses arbitrary features to provide a framework beyond simple term occurrences.

In this thesis, I experiment with term-based query expansion models on complex queries. I make substantial contributions by developing new expansion models based on PLM-generated and ranked content.

2.5.2 Entity Expansion

Query expansion has been extended to include not just terms but also external knowledge, such as entities, to improve retrieval accuracy. Extensive research has focused on incorporating entity-based representations into document ranking systems [115, 116, 135, 180, 180, 207–209]. Prior work typically uses entity mentions present in the query and documents to help ground the task to an external knowledge base, where entity linking is the process of identifying entity mentions within documents [53, 77, 103, 163].

Specifically, prior work has effectively applied entity-based query expansion methods to enrich the query with useful concepts to help retrieve relevant documents [135, 208, 214]. For example, [171] developed an entity-based language model for document retrieval by representing queries and documents as bag-of-words and bag-of-entity-links. Furthermore, EQFE [42] enriches the query with KG entity-based features to improve document ranking, improving the hardest topics. Moreover, the Word-Entity Duet [210] framework uses word-based and entity-based representations to embed documents and queries for ad-hoc retrieval.

There has also been work that has leveraged pre-trained language models. For example, [180] shows that enriching queries and documents using an end-to-end PLM entity linking system [103] can provide knowledge-grounded context and improve initial retrieval. While [193] introduced a dense retrieval method that clusters entities within documents to produce multiple entity “views”, enhancing the understanding and interpretation of various facets of a document.

In my work, I explore entity expansion models on complex queries, and use pre-trained language models to rank precise feedback to enable the creation of new entity representations for retrieval.

2.5.3 Dense Expansion

The rise of dense retrieval has brought variants of complementary pseudo-relevance feedback and query augmentation techniques. For example, ANCE-PRF [219] leverages feedback documents retrieved by ANCE to enrich query representations by training a new PRF query encoder to output an updated query embedding. Li et al. [105] run a study on vector pseudo-relevance feedback and show that using Rocchio feedback is highly effective, where the mean of the passage representations directly updates the query representations based on a weighted combination of vectors. They also demonstrate that ColBERT-TCT PRF [105] is highly effective and propose a robust default experimental setup.

Another method of leveraging PLMs for query expansion, CEQE [145], extends the RM3 pseudo-relevance feedback technique that uses contextual embeddings to compute the probability of candidate expansion terms given both a query and a document. Furthermore, BERT-QE [223] incorporates the pseudo-relevance feedback context as part of the relevance prediction to improve re-ranking effectiveness. Lastly, DREQ [27] proposes to learn query-specific weights for entities within candidate documents for document re-ranking.

Other related methods include ColBERT PRF [201], which extends the multi-representation dense expansion identifying representative and important embeddings from the pseudo-relevant set to expand the original query embeddings. Similarly, work has leveraged pseudo-relevance feedback with learned sparse representations [96] to improve retrieval effectiveness. Overall, this work shows that traditional query expansion and PRF techniques can be extended to the pre-trained language model retrieval settings across dense, Colbert, and learned sparse retrieval. In this thesis, I show that by leveraging PLMs to rank and generate relevant feedback can develop new dense retrieval pipelines.

2.5.4 Generative Expansion

Pre-trained language models are known for their ability to generate coherent and fluent text [22, 192]. Unsurprisingly, these capabilities have been leveraged in prior work to improve core retrieval effectiveness. For example, encoder-decoder PLMs have been used to generate facets for queries [121, 176], improving search diversity and effectiveness. Encoder-decoder models have also been applied to document expansion through synthetic query generation, as demonstrated in Doc2Query [151] and DocTTTTTquery [151]. Additionally, research by Bonifacio et al. [21] on Inpars highlights the potential of PLMs for few-shot query generation, enabling efficient and scalable dataset creation. Lastly, Datta et al. [46] using a transformer bi-encoder improves retrieval effectiveness by combining original and selective expanded queries to mitigating query drift across sparse, generative, and ColBERT–PRF methods.

Moreover, Liu et al. [114] extract contextual cues from PLMs, augmenting and merging multiple questions to enhance the effectiveness of question answering. This is based on the idea that PLMs can act as knowledge bases [162] and provide useful content for retrieval. In the domain of passage ranking, HyDe [59] employs InstructGPT [154] to generate hypothetical document embeddings for dense retrieval. Other notable works focus on leveraging pre-trained language models for query-specific reasoning, contributing to improved ranking effectiveness [54, 159].

In this thesis, I present the first published work on query expansion using text generated from pre-trained language models. I show that PLMs can generate relevant content, exploring complex multi-turn prompting techniques, including subtopic generation for complex topics. I also explore PLM hallucinations (or non-relevant content) and methods to identify hallucinations and weight expansion models effectively.

Chapter 3

General Framework for Query Expansion

In this chapter, I outline this thesis's methodology and experimental setup. This includes defining the task setup, where I focus on document retrieval and formally define query expansion. In essence, I aim to develop (1) new query expansion models and (2) useful query context to improve document retrieval effectiveness. This setup is consistent across the research threads where I incorporate PLMs into novel query expansion pipelines. Lastly, I define the secondary task of entity linking, where relevant entities are ranked, and this is relevant to the work on latent entity expansion.

I outline the evaluation methodology of this thesis, which is based on the Cranfield evaluation paradigm. This paradigm uses an evaluation dataset containing a set of queries and known relevant and assumed non-relevant documents. Thus, a search system is evaluated based on a ranked list of documents, and evaluation measures depict system behaviour and allow system comparison. As I focus on query expansion, recall-oriented measures are the primary evaluation criteria, such as R@1000, NDCG, and MAP, and I use a 95% confidence paired-t-test for significance testing.

I define the complex query criteria, that I'll fully explore in Chapter 4, that classifies a query as complex if it is multifaceted, concerns multiple entities and concepts, and requires significant amounts of knowledge and comprehension to contextualise. Based on this criteria, TREC Robust 2004 and CODEC are the two core datasets for this thesis. Additionally, I outline how entity linkers and knowledge base, KILT KB, is used for entity-centric baselines and new methods in Chapter 5

Lastly, I define the experimental setup and baselines for this thesis. This includes term-based methods, entity-centric methods, dense retrieval models, neural re-ranking, and adaptive re-ranking. These systems allow us to understand where current systems fail and effectively evaluate the new query expansion pipelines that leverage PLMs. This strong experimental setup and methodology lays the groundwork for the next chapters.

3.1 Notations

Table 3.1: List of notation used throughout the thesis.

Symbol	Description
q	A query
d	A document
e	An entity
p	A passage
w	A term
Q	A query set
D	A document collection
R	An assumed relevant set of documents
R_k^D	A relevance-ranked document run of depth k
R_k^E	A relevance-ranked entity run of depth k
k	The depth of run
N	The number of documents in a collection
E	A knowledge base of entities
$s(q,d)$	Estimated relevance score for a query–document pair
$s(q,e)$	Estimated relevance score for a query–entity pair
q^o	The original query
D^o	Feedback document list
q^{exp}	The expanded query
D^{exp}	Second-pass document list based on expanded query
$\mathscr{P}_{QE}(.)$	A query expansion process that takes in an original query and the feedback set
$Rel(q)$	The set of documents relevant to a query
rel_d	A relevance score of a document
$\vec{q}$	A query vector
$\vec{p}$	A passage vector
$LS(.)$	A learned sparse representation
$\|R\|$	The number of feedback documents
$\|W_R\|$	The number of feedback terms
W_R	The set of unique feedback terms
β	The weighting of the original query for expansion
q_{hard}	A hard query
$q_{complex}$	A complex query

3.2 Task

This thesis proposes new query expansion methods to improve document ranking effectiveness. **Document ranking** systems return a relevance-ranked list of k documents $R_k^D = [d_1, ..., d_k]$, from a document collection, D, for a given natural language query, q. The core component of any search system is a scoring function, $s(q,d)$, which estimates a document's relevance with respect to the query.

Document retrieval algorithms can efficiently calculate these scores across the entire document collection, D. Thus, these models require efficient scoring functions that can be run across millions or billions of documents. On the other hand, document re-ranking is an optional second, more computationally costly stage where a scoring function re-orders a pool of documents.

Query expansion, $\mathscr{P}_{QE}(.)$ is a function that takes in the original query, q^o, and feedback documents, D^o, and produces an expanded query, q^{exp}. This expanded query is then re-issued to the ranking system to return a new list of documents D^{exp}. Note, I have generalised D^o here to be documents, however, this would be other content or structured information. This thesis looks to improve ranking effectiveness by incorporating PLMs into query expansion pipelines, which includes both improving the feedback, D^o, and the expansion models themselves, $\mathscr{P}_{QE}(.)$.

A secondary task that I investigate is **Entity ranking**. Entity ranking systems have to return a relevance-ranked list of k entities, $R_k^E = [e_1, ..., e_k]$, from an entity knowledge base, E, for a given natural language query, q. Similar to document ranking, entity ranking systems rely on a scoring function, $s(q,e)$, to produce an entity ranking. I specifically focus on leveraging entity ranking as a component of an expansion model to improve document ranking.

3.3 Evaluation

My evaluation primarily focuses on the effectiveness of ranking systems based around the Cranfield evaluation paradigm [165]. This paradigm is based on a sample of real user applications through building realistic information needs (queries). Then, measuring system performance by searching for relevant documents from a representative corpus of documents. Specifically, the search system that is evaluated returns a top-k set of documents, $R_k(q)$, in response to a given query, q, where $Rel(q)$ are those judged relevant documents.

The two most common measures for information retrieval effectiveness are precision and recall. Equation 3.1 shows how Precision [7] is simply the fraction of relevant retrieved documents. Thus, this measure focuses on achieving high-quality results with less concern about missing relevant documents.

$$Precision = \frac{|Rel(q) \cap R_k(q)|}{|R_k(q)|} \tag{3.1}$$

Conversely, **Recall** [7], as shown in Equation 3.2, is the fraction of relevant documents that are retrieved. Thus, this measure focuses on not missing relevant documents and is less concerned about retrieving more non-relevant documents.

$$Recall = \frac{|Rel(q) \cap R_k(q)|}{|Rel(q)|} \tag{3.2}$$

There are also metrics, such as **F1** [194], that combines precision and recall to identify methods that perform well across both. Specifically, F1 calculates the harmonic mean of precision and recall.

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3.3}$$

Unlike precision, recall or F1 that works over sets, average precision, **AP** [64], takes document order into account within the measure. This is important as having relevant documents in high ranks is essential for most information retrieval tasks. In Equation 3.4, Rel_i is 0 for non-relevant documents and 1 for relevant documents, and $Precision(q,i)$ is simply the precision at cutoff, i.

$$AP(q,k) = \frac{\sum_{i=1}^{k} Precision(q,i) \times rel_i}{|Rel(q)|} \tag{3.4}$$

Mean Average Precision, **MAP** [23], is simply the mean of AP across the query set, as show in Equation 3.5.

$$MAP(Q,k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} AP(q_j,k) \tag{3.5}$$

A more precision-focused measure would be, Mean Reciprocal Rank (**MRR**) [166]. Equation 3.6 shows how this is the average multiplicative inverse of the rank of the first correct answer across queries.

$$MRR(Q,k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{rank_j} \tag{3.6}$$

Unlike the prior measures that are based on binary relevance judgments, cumulative gain [83] and normalised discounted cumulative gain (NDCG) [83] are designed for situations of non-binary notions of relevance. Thus, the ranking and the relevance magnitude are incorporated into a single measure, i.e., showing if a search system can rank highly relevant documents in the top ranks.

I start by building up discounted cumulative gain, DCG [83], which includes a logarithmic discount factor to assign higher weights to the highly relevant documents in top ranks. Equation 3.7 depicts this, where Rel_i is the graded relevance of the i-th document in the ranked list.

$$DCG(q,k) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{3.7}$$

Normalised discounted cumulative gain, **nDCG** [83], extends DCG so that the measure is comparable across different systems and different queries. This is done by applying a discounting factor, $IDCG(q,k)$, which calculates the DCG for the perfect ranking of a query q (where the most relevant documents are in the highest ranks).

$$nDCG(q,k) = \frac{DCG(q,k)}{IDCG(q,k)} \tag{3.8}$$

As is convention with information retrieval evaluation, Precision, Recall, F1, and nDCG are computed for each query individually, whereas MAP and MRR are explicitly defined as the mean values across all queries. This distinction reflects that only MAP and MRR carry formal names for their averaged forms.

In this thesis, I primarily focus on recall-oriented measures because query expansion is a first-pass retrieval method. Furthermore, the target queries assume a user model in which many documents are required to understand the information need fully, unlike more question answering or conversational systems, where a depth of 3 [42] or 10 [33, 34] might be more appropriate. Thus, I assess the system runs to a run depth of 1,000, with my primary measures being R@1000, MAP, and nDCG. At times, I also examine more precision-oriented measures such as nDCG@10 and nDCG@20. I use ir-measures [122] for all my evaluations and a 95% confidence paired-t-test for significance using Spacy [72].

3.4 Complex Query Criteria

In Chapter 4, I conduct detailed analysis of the information needs of current and new datasets and system performance. Based on this analysis, I define a criteria for categorising queries of focus: "complex" queries.

Complex queries are defined as being multifaceted, concerning multiple entities and concepts, and requiring a significant amount of knowledge and comprehension to contextualise the topic. For instance, the query: "How is the push towards electric cars impacting demand for raw materials?". This information need requires knowledge from various documents, e.g., raw materials that constitute electric cars, supply chain dynamics, and sales growth. While simultaneously understanding critical entities and relationships (e.g. [Electric car], [Raw Materials], [Cobalt], [Lithium-ion battery] and [Demand]) becomes vital to grasping the intricacies of the topic. This query is complex for a human to research and comprehensively understand.

Specifically, **complex queries**, $q_{complex}$, are defined using the following criteria:

- **Multifaceted**: The topic elicits debate, contains multiple points of view, or is multifaceted. Specifically, open-ended and discursive questions are good examples of this criterion, where a good response would require an understanding of each of these facets.

- **Entities**: The topic requires understanding across multiple central entities and concepts (people, things, events, etc.). Understanding these concepts is vital to creating an effective knowledge-heavy response.

- **Comprehension**: The topic is sufficiently complicated and requires an educated adult to understand and reason across the information. For example, there are complicated connections and intricacies within the topic that need to be understood and explained as part of a response.

- **Knowledge**: The topic requires deep knowledge to contextualise satisfactorily. For example, the user would require significant research to expose the knowledge needed to understand the topic fully.

I define an annotation process where each query is assessed on the 4 complex criterion based on the following scoring range, i.e., strongly, partially, or does not meet the criterion. Queries are independently scored by each criterion (i.e., multifaceted, entities, comprehension, knowledge). Then, the overall complexity classification is annotated: *complex*, *partially complex*, and *not complex*.

In Chapter 4, I present CODEC, as a dataset that is built from the ground up based on complex topics. I also find that TREC Robust 2004 contains a large proportion of complex topics unlike other question answering-style datasets. I believe that new query expansion methods leveraging PLMs can significantly the effectiveness of complex queries through improved contextualisation.

3.5 Evaluation Datasets

Based on the complex query criteria, the core datasets for this thesis are TREC Robust 2004 and CODEC (see Section 4 for more details).

TREC Robust 2004 [198] focuses on poorly performing document ranking topics. This dataset comprises 249 topics, containing short keyword "titles" and longer natural-language "descriptions" of the information needs. Scaled relevance judgments are over a collection of 528k long newswire documents (TREC Disks 4 and 5). I use 5-fold cross-validation with standard folds from previous work [79]. In Section 4, I demonstrate that Robust 2004 contains complex topics and is a suitable evaluation dataset for this thesis.

CODEC [131] focuses on the complex information needs of social science researchers (economists, historians and politicians). This resource contains 42 essay-style topics and rich narrative developed by domain experts. CODEC encompasses two aligned tasks: document ranking over a focused web corpus (750k long documents) and entity ranking grounded to KILT KB [161]. I use the folds outlined within the online resource for 4-fold cross-validation.[1] I discuss CODEC construction in detail in Section 4.

Table 3.5 shows the overview statistics for each of these document ranking datasets. This includes the size of the collections, ($|C|$), number of queries, ($|Q|$), number of relevant documents

[1] https://github.com/grill-lab/CODEC

per query, ($|Rel(Q)|/|Q|$), and average number of terms per query, ($|q|/|Q|$).

Table 3.2: Statistics of primary document ranking datasets. From right to left: size of collection, number of queries, number of relevant documents per query, and average number of terms per query.

	$\lvert C\rvert$	$\lvert Q\rvert$	$\lvert Rel(Q)\rvert/\lvert Q\rvert$	$\lvert q\rvert/\lvert Q\rvert$
Robust04 (titles)	528,155	249	69.9	2.8
Robust04 (descriptions)	528,155	249	69.9	15.6
CODEC	729,824	42	91.3	12.5

Furthermore, I conduct initial analysis and evaluate baseline methods on more general datasets such as MS MARCO and the TREC Deep Learning family (2019, 2020, and DL-HARD). Table 3.5 show the statistics for each of these document ranking datasets.

MS MARCO [148] is a leading benchmark for web search that uses web queries that are candidates from Bing's web question answering system. MS MARCO contains many queries with sparse relevance labels based on Bing click data. This resource contains two tasks: passage and document ranking, with 8.8 million passages that map to 3.2M unique documents (web pages) making up the collections. Around one judged relevant passage per query creates a large training dataset for neural search methods. MS MARCO release training, validation, and test data splits.

TREC Deep Learning 2019 [34] & 2020 [33] The TREC Deep Learning 19/20 document track is built upon the MS MARCO collection, but with NIST annotators to provide judgments pooled to a greater depth. These resources contain 43 topics for DL-19 and 45 topics for DL-20 for document ranking, and 43 topics for DL-19 and 54 topics for DL-20 for passage ranking.

DL-HARD [130] is built from analysis of TREC Deep Learning 19/20. This provides a challenging subset of 50 topics focusing on topics where current passage and document ranking systems fail. I use the folds outlined within the online resource for 5-fold cross-validation.[2] This analysis shows the need for datasets built from the ground up for evaluating complex information needs (detailed in Section 4).

Table 3.3: Statistics of supplementary document ranking datasets. From right to left: size of collection, number of queries, number of relevant documents per query, and average number of terms per query.

	$\lvert C\rvert$	$\lvert Q\rvert$	$\lvert Rel(Q)\rvert/\lvert Q\rvert$	$\lvert q\rvert/\lvert Q\rvert$
DL-19	3,213,835	43	153.4	5.5
DL-20	3,213,835	45	39.3	6.3
DL-HARD	3,213,835	50	67.2	6.0

[2] `https://github.com/grill-lab/DL-HARD`

3.6 Knowledge Base

I explore using entity knowledge as a means of improving query expansion methods for document ranking. Specifically, by constructing a dual Relevance Model using entity mentions as well as terms, and leveraging PLMs to weight the most important entities [128] effectively.

I use KILT KB [161] for all ranking experiments and to ground my entity linkers. KILT KB is based on the 2019/08/01 Wikipedia snapshot and contains 5.9M preprocessed articles that are freely available. The entity pages are primarily text-based, with minimal structure to indicate headings or passages. KILT contains related knowledge-grounded tasks (fact-checking, open-domain question answering, entity linking, etc.) and provides inter-entity entity links based on Wikipedia mentions, which could be helpful when identifying how related entities are to each other.

3.7 Indexing

For initial retrieval, an efficient toolkit to index and retrieve documents for baselines and developed methods is required. I index text corpora using Pyserini version 0.16.0 [110], removing stopwords and using Porter stemming [205]. This allows the easy implementation of baselines and the development of term-based algorithms and expansion models leveraging standard Lucene inverted indexes.

Several developed methods are based on entity-centric retrieval, where a bag of entities represents documents or passages. Pyserini stores the respective entity mentions using the unique entity identifiers from the KILT KB as individual terms. For instance, the document “Apple just released the new iPhone. The Apple CEO is called Tim Cook”, and would be indexed as "[Apple], [iPhone], [Apple], [CEO], [Tim Cook]". This allows for the easy experimentation of entity-centric expansion methods and fair comparison to term-based equivalents.

For dense retrieval, I use FAISS [50] and exhaustive search to find the semantically closest document vectors with respect to the query vector. Documents are stored in GPU memory to allow for efficient experiments.

3.8 Entity Linking

Entity linking [181] is the task that involves identifying text spans and linking these entity mentions to their corresponding entities in a knowledge base (KB). For instance, consider the phrase “Apple just released the new iPhone”. In this example, an entity linker would identify the span [Apple] as an entity and link to the corresponding entity from the KB: https://en.wikipedia.org/wiki/Apple_Inc.. This task closely relates to Wikification [138], where both concepts and entities are linked to the KB, i.e., also linking [iPhone] in this example.

Retrieval systems have a long history of using entity links to improve effectiveness [42, 135, 208, 214]. Furthermore, entities play a critical role in complex topics, where core concepts span multiple documents and are an explicit means of representing knowledge. Thus, I conduct detailed analysis using entities and explore new expansion methods leveraging entities. Throughout this work, entity-linking systems are used:

REL is a lightweight neural entity linker that allows easy deployment and strong performance. I use the suggested setup for mention detection, i.e., Flair [4], which is a Named Entity Recognition (NER) model based on contextualized word embeddings. REL's pre-trained model is used for candidate selection that is based on the 2019-07 version of Wikipedia (i.e., closely aligns with the 2019/08/01 Wikipedia version for KILT).

WAT [163] is a wikification system to provide high-recall named entities and concept connections. WAT is the successor of TagME [53] and offers reasonable effectiveness that can be efficiently run over corpora-scale inputs. Previous studies show that high-recall information extraction techniques are required for successful usage in ranking tasks [89]. We use the entity linking API available on: https://sobigdata.d4science.org/web/tagme/wat-api .

BLINK [206] uses a two-stage approach for entity linking. First, BLINK performs retrieval in a dense space with a BERT bi-encoder [47] that independently embeds the mentioned context and the entity descriptions. Each entity candidate is then re-scored with a BERT cross-encoder and achieves state-of-the-art results on multiple datasets. We use the official repository: https://github.com/facebookresearch/BLINK.

GENRE [26] is a sequence-to-sequence entity linking model that generates entity links in an autoregressive fashion and is conditioned on the context. The authors use a pre-trained BART [101] model and show state-of-the-art performance across multiple entities linking datasets. Nonetheless, GENRE is impractical to run over large document collections, and its primary use is at the query level in this thesis. We use the repository provided by Meta authors: https://github.com/facebookresearch/GENRE.

ELQ [206] is an end-to-end entity linking model designed for questions, i.e., produces a list of relevant entities given an input question or query. ELQ uses a bi-encoder [47] to perform mention detection and linking in one pass jointly. This system is useful for downstream question-answering or search systems based on accurately generating relevant entities that may not be explicitly mentioned in the question. This package is available within BLINK package.

Entity linking is essential to this thesis; for example, in Chapter 5 with the work on LEE [128], entity links contained within queries (ELQ) and documents (REL) are used to propose a new dual query expansion model. This weighting of terms and words is extracted using PLM re-rankers [150] to improve state-of-the-art retrieval effectiveness significantly. This work shows that entity representations can play a critical role in improving the effectiveness of complex topics.

3.9 Baseline Systems

I provide details of baseline systems and the hyperparameter tuning setup I use across methods. Having a robust experimental setup is vital to understand system differences. Overall, I use cross-validation and optimise R@1000 on standard folds for Robust04 [79], CODEC [131], and DL-Hard [130]. On DL-19, I cross-validated on DL-20 and use the average parameters zero-shot on DL-19 (and vice-versa for DL-20).

3.9.1 Term-Based Methods

Term-based initial retrieval models are the backbone of information retrieval literature due to their efficiency, strong baseline performance, and explainability. This research uses these term-based and PRF baselines, including benchmarking term-based adaptive expansion, LEE, GRF, and GRM methods.

BM25 [172]: This is underlying term-based retrieval algorithm used within many of my sparse expansion approaches, with tuning $k1$ (0.1 to 5.0 with a step size of 0.2) and b (0.1 to 1.0 with a step size of 0.1).

Relevance Model (RM3) [1]: RM3 expansion is part of the pseudo-relevance feedback (PRF) paradigm, where an initial retrieved set of documents is assumed relevant, R, to help identify useful expansion terms, w, for query expansions. This approach allows automatic enrichment of a query with helpful information that can improve retrieval effectiveness. To identify valuable terms for expansion, Equation 3.9 shows how to estimate the probability of a term given the assumed relevant documents, $P(w|R)$:

$$P(w|R) = \sum_{d \in R} P(q|d)P(w|d) \tag{3.9}$$

Because I use BM25 as the initial retrieval algorithm rather than a probabilistic model, following standard practices [110], and obtain $P(q|d)$ by normalising the ranking scores from the initial retrieval. These scores are then converted into probabilities by dividing each normalised score by the sum over all documents, $\sum_{d' \in R} P(q|d')$. The probability of a term given a document, $P(w|d)$, is the term frequency divided by the document length. Thus, Equation 3.10 shows the expanded query, $\mathscr{P}_{QE}(q,R)$, as a combination of the language models of the original query, $P(w|q)$, and RM3 model, $P(w|R)$:

$$\mathscr{P}_{QE}(q,R) = \beta P(w|q) + \begin{cases} (1-\beta)P(w|R), & \text{if } w \in W_R. \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

This includes β (original query weight) as a hyperparameter to weight the relative importance of the two language models, and θ (number of expansion terms) limits the most probable pseudo-relevant terms, and $|R|$ (number of feedback documents) as the retrieval depth of the initial run. I tune *fb_terms* (5 to 95 with a step of 5), *fb_docs* (5 to 100 with a step of 5), and *original_query_weight* (0.1 to 0.9 with a step of 0.1). BM25 with RM3 expansion is the primary sparse expansion baseline of this thesis.

Latent Concept Expansion (LCE) [137]: This is a generalisation of Relevance Modeling that incorporates term dependencies during expansion. Furthermore, using arbitrary features provides a framework beyond simple term occurrence features that are implicitly used by most other expansion techniques. Following LCE [137], Equation 3.11 shows how I normalise the term distribution utilising the probability of a word given the collection, $P(w|D)$ (that I approximate for convenience with IDF(w,D)). Thus, this formulation will increase the relative weighting of less relevant terms from the collection. I expand and tune this model in the same manner as RM3.

$$P(w|R) = \sum_{d \in R} P(q|d)P(w|d)P(w|D) \tag{3.11}$$

I use the same query expansion methodology as shown in Equation 3.10, so it is comparable to RM3. Therefore, tuning *fb_terms* (5 to 95 with a step of 5), *fb_docs* (5 to 100 with a step of 5), and *original_query_weight* (0.1 to 0.9 with a step of 0.1).

3.9.2 Entity-Centric Methods

Entity-centric retrieval methods look to incorporate information for a knowledge graph to help factually ground or contextualise the retrieval process. Specifically, I use these methods as strong baselines for my work on latent entity expansion.

RM3-Entity: I follow the exact same setup and tuning as the term-based RM3 method. The core difference is that I use the query and documents entity vocabulary to build my expansion model. Specifically, this involves building entity language models of the query, $P(e|q)$, and documents, $P(e|d)$.

LCE-Entity: I follow the same methodology as the term-based LCE method. The difference is that I use the query and document entity vocabulary to build my expansion model. This involves building entity language models of the query, $P(e|q)$, and documents, $P(e|d)$. I also normalise the entity distribution utilising the probability of an entity given the collection, $P(e|D)$ (that I approximate for convenience with IDF(e,D)).

ENT [180]: I follow [180] work where queries and documents are enriched with entity context. I re-implement a method that follows their best standalone method, “Entities”, where I expand queries and documents with the unique names of linked entities, using ELQ for queries and WAT for documents. I parameter-tune BM25 in the same manner as my term-based BM25.

ENT ⇒ RM3: I extend ENT [180] to use RM3 expansion and tune parameters in the same manner as BM25 ⇒ RM3. Thus, creating a strong entity-centric PRF baseline.

Entity Query Feature Expansion (EQFE) [42]: EQFE shows that incorporating entity-based information from knowledge graphs (KG) improves the effectiveness of general ad-hoc ranking, with the most notable improvements made on complex topics [42]. I include the best-performing EQFE Robust 2004 run that is provided by the author, which is based on a pooled first-pass retrieval.

3.9.3 Neural retrieval methods

Contextualised embeddings from pre-trained language models allow vector-based search between query and document embeddings. I also include baselines focusing on learned sparse retrieval and comparable PRF methods. I use these methods to benchmark adaptive expansion and dense and learned sparse GRF methods.

ANCE [213]: I use an MS Marco fined-tune ANCE model and Pyserini's wrapper for easy indexing. Following the methodology in the ANCE paper, *ANCE+FirstP* takes the first 512 BERT tokens of each document to represent that document. For max passage aggregation, which is my default setup *ANCE*, shards the document into a maximum of four 512-token shards with no overlap, and the highest-scoring shard represents the document.

ColBERT-TCT (TCT) [112]: I adopt the TCT-ColBERT-v2-HNP's model using a max-passage approach for document retrieval. Documents are partitioned into passages of 10 sentences with a stride length of 5, with the title encoded within each passage.

ColBERT-TCT with PRF (TCT ⇒ PRF) [105]: Comparable to sparse PRF approaches, dense PRF uses feedback passage vectors to contextualise the original query vector. I adopt the Rocchio PRF approach [105] for ColBERT-TCT with PRF to allow different weighting of the query vector and the feedback vectors. Equation 6.3 illustrates that the updated query vector, $\mathscr{P}_{QE}(q,R)$, is the combination of the original query vector, $\vec{q}$, and mean of the retrieved passage vectors, $\frac{1}{|R|}\sum_{\vec{p}\in R}\vec{p}$.

$$\mathscr{P}_{QE}(q,R) = \alpha\vec{q} + \beta\frac{1}{|R|}\sum_{\vec{p}\in R}\vec{p} \tag{3.12}$$

The hyperparameters for feedback documents α and β to weight the relative importance of query and feedback vectors, $|R|$ controls the number of feedback vectors. I tune Rocchio PRF parameters: $|R|$ (2,3,5,7,10,15), α (between 0.1 and 0.9 with a step of 0.1), and β (between 0.1 and 0.9 with a step of 0.1).

SPLADE [55]: I index the term vectors using Pyserini [110] and use their "impact" searcher for max-passage aggregation. Similar to my dense setup, documents are partitioned into passages of 10 sentences with a stride length of 5, with the title encoded within each passage.

SPLADE ⇒ RM3 [55]: I draw on prior work combining pseudo-relevance feedback with learned sparse representations [96]. Equation 3.13 shows how I combine the normalised learned sparse representations of the query, $LS(w|q)$, with the representation of the combined feedback passages, $LS(w|R) = \frac{1}{|R|}\sum_{p\in R} LS(w|p)$. For the combined feedback passages, similar to [146], I normalise each sparse passage representation and aggregate them together before normalising again. Similar to RM3, I tune my parameters $|R|$ feedback passages (5,10,15,20,25,30), number of feedback terms $|W_R|$ (20,40,60,80,100), and original query weight β (between 0.1 and 0.9 with a step of 0.1).

$$\mathscr{P}_{QE}(q,R) = \beta LS(w|q) + \begin{cases} (1-\beta)LS(w|R), & \text{if } w \in W_R. \\ 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

CEQE [145]: I use the most effective runs: CEQE-MaxPool(fine-tuned) for initial retrieval comparison and (BM25+CEDR)+CEQE-MaxPool+CEDR for query expansion post-re-ranking comparison.

3.9.4 Neural Re-Ranking Systems

Neural re-ranking systems are more computationally expensive processes that cannot be effectively run over the entire collection of documents. Thus, these systems take in a set of documents and produce a new ranking. I use PLM re-ranking as a critical component within novel expansion pipelines to rank relevant content in adaptive expansion, and LEE and better weight the generated documents in GRM. I also use PLM re-ranked systems as strong baselines throughout this research.

T5-3B [150]: Following the paper [150], I shard documents in passages of 10 sentences with a stride length of 5 and use a max-passage aggregation approach. This checkpoint (`castorini/monot` is trained on the MS MARCO passage dataset for 10k steps (or 1 epoch), which has shown good zero-shot effectiveness. I use T5-3B for many experiments exploring "neural re-ranker" as part of the expansion pipeline, and for generality, I will use: **PLM**.

CEDR [126]: I use the CEDR-KNRM runs with BERT-base embedding [47] for document ranking that the authors provide.

PARADE [104]: This is a strong document re-ranking system, and I use the run from the ELECTRA-Base variant with attention [29], provided by the author.

3.9.5 Adaptive Re-ranking Systems

Using the building blocks from the initial work on passage ranking, I transfer to the document ranking setting. I follow the same experimental setup as [124] to allow a fair comparison. This allows for assessing adaptive re-ranking on long documents and benchmarking my proposed "adaptive expansion" methods.

Algorithm 1 details the implementation of the adaptive-ranking baseline algorithm. Specifically, I start with an initial BM25 retrieval set (R_1), use a batch size $b = 16$, a total re-ranking budget of $c = 1000$ documents, and implement three graph-based document similarity methods, denoted as G. These methods are outlined below and employ terms, entities, and neural

Algorithm 1 Graph-based Adaptive Re-Ranking [124]

```
Input: Initial ranking $R_0$, batch size $b$, budget $c$, corpus graph $G$
Output: Re-Ranked pool $R_1$
$R_1 \leftarrow \emptyset$                                        ▷ Re-Ranking results
$P \leftarrow R_0$                                               ▷ Re-ranking pool
$F \leftarrow \emptyset$                                         ▷ Graph frontier
repeat
    $B \leftarrow \textsc{Score}(\text{top } b \text{ from } P, \text{ subject to } c)$    ▷ e.g., monoT5
    $R_1 \leftarrow R_1 \cup B$                                  ▷ Add batch to results
    $R_0 \leftarrow R_0 \setminus B$                             ▷ Discard batch from initial ranking
    $F \leftarrow F \setminus B$                                 ▷ Discard batch from frontier
    $F \leftarrow F \cup (\textsc{Neighbours}(B,G) \setminus R_1)$   ▷ Update frontier
    $P \leftarrow \begin{cases} R_0 & \text{if } P = F \\ F & \text{if } P = R_0 \end{cases}$   ▷ Alternate initial ranking and frontier
until $|R_1| \geq c$
$R_1 \leftarrow R_1 \cup \textsc{Backfill}(R_0, R_1)$            ▷ Backfill remaining items
```

similarity within a document context, as opposed to passages.

In the implementation, each batch is scored using a PLM re-ranker (T5–3B), alternating between the initial BM25 retrieval and the application of G. Specifically, G finds the most similar unscored b number of documents to the highest-scoring document in R_1. I evaluate several G methods for identifying similar documents within the adaptive re-ranking framework using a PLM:

GAR-BM25 ⇔ PLM [124]: I modify this graph-based adaptive re-ranking (GAR) approach for passage by [124] for document ranking. Specifically, based on the highest currently re-ranked document, I issue the document terms as a BM25 query against the document index to identify the most similar documents. This batch of retrieved documents is then re-scored, and the document frontier is updated.

GAR-TCT ⇔ PLM [124] I use ColBERT-TCT dense representations to calculate document-to-document similarity. Thus, based on the highest-ranking document, I issue the mean of each document's passage vectors as the query vector and perform a max-passage exhaustive search over the FAISS index. Thus, a batch of 16 documents is returned for re-ranking.

GAR-ENT ⇔ PLM I also extend the GAR framework to represent documents using the WAT document entity links. Similar to GAR-BM25, I issue a BM25 query of the entities linked within the top-ranked documents against the document entity index. Thus, this method will find documents that are similar to those that are highly ranked based on overlaps in the mentioned entities.

For "adaptive expansion", I use the same setup and parametrisations for fair comparison. I replace these GAR methods with my expansion models leveraging PLM content. I replace the query-independent corpus graph (G) with our novel PLM-based query expansion models.

Chapter 4

Datasets with Complex Queries

The development of new machine learning models for ranking is an important area of Information Retrieval research. Based on recent advancement in ranking performance from pre-trained language models [217], both for research [104, 149, 150] and commercial applications[1], it becomes essential to be able to measure model improvement adequately. In this chapter, I focus on understanding what makes queries "complex" based on analysis of system performance and deep annotations of the TREC Learning tracks [33, 34]. These learnings help to establish the complex query criteria to identify topics where these new classes of models struggle. I go on to show that TREC Robust 2004 [198] and a new dataset I develop, Complex Document and Entity Collection (CODEC) [131], have complex queries that my thesis will use to build and evaluate new query expansion models.

Specifically, through my DL-HARD work [130], I augment TREC Deep Learning [33, 34] with detailed query annotations, covering aspects like question intent categories, answer types, wikified entities, topic categories, and result type metadata from a commercial web search engine. My findings indicate that current ranking models exhibit effectiveness primarily on specific query categories, particularly on short factoid-based questions. However, substantial room for improvement exists for other categories, such as those involving multiple facets or requiring long-form answers. This analysis supports the development of my complex criteria.

I review several datasets using the complex criteria and annotation guidelines to identify those suitable for benchmark systems, finding that the TREC Deep Learning datasets (19/20) contain only 2% and 7% of complex queries. I also analyse DL-HARD, but even this harder subset contains only 6% of complex queries. However, after considering multiple other datasets, I show that TREC Robust 2004 is a strong candidate for evaluating complex information with 51% partially and 46% strongly complex queries. Thus, TREC Robust 2004 is considered a core dataset for this thesis, and I show headroom on baseline systems.

Therefore, I develop a new dataset to allow the development and evaluation of new query expansion models. CODEC focuses on complex information needs from social science do-

[1] https://blog.google/products/search/search-language-understanding-bert/

mains, such as history, finance, and politics. This dataset includes essay-style queries, golden "narratives" encapsulating information needs, query facets, and dense annotations of relevant documents and entities. The resource builds complex topics using my complex criteria to facilitate developing and evaluating new retrieval models. Notably, I show that even basic query expansion methods using entities and query facets improve ranking effectiveness, serving as motivation for future advancements in this thesis.

4.1 Motivation based on DL-HARD

I conduct qualitative and quantitative analyses to ground the complex criteria in improving opportunities for current retrieval systems. I begin by focusing on TREC Deep Learning [33, 34], a prominent dataset within modern Information Retrieval. Specifically, deeply annotating the datasets to understand what makes queries challenging across systems. I use this learning to inform my complex query criteria.

4.1.1 Background

The MS MARCO passage and document collections [148] consist of queries, web passages or documents, and sparse relevance judgments between them. This dataset derives passage-level relevance judgments from the MS MARCO Question Answering dataset by treating any passage containing a correct answer for the query to be relevant. These judgments are transferred to the document level by labelling any document containing a relevant passage as also relevant.

In contrast to MS MARCO's sparse relevance labels, TREC Deep Learning tracks [33, 34] have more deeply assessed, albeit a smaller, subset of queries. The large volume of sparse MS Marco data for training and high-quality NIST judgments for evaluation, result in TREC Deep Learning being a significant step forward for evaluation. In recent years, this family of datasets have become the most prominent general ranking dataset in the Information Retrieval community.

Nonetheless, using TREC Deep Learning to test modern ranking algorithms is challenging. First, the reported system effectiveness is high, even for existing baseline systems, with a median mean reciprocal ranking above 0.8 for both 2020 document ranking and passage ranking tasks. This would suggest that current deep learning methods leave little headroom for improvement. Second, queries exhibit variations in difficulty, intent, and answer types, raising questions about the ideal queries to prioritize when evaluating state-of-the-art neural models. For these reasons, analysing TREC Deep Learning 2019 [34] and 2020 [33] tracks is an obvious choice to inform the complex criteria.

As part of this analysis, I release the Deep Learning Hard (DL-HARD) dataset[2] that builds

[2] DL-HARD is available at `https://github.com/grill-lab/DL-HARD`

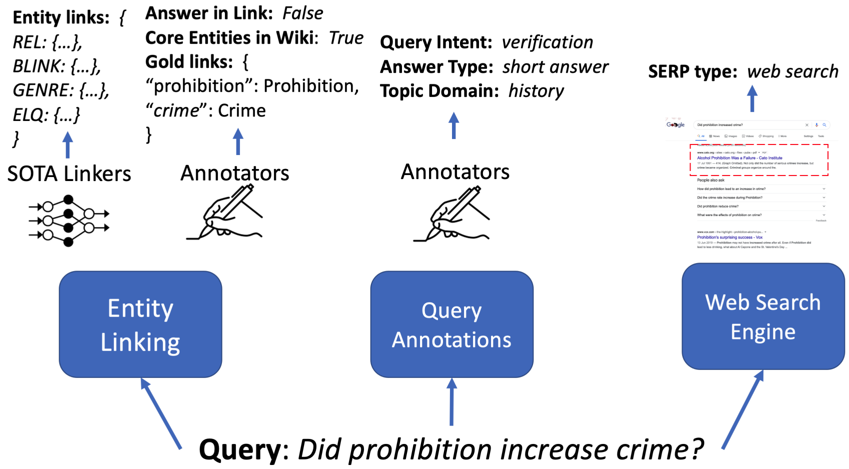


Figure 4.1: DL-HARD annotation process overview.

on TREC Deep Learning topics by extensively annotating them with question intent categories, answer types, wikified entities, topic categories, and result type metadata from a commercial web search engine. These rich annotations on the full four hundred queries, combined with TREC Deep Learning track official model submissions, enable an understanding of what makes queries challenging for current search systems. Thus, motivating the complex query criteria.

My analysis shows that short, factoid-based questions are substantially easier than questions involving multiple facets or requiring long-form answers. I perform experiments using the official submitted runs to TREC Deep Learn on DL-HARD and find substantial differences in metrics and the ranking of participating systems. Overall, DL-HARD is a new resource that promotes research on neural ranking methods by focusing on challenging topics. This work helps to lay the foundation of my complex query criteria and new datasets later in this chapter.

4.1.2 DL-Hard

This section outlines the methodology of constructing the DL-HARD dataset. I take the 400 queries from TREC Deep Learning 2019/2020 [33, 34], roughly 50 NIST annotated queries and 150 sparse annotated queries for each year, and richly annotate each query. These annotations include query intent, SERP results type, answer types, query entity links, and coarse topic categories.

I identify the "hardest" 50 queries for state-of-the-art retrieval and re-ranking models using these annotations and the official TREC track runs. Considering that around 25 of these queries have sparse annotations, I annotate these queries with additional judgements to approximate evaluations.

Specifically, by utilising the official system runs, I compute the median nDCG@10 (indicative of precision) and the median R@1000 (indicative of recall) for document ranking. These metrics enable establishing an approximate ranking, Figure 4.2 shows how to differentiate the

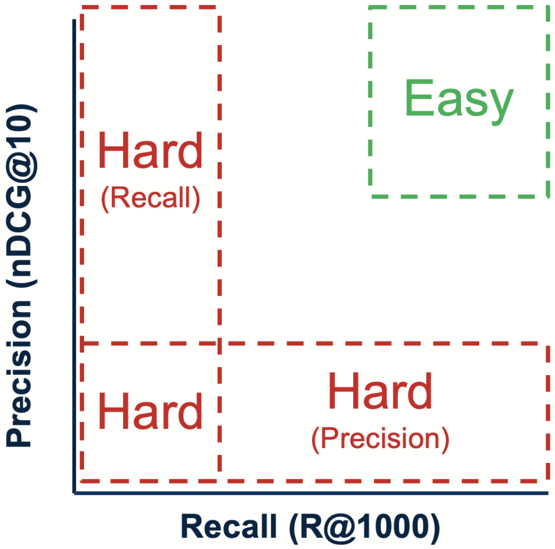


Figure 4.2: Hard vs easy queries based on retrieval effectiveness

"hardest" queries with lower recall and precision from the "easiest" queries with higher recall and precision.

The annotators consider both when and how systems struggle. I consider behaviour in a first-pass candidate retrieval (candidate recall) and second-pass re-ranking (retrieval in top ranks). Queries with either type of failure are candidates for inclusion in DL-HARD. Each candidate topic is individually labelled and resolved across all annotators. Annotators primarily focus on effectiveness with some consideration of system failures for non-interesting reasons. For example, poor system performance due to missed stopwords or tokenization issues, i.e. "why did the us volunterilay enter ww1". Furthermore, under-specified queries pose difficulty for both assessors and search engines in providing definitive answers, i.e. "who is robert gray" (multiple Robert Grays) or "cost of interior concrete flooring" (local and ambiguous). Thus, after multiple rounds of qualitative and quantitative annotations, the 50 hardest queries are selected on which to base the complexity analysis.

Relevance Assessments

DL-HARD uses the full NIST assessments for previously judged topics. There are also new passage and document level judgments provided for unjudged queries from TREC Deep Learning. I perform relevance assessment on a graded scale using the same guidelines as the original track for the new judgments. This ensures a reasonable density of annotations for the complexity analysis. I assess passages returned in the MS MARCO QA corpus and the documents from which they are drawn. Converse to MS MARCO sparse judgments, which generally include only one relevant passage per query, assessing all of the top ten responses. Experienced IR researchers perform the annotations.

Table 4.1: DL-HARD Judgment statistics including DL labels and newly developed assessments.

Rel. label	DL Doc	New Doc	DL Psg	New Psg
0	5,164	145	2,403	292
1	2,498	180	707	379
2	454	93	570	149
3	309	146	305	189

Table 4.2: Top 20 systems' effectiveness on DL-HARD compared with TREC DL for the 2020 document ranking task.

	NDCG@10			Recall@1000		
	DL-HARD	**DL-20**	**% Diff**	**DL-HARD**	**DL-20**	**% Diff**
Mean	0.419	0.626	-20.8%	0.545	0.736	-19.1%

To calculate agreement with the NIST assessors, I additionally judge the top QA passage responses for 24 queries from Deep Learning (12 from each year). I find Krippendorff's alpha is 0.47 on the passage judgments for these queries and 0.43 on the document judgments, which indicates moderate agreement. Krippendorff's alpha drops to 0.12 when transferring passage assessments to documents, illustrating the difficulty of automatically transferring passage assessments. For this reason, I adopt document-level relevance judgments for the official DL-HARD document ranking task.

System Performance

To show the effectiveness difference, I measure official TREC 2020 document run submissions on DL-HARD and compare to the original Deep Learning Track to (1) determine whether the dataset differs in system behaviour and (2) measure differences in system rankings (swaps) on this dataset. For binary metrics, I consider labels of two or greater to be relevant.

The 2020 system effectiveness for Deep Learning Track, DL-HARD and the relative differences is shown in Table 4.2. On an average relative basis, DL-HARD NDCG@10 is 21.1% lower and Recall@1000 is 19.6% lower. There are similar findings when evaluating the 2019 document task, and it shows headroom for system improvement.

Additionally, many system swaps occur when comparing the Deep Learning system rankings to DL-HARD rankings. This includes each system changing on average 4.6 places, with some systems changing as many as 12 places. This is supported by Kendall's Tau coefficients of 0.696 (2019) and 0.641 (2020) when comparing the TREC Deep Learning Track and DL-HARD system rankings. This large number of swaps supports that removing easier queries allows for greater differentiation and more precise system comparison.

Similarly, I evaluate the 2019 and 2020 Deep Learning systems on the 25 new sparse annotations using the official runs. These results cannot be directly compared to the DL Track, as these queries have new judgments. Nonetheless, the top 10 systems only have an NDCG@10 of 0.314 and RR of 0.452, indicating DL-HARD topics with new judgments are challenging for modern systems.

These results support that DL-HARD is more challenging for current ranking systems. I now show my methodology for deeply annotating these queries. This will allow the development of complex query criteria that should generalise across datasets and ranking models.

4.1.3 Annotations

To understand different aspects of what makes queries complex, I richly annotated all 400 TREC Deep Learning queries. I detail the annotation taxonomies below, including question intent, SERP result type, answer type, entity annotations, and coarse topic category.

Question Intent Annotation

Leveraging the question intent taxonomy designed for MS MARCO web questions [25], I adopt a fine-grained bottom-up taxonomy in contrast to other frameworks. Applying the Query Intent Categories and guidelines, I annotate all 400 official TREC Deep Learning queries, encompassing 13 intents: Advice, Attribute, Calculation, Description, Entity, Language, List, Location, Opinion, Process, Quantity, Reason, Resource, Temporal, Verification, and Weather. Each annotation is conducted by at least one annotator, with ambiguous instances resolved through majority agreement among the three annotators. To my knowledge, this is the first resource to make these annotations publicly available.

SERP Result Types

To retrieve the Search Engine Results Page (SERP), I manually issue every query on a Desktop browser to an English-language Google search engine from the United Kingdom in “incognito mode”. I manually examine the results, preserving the raw HTML content for inclusion in the resource. Queries susceptible to localization issues, encompassing location, region-specific language, or time concerns, are flagged. These queries may be excluded due to their unanswerable nature without access to local context, which is not provided at query time.

For each query, I annotate the type of rich results returned in the SERP and whether the Knowledge Graph [152] is used (the raw HTML shows the schema elements). Although many possible types of rich results may be present in a SERP, the ones highlighted below are the most prevalent for TREC Deep Learning queries:

- *Spell correct or suggestion*: Shows a suggested spelling correction or alternative query.

- *Knowledge Graph (KG)*: Returns a specific answer entity, list of entities, or their attributes from structured entity data. This includes media results for television, movie, and music entity information.

- *Dictionary*: Provides a dictionary definition of one or more words.
- *Weather*: Shows the weather forecast for a locale via an embedded panel.
- *Map*: Shows a Maps vertical result, optionally with possible driving directions.
- *Web Short Answer*: Shows a specific string short answer, possibly with a separate supporting evidence passage from a web result.
- *Web Passage*: Shows a passage (or portion of a list or table) from a web result. It may highlight possible answers.
- *Web Search*: Shows a standard list of '10 blue links'.

Answer Type Annotations

Previous manual and automatic annotations focus on the type of question intent or the SERP result type. Therefore, I create a new target answer type for MS MARCO web queries. The manual answer type labels are from all authors with a majority vote resolution. To develop the types, I follow a bottom-up multi-round curation similar to that used for query intents [25]. The answer types are:

- *Definition* - A single passage precisely and completely answers the information need. These are most commonly associated with the Description and Language query intents.
- *Factoid* - A specific short fact answer to a question. These are often associated with Entity, Attribute, Quantity, and Location intent types.
- *Short answer* - A short passage (approximately a sentence) generally satisfies most information needs. These are usually associated with Description and other factoid-like intents.
- *Long answer* - A long passage or full document is needed to answer the query. These are associated with Description, List, and Process intents.
- *List* - More than one answer, passage, or entity with justification is needed to answer the query.
- *Maps* - A structured map answer is needed; this is associated with Location and local Calculation intents.
- *Weather* - A structured weather result; corresponds to the Weather intent type.
- *Comparison* - A comparison of two or more entities. These are associated with Description intent types.
- *Guide* - A guide answer is a long semi-structured answer to satisfy the Process intent.

Query Entity Annotation

Entity linking [32] and semantic parsing [19] of question queries is an important component of modern QA systems. However, the existing TREC Deep Learning queries do not have standard automatic or manual annotations. I include four state-of-the-art entity linkers developed for documents and queries: REL [77], Blink [206], GENRE [26], and ELQ [206]. I run these annotators with high-recall setups, preserving score information for downstream applications, which is important for entity-based retrieval models [42]. Based upon the automatic results, I create gold entity links to Wikipedia and metadata about the entities, i.e. (1) whether the query entity is in Wikipedia and (2) whether the Wikipedia entity satisfies the query.

Coarse Topic Categories

Following TREC Conversational Assistance Track (CAsT) [43], I provide a breakdown of topics by coarse subject domain. For example, Business & Finance, Education, Entertainment & Celebrity, Food & Travel, Health, History, Language & Literature, Law & Politics, Local, Mathematics & Science, Sports, and Technology.

4.1.4 Research Questions

As part of the analysis on complex queries, I set a research question to better understand what aspects make a query challenging for current systems. Specifically, my work focuses on the following question:

- **RQ1.1: What identifying features do hard queries have?** This research question analyses the differences between normal (TREC Deep Learning 2019/20) and hard queries (DL-HARD). I cover question intents, SERP result type, answer type, and topic categorisation. This helps build a picture of what makes a query more generally complex.

4.1.5 RQ1.1: What identifying features do hard queries have?

In this research question, I use the deep query annotations and DL-HARD (i.e., the classified 50 most challenging queries) to identify specific attributes that are important to overall query complexity.

Question Intent Annotation

The distribution of the query intents on the complete TREC Deep Learning datasets as well as DL-HARD is shown in Table 4.3. The most notable difference is the increase in List intents from TREC Deep Learning (10.2% across 2019 and 2020) to DL-HARD (34.7%). The annotators note that list queries are harder as the user seeks multiple entities or facts that could span many documents. For example, topic 877809, “What metal are hip replacements made of?”, requires an understanding of different types of hip replacements and the composite materials contained in each. Having to find diverse knowledge spanning multiple sources is an important characteristic to consider when it comes to query complexity.

The proportion of Quantity intents in DL-HARD is much lower as most of these queries are either simple factoid-QA questions (“hydrogen is a liquid below what temperature”) or highly underspecified and should be clarified (e.g. “cost of interior concrete flooring”). Language (e.g., “definition of laudable”) and Weather (“how is the weather in jamaica”) intents are also filtered out. Thus, general and open-ended queries are likely a more useful type of complexity for evaluating current retrieval systems versus queries that require very specific knowledge.

Table 4.3: Query Intent Categories for DL and DL-HARD (document ranking).

Intent Category	**DL-2019**	**DL-2020**	**DL-HARD**
Attribute	1	5	1
Description	21	20	20
Entity	3	4	3
Language	0	2	0
List	7	2	17
Process	1	1	0
Quantity	5	6	3
Reason	3	4	4
Verification	1	1	2
Weather	1	0	0

Table 4.4 shows whether query intent is a strong heuristic for systematically identifying hard queries by mapping the SERP annotations onto 2019/2020 TREC Deep Learning runs. List and Entity intents have the largest negative Pearson correlation Coefficients (PCC) with Recall@100 and nDCG@10. At the same time, Attributes (e.g., “when was the salvation army founded”) and Verification (e.g., “do google docs auto save”) have the largest PCC, suggesting they are not hard. Overall, these findings support open-ended entity-centric queries, where knowledge spans many documents, should be considered part of the complexity criteria.

SERP Result Types

The distribution of the response types for assessed TREC Deep Learning and DL-HARD topics is shown in Table 4.5. The most frequent response type is a Web Passage (e.g., “is cdg airport in main paris”), which is unsurprising given that the queries are questions originally used for QA. It shows that over 20% of the TREC Deep Learning queries are answered directly with short factoid answers (e.g., “where is the show shameless filmed”), with 12.5% of results from a structured source (e.g., “how many sons robert kraft has”). This important point is unpacked later when I show that TREC Deep Learning (even DL-HARD) does not contain a large density of complex query criteria.

Table 4.4: Query Intent Category on DL 2019/20 document ranking systems (systems above median). Mean and Pearson Correlation Coefficient (PCC) across DL assessed queries.

	NDCG@10		Recall@100	
Query Intent	**Mean**	**PCC**	**Mean**	**PCC**
Attribute	0.699	0.10	0.842	0.22
Description	0.615	-0.02	0.750	-0.04
Entity	0.567	-0.07	0.616	-0.02
Language	0.778	0.12	0.862	0.00
List	0.568	-0.07	0.726	-0.20
Process	0.758	0.09	0.727	0.09
Quantity	0.594	-0.04	0.791	0.00
Reason	0.596	-0.03	0.686	0.01
Verification	0.691	0.05	0.726	0.09
Weather	0.735	0.05	0.930	-0.06

Table 4.5: SERP result types distribution for DL Track and DL-HARD (document ranking).

SERP Result	DL-2019	DL-2020	DL-HARD
Dictionary	1	3	1
KG	1	5	2
Weather	1	0	0
Web Passage	24	25	28
Web Search	12	9	18
Web Short Answer	4	3	1

DL-HARD contains fewer Dictionary and Weather result types on average compared to TREC Deep Learning. However, DL-HARD contains 36% Web Search response types compared to 24% on TREC Deep Learning, highlighting that multiple long-form documents are likely required to satisfy the query. For example, you could imagine a user needing to read many sources to understand the "causes of military suicide". The fact that a query requires a long-form answer to satisfy the information need is an important aspect of the type of query complexity I will focus on and aligns with requiring lots of diverse information across sources on what to ground the response.

Although the Google answer quality is not explicitly assessed, I observe only 2 instances of clear failure due to imprecise or ambiguous queries. This indicates that existing models (neural or otherwise) can adequately satisfy most of these "easy" factoid queries.

Table 4.6 shows the SERP categories and mean and correlation with respect to nDCG@10 and Recall@100. This supports the idea that Web Search is a strong feature for identifying hard queries. For example, "Medicare's definition of mechanical ventilation" results in a Web Search as Google question-answering system cannot directly answer the question. The median effectiveness of TREC categories (traditional, neural, and pre-trained language model systems) all have nDCG@10 and R@100 under 0.25. Conversely, KG, Web Passage, Dictionary, and Web Short Answer are all neutral to easy across recall and precision measures. This further supports the idea that queries requiring a long-form response are an interesting complexity feature on which to focus my research.

Table 4.6: SERP result type on DL 2019/20 document ranking systems (systems above median). Mean and Pearson Correlation Coefficient (PCC) across DL assessed queries.

	NDCG@10		**Recall@100**	
SERP Result	**Mean**	**PCC**	**Mean**	**PCC**
KG	0.577	-0.05	0.794	0.11
Dictionary	0.748	0.13	0.722	0.04
Weather	0.735	0.05	0.464	-0.06
Web Passage	0.647	0.13	0.684	0.04
Web Search	0.535	-0.20	0.581	-0.16
Web Short Answer	0.621	0.00	0.731	0.05

Answer Type Annotations

The answer types have strong associations with query intent types. However, I find that the Description intent is often quite general and does not provide guidance on the type of information needed for the answer. For example, more granular categorising of Description intent queries: “the difference between a company’s strategy and business model is” as Comparison or “what is the most popular food in Switzerland” as Factoid answer types. This fine-grained annotation is essential because these answer types are useful features for topic hardness.

Table 4.7 shows the answer type breakdown for the assessed Deep Learning and DL-HARD topics. It is clear that there are fewer Factoid responses (41% across TREC Deep Learning and only 10% on DL-HARD) and more List answers (10% across TREC Deep Learning and only 30% on DL-HARD) within DL-HARD. DL-HARD also has more Short Answer (+15%) and Long Answer (+8%) types compared to TREC Deep Learning. This further supports that QA-style queries (Factoid) are less interesting for my complexity criteria compared to longer written responses (Long Answer) or multiple aspects of the answer (List).

Coarse Topic Categories

Following TREC Conversational Assistance Track (CAsT) [43], I provide a breakdown of topics by coarse subject domain in Table 4.8. For 2019 I observe frequent Deep Learning topic categories to be Health, Science, and History. In 2020 there is a shift to more Entertainment, Business and Finance topics, and less Science and Health. There is also an increase in Language topics, with predominately definition queries. The largest category for DL-HARD is Health, a challenging category that often requires long answer responses. However, it’s unclear whether the domain itself provides the complexity or the query within the domain, and something I don’t explicitly focus on.

Table 4.7: Answer Type distribution for DL Track and DL-HARD (document ranking).

Answer Type	**DL-2019**	**Dl-2020**	**DL-HARD**
Comparison	3	2	0
Definition	9	7	7
Factoid	12	24	5
Guide	0	1	0
List	9	0	15
Long Answer	6	10	13
Multi-Answer	0	1	0
Short Answer	3	0	9
Short Description	0	0	1
Weather	1	0	0

Table 4.8: Topic domain category for DL Track and DL-HARD (document ranking).

Topic Domain	**DL-2019**	**DL-2020**	**DL-HARD**
Business & Finance	1	6	3
Entertainment & Celebrity	0	9	0
Food & Travel	5	3	4
Health	10	4	20
History & Education	5	6	8
Language & Literature	1	4	2
Law & Politics	2	2	2
Local	1	0	1
Mathematics & Science	13	6	10
Sports	1	3	1
Technology	4	2	0

4.1.6 Conclusion

In this section, I conduct analysis on TREC Deep Learning [33,34] to help motivate the development of the complex criteria. I deeply annotate TREC Deep Learning queries and classify “easy” and “hard” queries based on system performance. These annotations include query intent, SERP results type, answer types, query entity links, and coarse topic categories.

Based on this analysis of these categorised queries, I form my understanding of characteristics that I can build into my complex criteria. Specifically, to avoid focusing on QA-style queries where a user is looking for a specific fact to satisfy their information need. Although this type of query can be challenging and have some complexity, for example, “Can fever cause miscarriage in early pregnancy” is part of DL-HARD, I feel there is more research headroom within other query types.

Instead, I should focus the complex criteria on more open-ended and discursive topics, where information and knowledge span multiple sources. My analysis supports that the query should require a long-form response and contain many entities or facts to satisfy the information need. Such queries would be a strong indicator of complexity and require new search systems to achieve success.

Nonetheless, I identify limitations with TREC Deep Learning and DL-HARD dataset construction that make them poor candidates for evaluating complex queries. These limitations are:

- **Judgments**: The MS MARCO passage and document collections [148] consist of queries, web passages or documents, and sparse relevance judgments between them. This dataset derives passage-level relevance judgments from the MS MARCO Question Answering dataset by treating any passage containing a correct answer for the query to be relevant. These judgments are transferred to the document level by labelling any document containing a relevant passage as also relevant.

  While this label transfer method offers efficiency and dataset scalability, it introduces potential issues, including (1) the fact that all QA queries have an associated single passage answer and (2) the document corpus is limited to passage candidates for the queries. Despite MS MARCO being a cornerstone of information retrieval research, these aspects makes MS MARCO artificially easy.

- **QA-Style Queries**: As already discussed, the queries for the TREC Deep Learning family [33, 34, 130] originated from the MS MARCO Question Answering dataset. It is thus unsurprising that almost all queries are QA-style and looking for specific facts. This is supported by over 50% of Deep Learning 19/20 and DL-HARD being answerable by a Web Passage via Google's search engine and large proposition of Answer Types being Definition, Factoid, or Short Answer. This is supported by other literature [12].

  The analysis supports that my complex criteria should focus on queries that require knowledge spanning multiple sources and a long-form response containing many entities or facts to satisfy the information need. This is the opposite of the vast majority of the queries in MS MARCO and TREC Deep Learning. Even for DL-Hard, it is clear that a large proposition of the queries likely do not satisfy this criteria. This is confirmed later in this chapter.

- **DL-HARD Judgments**: Although DL-HARD's queries focus on challenging topics, around 50% of queries still have relatively sparse judgments. Specifically, the non-NIST queries were only judged to a depth of 10 compared to NIST topics that contain hundreds or thousands of judgments per query. These sparse judgments were adequate to support my query analysis; however, for core datasets for this thesis, DL-HARD would require more annotation density. I ultimately chose to focus my efforts on CODEC, a new dataset focused specifically on complex queries.

Overall, this analysis of query effectiveness provides the basis for my complex criteria, which will be formalised in the next section. I need to find two core datasets to assess new query expansion models, either within the literature or develop new datasets myself. Based on my findings, TREC Deep Learning, or even DL-HARD, will not be a good fit for evaluating new query expansion models on complex queries.

4.2 Complex Criteria and Current Datasets

Based on the findings of my analysis on TREC Deep Learning, I develop a formal criteria for complex queries that I wished to focus my research on. Specifically, complex queries are multifaceted, concerning multiple entities and concepts, requiring significant knowledge and comprehension to contextualise the topic. For example, open-ended knowledge-heavy queries such as, “How is the push towards electric cars impacting demand for raw materials?”. I also define a rigorous annotation process to assess whether queries and overall datasets are complex.

In this section, I outline the complex criteria and annotation process in detail. I annotate the family of TREC Deep Learning dataset - 2019, 2020, and DL-HARD - and analyse the results. Deep Learning 2019 and 2020 contain only 7% and 2% of complex and 25% and 13% partially complex queries. Although DL-HARD is constructed to have more challenging queries, only containing 6% are complex and 42% are partially complex queries (over 50% are not complex). These findings are not surprising given that the TREC Deep Learning query set comes from QA-style query logs that generally are juxtaposed to the complex criteria definition. Thus, I do not include these datasets in my core resources for benchmarking new methods.

I conducted a thorough search across Information Retrieval resources to identify datasets that had the potential to match my criteria. I identified TREC Robust 2004 [199], which focuses on poorly performing document ranking topics over long newswire documents (TREC Disks 4 and 5). These have rich topics containing “titles”, “descriptions”, and “narratives” to show the information need. For example, many of these queries have many of the characteristics I am looking for, such as: “What are the arguments for and against an increase in gasoline taxes in the U.S.?”. Based on the annotation process, I find that 41% are complex and 46% are partially complex queries (only 3% are not complex), concluding that TREC Robust 2004 is a suitable dataset for my research purposes.

Lastly, I outline how another dataset is required to properly evaluate my new query expansion methods and discuss the types of properties I would require. This motivates my development of CODEC in Section 4.3.

4.2.1 Complex Criteria

I learned much from the analysis of the family of TREC Deep Learning datasets and from developing DL-HARD. This included what makes queries “hard” for current retrieval systems, such as information spanning multiple documents, longer and richer responses, reasoning, and domain expertise. I also discussed aspects that make queries generally “easy”, including QA-style questions, factoid on knowledge graph lookups, or basic definition and descriptions questions. I take these insights and develop a more focused criteria for the complex queries I will focus on.

Specifically, I focus model development on categories of queries that receive less research interest. Thus, not focusing on challenging specific queries, such as "how long will methadone stay in your system" or "medicare's definition of mechanical ventilation" from TREC Deep Learning. Instead, focusing on open-ended and knowledge-rich queries, such as "causes of military suicide" or "what drives poaching". This query category is multifaceted, concerning multiple entities and concepts, and requires significant knowledge and comprehension to contextualise the topic. You could imagine professional researchers (i.e., military historians, policy-makers, journalists) spending considerable time exploring sources to understand key arguments, concepts, and facts about these topics. For example, surveys show that many legal researchers, recruitment professionals, and healthcare researchers require high-recall Boolean or structured queries over domain-specific collections to ensure they find the relevant information [174].

Thus, improving model effectiveness in this query category has the potential for significant research impact and practical application. In particular, I hypothesise that integrating pre-trained language models into multi-stage query expansion pipelines will improve document retrieval effectiveness on these types of complex queries. This will leverage PLM's capabilities to both rank and generate relevant content, serving as the foundational elements of new query expansion that can provide additional reasoning or contextualization over multiple documents and entities.

Now, I will outline my explicit criteria and annotation process for complex queries. Based on my analysis and multiple rounds of discussions amongst multiple information retrieval researchers, I identify four core criterion:

- **Multifaceted**: The topic elicits debate, contains multiple points of view, or is multi-faceted. Open-ended and discursive questions are good examples of this criterion, where a good response requires an understanding of each of these facets.

- **Entities**: The topic requires understanding across multiple central entities and concepts (people, things, events, etc.). Understanding these concepts is vital to creating an effective knowledge-heavy response.

- **Comprehension**: The topic is sufficiently complicated and requires an educated adult to understand and reason across the information. For example, the topic has complicated connections and intricacies that need to be understood and explained as part of a response.

- **Knowledge**: The topic requires deep knowledge to contextualise satisfactorily. For example, the user would require significant research to expose the knowledge needed to understand the topic thoroughly.

Annotation Process

I define an annotation process to understand what queries are complex and which datasets are suitable for this research. I use experienced information retrieval researchers to assess the four complex criterion based on the following scoring range:

- **Strongly meets criterion (2)**: The topic is a strong example for this criterion, e.g., for "Multifaceted", the topic would contain many distinct facets or points of view.
- **Partially meets criterion (1)**: The topic is a partial example for this criterion, e.g., for "Multifaceted", the topic would contain a small number of distinct facets or points of view.
- **Does not meet criterion (0)**: The topic is a bad example for this criterion, e.g., for "Multifaceted", the topic would contain a single answer or point of view.

Queries are independently scored on each criterion by experts (i.e., multifaceted, entities, comprehension, knowledge). If there is uncertainty, these are discussed and resolved as a group. Next, leveraging these four complex criterion scores, the annotators add an overall complex query classification:

- **Complex query (2)**: The topic is a good example of a complex query.
- **Partially complex (1)**: The topic is not comprehensively complex but contains some strong aspects of complexity.
- **Not complex (0)**: The topic is a bad example of a complex query.

This complex criteria and annotations process allows selecting the datasets that suit this thesis. In Section 4.3, I present CODEC, a dataset built from the ground up based on complex topics leveraging this criteria.

4.2.2 Research Questions

These research questions focus on annotating current datasets to assess whether they contain a reasonable number of complex queries, and ensuring I focus on representative datasets during this thesis. I also conduct qualitative analysis on annotated queries present in these datasets to motivate future model improvements.

- **RQ:2.1: Are TREC Deep Learning datasets complex?** This research question focuses on annotating TREC Deep Learning 2019, 2020, and DL-HARD datasets. I discuss the findings attributed from the annotations and the difference between "hard" and "complex" queries.

- **RQ:2.2: Is TREC Robust 2004 dataset complex?** This research question focuses on annotating TREC Robust 2004, which I identify as a possible dataset containing the desired properties. I discuss insights based on annotations and conduct qualitative analysis on TREC Robust to motivate future model development.

- **RQ:2.3: How do baselines systems perform on TREC Robust 2004?** Evaluate strong baseline systems that include sparse, dense, learned sparse, and entity-centric retrieval methods, with and without re-ranking. Look into query failures to motivate expansion methods I develop in the coming Chapters.

4.2.3 RQ:2.1: Are TREC Deep Learning datasets complex?

In this research question I discuss the complex criteria annotations on TREC Deep Learning 2019, 2020, and DL-HARD datasets. I follow the process discussed previously and full results can be seen in Appendix A. I also discuss these findings and conduct qualitative analysis on a number of interesting queries.

TREC Deep Learning 2019

Table 4.2.3 shows aggregated annotation for the 43 queries present in TREC Deep Learning 2019. This shows the four complex criterion (i.e., Multifaceted, Entities, Comprehension, Knowledge) and the overall complex classification, and the number of and percentage of queries that fall within each classification (i.e., does not meet, partially meets, strongly meets).

Table 4.9: Aggregated annotation of complex criteria on TREC Deep Learning 2019. Number of queries and percentage of query set within each criterion and annotation class.

	Multifaceted	**Entities**	**Comprehension**	**Knowledge**	**Complex**
Does not meet (0)	**27 (62.8%)**	18 (41.9%)	17 (39.5%)	**20 (46.5%)**	**29 (67.4%)**
Partially meets (1)	12 (27.9%)	**22 (51.2%)**	**21 (48.8%)**	18 (41.9%)	11 (25.6%)
Strongly meets (2)	4 (9.3%)	3 (7.0%)	5 (11.6%)	5 (11.6%)	3 (7.0%)

These results show that very few queries are strongly Multifaceted, and in fact, 63% do not meet this criterion at all. This is driven by many specific description or definition queries like “exons definition biology” and “who is Robert Gray”. More queries partially meet the Entities criterion (51%), where understanding several entities in detail is needed to comprehend the topic, for example, “What is an all surveillance analyst?”. However, 42% of queries are not heavily entity-centric, such as “How long is the life cycle of a flea” where you do not require a deep understanding of multiple entities.

Although still a high proportion does not meet the Comprehension criterion (39.5%), there are slightly more queries that partially meet (49%) and strongly meet (12%) the criterion. For instance, a query like “axon terminals or synaptic knob definition” requires a relatively complicated understanding of the area to come up with a precise definition. Lastly, only 12% meet the

Knowledge criterion, with most of the queries lacking the required research levels to strongly meet the criteria.

Overall, 67% of queries are classified as not complex, 25% as partially complex, and only 7% of queries are complex. Specifically, the only three complex queries are “causes of military suicide”, “why did the us voluntarily enter WWI”, and “what are the social determinants of health”.

TREC Deep Learning 2020

Table 4.2.3 shows aggregated annotations for the 45 queries present in TREC Deep Learning 2020. This shows the four complex criteria, the overall complex classification, and the number and percentage of queries that fall within each classification.

Table 4.10: Aggregated annotation of complex criteria on TREC Deep Learning 2020. Number of queries and percentage of query set within each criterion and annotation class.

	Multifaceted	**Entities**	**Comprehension**	**Knowledge**	**Complex**
Does not meet (0)	**40 (88.9%)**	17 (37.8%)	**30 (66.7%)**	**34 (75.6%)**	**38 (84.4%)**
Partially meets (1)	4 (8.9%)	**26 (57.8%)**	13 (28.9%)	10 (22.2%)	6 (13.3%)
Strongly meets (2)	1 (2.2%)	2 (4.4%)	2 (4.4%)	1 (2.2%)	1 (2.2%)

Overall, these queries follow a relatively similar distribution to the 2019 Deep Learning. Although, these are even less complex. Specifically, almost 90% are not partially or strongly Multifaceted. This includes queries like “definition of laudable” and “how often do button quail lay eggs”. Around 58% of queries partially meet the Entities criteria, where a deep understanding of a few entities is required to satisfy the topic. For example, “what metal are hip replacements made of”, where I need to understand the different metal compositions.

Nonetheless, TREC Deep Learning 2019 has 67% of queries that do not meet Comprehension and 76% that don’t meet Knowledge criterion. Therefore, it is not surprising that 84% of queries are not complex, 13% are partially complex, and only 2% are complex. The only identified complex query is “what does a psychological screening consist of for egg donors”. These results support that Deep Learning 2019 and 2020 are not complex datasets based on the criteria.

DL-HARD

Table 4.2.3 shows the annotation statistics of DL-HARD in the same table structure as I have previously described. Despite these queries being selected because they are hard for current search systems, many of these queries do not have complex query characteristics. For example, 58% do not meet the Multifaceted criteria, 30% partially meet, and 12% meet the criterion. Most queries (68%) partially meet the Entities criterion, with 18% not meeting and 14% meeting this criterion. Comprehension and Knowledge criteria similarly have large proportions of queries that do not meet the criterion (30% and 46%) with a smaller proportion of queries that strongly meet the criterion (12% and 20%).

Table 4.11: Aggregated annotation of complex criteria on DL-HARD. Number of queries and percentage of query set within each criterion and annotation class.

	Multifaceted	**Entities**	**Comprehension**	**Knowledge**	**Complex**
Does not meet (0)	**29 (58.0%)**	9 (18.0%)	15 (30.0%)	**23 (46.0%)**	**26 (52.0%)**
Partially meets (1)	15 (30.0%)	**34 (68.0%)**	**29 (58.0%)**	17 (34.0%)	21 (42.0%)
Strongly meets (2)	6 (12.0%)	7 (14.0%)	6 (12.0%)	10 (20.0%)	3 (6.0%)

Based on these individual criterion proportions, it is unsurprising that 52% of DL-HARD queries are not complex, 42% are partially complex, and only 6% are complex. Despite these queries being challenging relative to other Deep Learning queries, they are predominately not complex queries. For example, queries like “who is thomas m cooley”, “define: geon”, and “what is a alm” that may be hard for current systems but are clearly not complex. Only three queries in DL-HARD are classified as complex, i.e., “causes of military suicide”, “causes of stroke?”, and “what drives poaching”.

Overall, these results support the notion that DL-HARD, like Deep Learning 2019 and 2020, does not contain a sufficient number of complex queries to be a core dataset for this thesis.

4.2.4 RQ:2.2: Is TREC Robust 2004 dataset complex?

I identified TREC Robust 2004 [198] as a dataset that contains queries closer to the desired characteristics. In this research question, I discuss the Robust 2004 annotations of the complex criteria and conduct qualitative analysis. I follow the annotation process I previously outlined, and full results can be seen in Appendix A.

Table 4.2.4 shows annotation statistics across the 249 queries present in TREC Robust 2004. This shows the four complex criterion (i.e., Multifaceted, Entities, Comprehension, Knowledge) and the overall complex classification, and the number of and percentage of queries that fall within each classification (i.e., does not meet, partially meets, strongly meets).

Table 4.12: Aggregated annotation of complex criteria on TREC Robust 2004. Number of queries and percentage of query set within each criterion and annotation class.

	Multifaceted	**Entities**	**Comprehension**	**Knowledge**	**Complex**
Does not meet (0)	7 (2.8%)	0 (0.0%)	12 (4.8%)	2 (0.8%)	7 (2.8%)
Partially meets (1)	102 (41.0%)	100 (40.2%)	111 (44.6%)	78 (31.3%)	114 (45.8%)
Strongly meets (2)	**140 (56.2%)**	**149 (59.8%)**	**126 (50.6%)**	**169 (67.9%)**	**128 (51.4%)**

When I consider the Multifaceted criterion, 56% strongly meets and 41% partially meets and only 3% does not meet this criterion. For example, “What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?”, is a query that contains multiple dimensions across chemistry, medical care, and drug development. Similarly, the topics are very knowledge-rich, with 60% strongly meeting the Entities criterion and 40% partially meeting this criterion. For instance, "What are the pros and cons of Great Britain's universal health care system?", the user would need to understand concepts like the National Health Service (NHS), British politics, benefits systems, etc.

Similarly, a high proportion of queries strongly or partially meet the Comprehension and Knowledge criterion. Specifically, 51% strongly meet, 45% partially, and only 5% do not meet the Comprehension criterion. For example, "What is the extent of U.S. raw timber exports to Asia, and what effect do these exports have on the U.S. lumber industry?" requires complex reasoning to respond reasonably. Nonetheless, some queries in TREC Robust require minimal reasoning and more basic research, like "Find accounts of selfless, heroic acts by individuals or small groups for the benefit of others or a cause.". For the Knowledge criterion, 68% strongly meet, 31% partially, and 1% does not meet, highlighting the deep research required for the vast majority of the topics.

Overall, 51% of queries are complex, 46% are partially complex, and only 3% not complex. There are significantly more complex queries than general information retrieval datasets, such as TREC Deep Learning. For example, "How is Attention Deficit Disorder (ADD) diagnosed and treated in young children?" or "What were the causes for the Islamic Revolution relative to relations with the U.S.?". These topics are open-ended, contain multiple facets and central entities, and require significant knowledge and comprehension.

4.2.5 RQ:2.3: How do baselines systems perform on TREC Robust 2004?

Table 4.13 shows the effectiveness of current sparse, dense, entity-centric and PRF initial retrieval methods on TREC Robust 20004. These results highlight that dense retrieval methods have considerably worse recall effectiveness when compared to sparse methods. For example, ColBERT-TCT with PRF has 14-21% lower R@1000 compared to a tuned BM25 with RM3 expansion. Moreover, BM25 with RM3 expansion, i.e. simple sparse term expansion, has the highest R@1000 on Robust 2004 title queries, outperforming more complex embedding-based expansion techniques such as CEQE. In fact, no method significantly outperforms BM25 with RM3 in recall on these datasets. Interestingly on Robust 2004 descriptions, the entity-centric RM3 expansion method (ENT+RM3) has the best R@1000, highlighting the usefulness of entities, particularly within long queries with multiple entities mentioned.

When combined with PLM re-ranking, there's large absolute improvements in NDCG@20, for example, increasing on average by +0.10 on Robust 2004 titles across my first pass retrieval. Furthermore, there is even more improvement in description queries (+0.19), where long natural language queries benefit from PLM effectiveness. Thus, PLM re-ranking models offer an opportunity to improve the precision in early ranks that are the basis for effective query expansion methods.

Overall, these results show that current sparse and entity-centric expansion methods are still

Table 4.13: The effectiveness of initial retrieval methods and then combined with PLM re-ranking (T5-3b). For first-pass retrieval significance testing is against BM25+RM3. For PLM re-ranking it is BM25+RM3 ⇒ PLM. “$^{+}$” significantly better and “$^{-}$” significantly worse, *bold* depicts best system.

		Robust04 - Title			Robust04 - Description		
		NDCG@20	MAP	R@1000	NDCG@20	MAP	R@1000
Retrieval	SPLADE	0.420	0.224^{-}	0.597^{-}	0.431	0.231^{-}	0.617^{-}
	TCT	0.428	0.233^{-}	0.637^{-}	0.392^{-}	0.214^{-}	0.595^{-}
	TCT+PRF	0.464	0.272	0.681^{-}	0.392^{-}	0.246^{-}	0.619^{-}
	ENT	0.416	0.252^{-}	0.714^{-}	0.434	0.257^{-}	0.722^{-}
	ENT+RM3	0.427	0.276	0.745^{-}	0.432	0.279	0.759
	BM25	0.422	0.252^{-}	0.705^{-}	0.392	0.227^{-}	0.664^{-}
	BM25+CEQE	0.458^{+}	0.310^{+}	0.764	-	-	-
	BM25+RM3	0.435	0.292	0.777	0.425	0.278	0.750
Re-Ranking	SPLADE ⇒ PLM	0.539^{-}	0.309^{-}	0.597^{-}	0.590^{-}	0.357^{-}	0.617^{-}
	TCT ⇒ PLM	0.524^{-}	0.327^{-}	0.637^{-}	0.589^{-}	0.346^{-}	0.595^{-}
	TCT+PRF ⇒ PLM	0.530	0.345^{-}	0.681^{-}	**0.575**$^{-}$	0.364^{-}	0.619^{-}
	ENT ⇒ PLM	0.530	0.357^{-}	0.714^{-}	0.612	0.397	0.722^{-}
	ENT+RM3 ⇒ PLM	0.532	0.366^{-}	0.745^{-}	0.607	**0.407**	**0.759**
	BM25 ⇒ PLM	0.532	0.351^{-}	0.705^{-}	0.602	0.375^{-}	0.664^{-}
	BM25+CEQE ⇒ PLM	0.531	0.373	0.764	-	-	-
	BM25+RM3 ⇒ PLM	**0.536**	**0.377**	**0.777**	0.605	0.406	0.750

more effective than newer dense or learned sparse. Gains from PLM re-ranking passages provide significant gains in effectiveness, particularly in early ranks. We, therefore, PLMs these as a building block to improve expansion methods in Chapter 5 and Chapter 6.

4.2.6 Conclusion

In the section, I defined the complex criteria and annotation process. This allows a clear and demonstrable focus for this thesis on complex topics that are open-ended with multiple facets, concerning many central entities, requiring complicated reasoning, and requiring significant research to understand. Using this criteria, I show that TREC Deep Learning 2019, 2020, and DL-HARD contain few complex or partially complex queries. However, I show that TREC Robust 2004 contains many complex and partially complex queries and, thus, is a suitable dataset for this thesis. I show that baseline sparse and entity expansion methods are the most effective, and I will build on these directions in later Chapters. Moreover, another dataset is required to complement TREC Robust and ensure my proposed new methods generalise.

4.3 CODEC: Complex Document and Entity Collection

In the last section, I develop a complex criteria to categorise queries that are open-ended with multiple dimensions and central entities, require large amounts of reasoning and research to understand. Unfortunately, few public datasets have these aspects, and I solely identify TREC Robust 2004 with these characteristics. Considering this thesis focuses on developing new query expansion methods using PLMs, I require a second core dataset to evaluate current and proposed methods.

Based on these learnings, I build a dataset for complex topics from the ground up. CODEC (COmplex Document and Entity Collection) is a new resource that focuses on open-ended essay questions within the fields of history, economics, and politics. I worked with domain experts to develop topics based on the complex criteria and identify high-quality sources to build a focused document collection. This dataset supports both document ranking and entity ranking tasks, allowing development and benchmarking of new entity-centric query expansion methods.

In this section, I detail the construction of CODEC, which includes the topic creation, corpus developments, and extensive document and entity annotations. I then evaluate a number of state-of-the-art retrieval and re-ranking systems to show there is significant headroom to improve on complex topics. Specifically, I also show that the ground-truth method of expanding queries with rich text content and entities can improve effectiveness. This motivates the new query expansion methods I introduce in Chapter 5 and Chapter 6.

4.3.1 Methodology

Task Setup

CODEC supports both document ranking and entity ranking tasks. Having these tasks aligned in a single resource is rare and allows the development and evaluation of entity-centric document retrieval methods. The formulation of these tasks follows the same setup as those outlined in Methodology in Chapter 3.

Both document and entity ranking tasks use the same 42 topics, with dense annotations of both document and entity relevance. I develop a focused document corpus based on high-quality web documents curated using domain experts for the document ranking task. KILT KB [161] is the entity knowledge base for the entity ranking task, where each entity is a unique Wikipedia page. I also provide entity links for the document corpus to align both tasks.

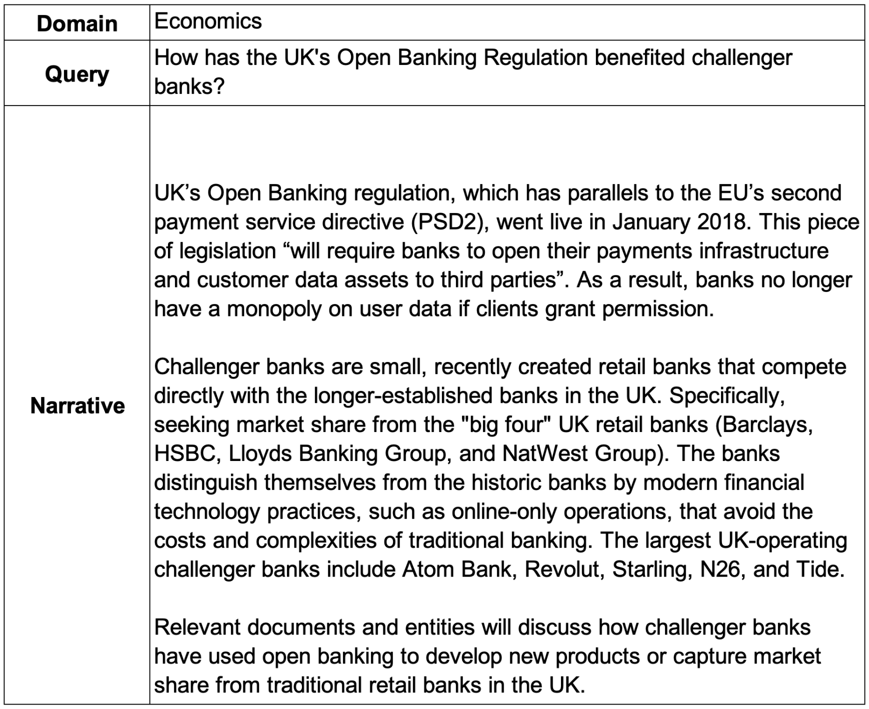
For the experimental setup, I provide four pre-defined "standard" folds for k-fold cross-validation to allow parameter tuning. CODEC is suitable for initial retrieval or re-ranking evaluation using this test collection. This setup encourages exploration of the related entity and document ranking tasks; however, both tasks can also be undertaken in isolation.

Topic Generation

CODEC provides complex topics that intend to benchmark the role of a researcher. Understanding these topics requires deep knowledge and investigation to identify the key documents and entities. Social science experts from history (History teacher, published History scholar), economics (FX trader, accountant, investment banker), and politics (political scientists, politician) help to generate interesting and factually-grounded topics.

The domain experts follow the complex criteria to write 42 topics with minimal post-processing from the annotators to align styles or correct spelling or grammatical errors. There is an equal number of topics per target domain, i.e. 14 History topics, 14 Economics topics, and 14 Politics topics.

Each topic contains a query and narrative. The query is the question the researcher seeks to understand by exploring documents and entities, i.e., the text input posed to the search system. The narratives provide an overview of the topic (key concepts, arguments, facts, etc.) and allow non-domain experts to understand the topic. An example of a topic can be seen in Figure 4.3.



Domain	Economics
Query	How has the UK's Open Banking Regulation benefited challenger banks?
Narrative	UK's Open Banking regulation, which has parallels to the EU's second payment service directive (PSD2), went live in January 2018. This piece of legislation "will require banks to open their payments infrastructure and customer data assets to third parties". As a result, banks no longer have a monopoly on user data if clients grant permission. Challenger banks are small, recently created retail banks that compete directly with the longer-established banks in the UK. Specifically, seeking market share from the "big four" UK retail banks (Barclays, HSBC, Lloyds Banking Group, and NatWest Group). The banks distinguish themselves from the historic banks by modern financial technology practices, such as online-only operations, that avoid the costs and complexities of traditional banking. The largest UK-operating challenger banks include Atom Bank, Revolut, Starling, N26, and Tide. Relevant documents and entities will discuss how challenger banks have used open banking to develop new products or capture market share from traditional retail banks in the UK.

Figure 4.3: Example CODEC topic: economics-1.

Due to the complexity of these topics, the narratives are not completely comprehensive but provide a useful starting point for annotators. They also review pooled runs to assess whether topics are too easy (i.e. lots of highly ranked relevant documents) or do not align with the corpora (i.e. not enough relevant documents or entities to satisfy the information need).

Table 4.14 shows the average number of words and entities in topic queries and narratives. Entity statistics are calculated by running GENRE [26] over the queries and narratives. An average of 12.5 words and 2.4 entities per query supports long natural language queries that include entities. Narratives provide a good proxy for the complexity of the underlying information need, and 143.4 words and 23.7 entities support this complexity.

Document Corpus

The CODEC document corpus aims to have enough high-quality coverage of current social science topics. I initially explore the standard document collections (MS MARCO, TREC Wash-

Table 4.14: Topic Statistics across 42 CODEC topics.

	Total	**Avg. Length**
Query (Words)	524	12.5
Query (Entities)	102	2.4
Narrative (Words)	6,021	143.4
Narrative (Entities)	994	23.7

ington Post, etc.) with CODEC topics but find critical coverage gaps within the required research content. History topics have particularly low coverage and would require augmentation from historical authority sites. This motivates building upon a subset of Common Crawl to create a new focused document corpus for the target domains. Table 4.15 shows the distribution of documents.

Table 4.15: Distribution of Top 15 Websites in Document Corpus.

	Count
reuters.com	172,127
forbes.com	147,399
cnbc.com	100,842
britannica.com	93,484
latimes.com	88,486
usatoday.com	31,803
investopedia.com	21,459
bbc.co.uk	21,414
history.state.gov	9,187
brookings.edu	9,058
ehistory.osu.edu	8,805
history.com	6,749
spartacus-educational.com	3,904
historynet.com	3,811
historyhit.com	3,173
TOTAL	**729,824**

Leverage domain experts, they recommend suitable seed websites or sections of websites. The pool contains a mixture of clearly specialized websites (i.e. economicsdiscussion.net, history.com, brookings.edu) and several general newswire websites (bbc.co.uk, latimes.com, etc.). Social science experts requested up-to-date newswire websites to contextualise current economic and political topics. I also run the topics through a commercial search engine to ensure appropriate coverage and that each domain has enough representation.

The document corpus pipeline takes the focused seed websites and uses Common Crawl and URL pattern matching to extract 300GB of HTML across recent crawls in 2021 (CC-MAIN-2021-[21,17,10,14]). 24 custom BeautifulSoup HTML parsers extract text and metadata while removing any advertising and formatting. This creates documents with fields:

- **id**: Unique identifier is the MD5 hash of URL.
- **url**: Location of the webpage (URL).
- **title**: Title of the webpage if available.
- **contents**: The text content of the webpage after removing any unnecessary advertising or formatting. New lines provide some structure between the extracted sections of the webpage while still being easy for neural systems to process.

I then run multiple filtering stages to ensure the documents are of suitable length and unique. First, the extracted text has to contain at least 30 words, approximately a paragraph. Second, I identify several websites that contain the same (or very similar) webpages hosted on different URLs. Thus, I run a de-duplication step by grouping webpages from the same website that (1) have the same *title* and (2) cosine similarity between document tokens greater than 95%. I solely include the document with the longest *contents* in the final corpus. This removes 96,900 duplicates and results in a final corpus containing 729,824 documents. The corpus is released in jsonlines format.

Entity KB

CODEC uses KILT's [161] Wikipedia KB for the entity ranking task, which is based on the 2019/08/01 Wikipedia snapshot. KILT KB contains 5.9M preprocessed articles that are freely available for use. The entity pages are primarily text-based with minimal structure to indicate headings or passages, i.e. similar to Document Corpus. KILT is selected for CODEC's KB because it aligns with related knowledge-grounded tasks (i.e., fact-checking, open-domain QA, entity linking, etc.). KILT KB also provides inter-entity entity links based on Wikipedia mentions, which could help identify how related entities are to each other.

Entity Linking

I run the REL [77] entity linker, in the default setup described in Chapter 3, over the entire 729,824 document corpus to provide structured connections between documents and entities. For each document, I provide a list of entity links containing fields:

- **mention**: Text span in the document that is linked to an entity.
- **prediction**: Top predicted entity link (Wikipedia title)
- **prediction_kilt**: I map *prediction* entity link to KILT id to align with entity KB and entity judgments.
- **candidates**: Top-k entity link candidates (Wikipedia title).

- **candidates_kilt**: I map *candidates* entity links to KILT ids to align with entity KB and entity judgments.

- **conf_ed**: Score of Flair NER model.

- **score**: Scores of REL candidate selection model.

I release the full 18GB of entity links in jsonlines format. This allows researchers to use entity links within documents to incorporate knowledge grounding and entity ranking. Table 4.16 shows breakdown of the 27.5m entity links (37.7/document) and 144.1m entity candidates (197.5/document).

Table 4.16: Entity Links on Document Corpus.

	Corpus Total	**Document Mean**
Entity Links	27,482,650	37.7
Entity Candidates	144,127,482	197.5

Relevance Criteria

I perform relevance assessment on a graded scale (between 0 and 3) using developed guidelines to ensure a consistent assessment process. Guidelines take inspiration from those of HC4 [97] and are adapted for my tasks.

Document Criteria: The key question for document relevance during annotation is: *How valuable is the most important information in this document?*

- **Very Valuable (3):** The most valuable information in the document would be found in the lead paragraph of a report written on the topic. This includes central topic-specific arguments, evidence, or knowledge. This does not include general definitions or background.

- **Somewhat valuable (2):** The most valuable information in the document would be found in the body of such a report. This includes valuable topic-specific arguments, evidence, or knowledge.

- **Not Valuable (1):** Although on topic, the information contained in the document might only be included in a report footnote or omitted entirely. This consists of definitions or background information.

- **Not Relevant (0):** Not useful or on topic.

Entity Criteria The key question for entity relevance during annotation is: *How valuable is understanding this entity to contextualize document knowledge?*

- **Very Valuable (3):** This entity would be found in the lead paragraph of a report written on the topic. It is absolutely critical to understand this entity to understand this topic.

- **Somewhat valuable (2):** The entity would be found in the body of such a report. It is important to understand this entity to understand this topic.

- **Not Valuable (1):** Although on topic, this entity might only be included in a report footnote or omitted entirely. It is useful to understand this entity to understand this topic.

- **Not Relevant (0):** This entity is not useful or on topic.

Assessment Process

CODEC uses a 2-stage assessment approach to balance adequate coverage of current systems while allowing annotators to explore topics using an iterative search system.

Initial Run Assessment: I generate pools from runs using state-of-the-art sparse and dense retrieval methods. For document runs I use top-100 BM25 [172], BM25 using RM3 expansion [1], ANCE [213], BM25 re-ranked with T5–3B [150], BM25 using RM3 expansion re-ranked with T5–3B, and ANCE re-ranked by T5–3B. I also use a commercial search engine where the top-100 search results are limited to the 24 corpus websites, and the URLs are mapped back to document ids.

For entity runs, I also use a pool of the top-100 results from BM25, BM25 using RM3 expansion, ANCE, BM25 re-ranked with T5–3B, BM25 with RM3 re-ranked with T5–3B, and ANCE re-ranked with T5–3B. I use ELQ [206], an end-to-end entity linking model for questions, to produce an entity run on the queries. GENRE [26], a sequence-to-sequence entity linking model, is also used to produce an entity run using the narrative. I again use a commercial search engine where top-100 search results are limited to Wikipedia and URLs mapped back to document ids.

I devise a weighting ratio for document and entity pooling based on an analysis of several topics across domains. This process takes (1) top-k for each initial system run, then (2) intersection across specified sub-groups, before (3) sampling until the required threshold is reached. The pooling method provides an initial 60 documents and entities for annotators to assess, which provides a reasonable starting point for annotation before the topic exploration stage. Experienced IR annotators (the authors) judge the top 60 documents before doing the same for the top 60 entities. Documents are deliberately judged before entities to provide the annotator with the necessary topic knowledge to assess entity relevance.

Topic Exploration: After the initial runs are assessed, annotators are allotted between two and three hours (although some topics took longer) to use live search systems to explore key dimensions of topics to find relevant documents or entities. Annotators need to construct a minimum of 6 new manual query reformulations. Figure 4.4 shows the query reformulations for the economics-1 topic.

Annotators are encouraged to run these queries through a commercial search engine for spell-checking and to evaluate whether the results are on topic. The live search systems use a

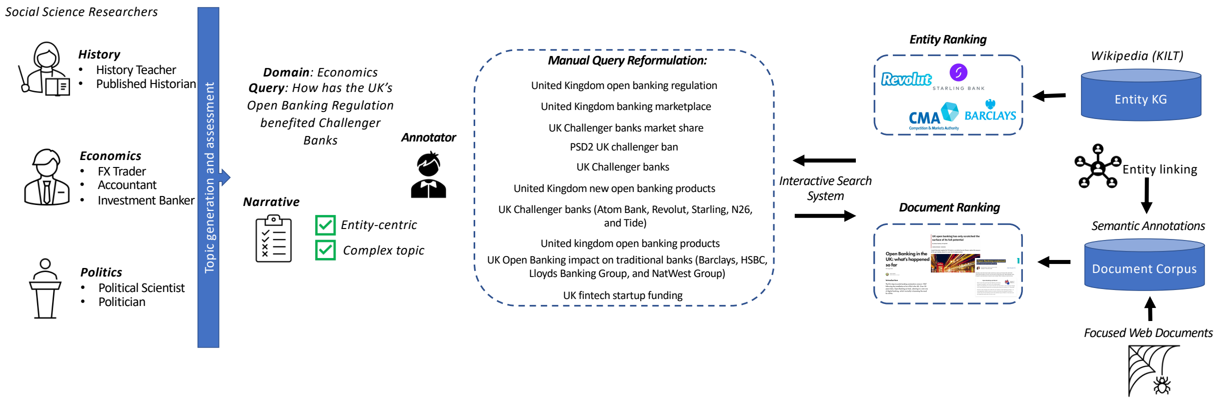

Social Science Researchers
History
History Teacher
Published Historian
Economics
FX Trader
Accountant
Investment Banker
Politics
Political Scientist
Politician
Topic generation and assessment
Domain: Economics
Query: How has the UK's Open Banking Regulation benefited Challenger Banks
Annotator
Narrative
Entity-centric
Complex topic
Manual Query Reformulation:
United Kingdom open banking regulation
United Kingdom banking marketplace
UK Challenger banks market share
PSD2 UK challenger ban
UK Challenger banks
United Kingdom new open banking products
UK Challenger banks (Atom Bank, Revolut, Starling, N26, and Tide)
United kingdom open banking products
UK Open Banking impact on traditional banks (Barclays, HSBC, Lloyds Banking Group, and NatWest Group)
UK fintech startup funding
Interactive Search System
Entity Ranking
Revolut
STARLING BANK
CMA
Competition & Markets Authority
BARCLAYS
Document Ranking
Open Banking in the UK: what's happened so far
Wikipedia (KILT)
Entity KG
Entity linking
Semantic Annotations
Document Corpus
Focused Web Documents

Figure 4.4: CODEC dataset overview.

hybrid BM25, BM25 with RM3 expansion and ANCE for initial retrieval, with re-ranking from T5–3B. This system returns the top 50 documents and top 50 entities to the assessor.

Similar to how a researcher would use commercial search systems to explore a topic iteratively, annotators do not need to assess all returned documents and entities. Annotators are encouraged to scan returned result lists using the title and keyword highlighting to decide whether the document or entity is worth considering before annotating. This process is designed to identify the highly relevant documents and entities not currently returned by baseline systems. Annotators are encouraged to keep searching until they cannot find new relevant documents or entities.

Judgments: Table 4.17 shows the distribution of judgments across the 42 judged topics, which includes 6,186 document judgments (147.3 per topic) and 11,323 entity judgments (269.6 per topic). *Highly Valuable (3)* only makes up 7% of document judgments and 7% of entity judgments. CODEC also releases the manual query reformulations, with the topic exploration phase providing around 74% of overall judgements. There are 387 additionally issued queries overall (9.2 per topic), which can be used to explore query performance prediction or system improvement via query reformulations.

Table 4.17: Judgment distribution across 42 topics.

Judgment	**Document Ranking**	**Entity Ranking**
0	2,353	7,053
1	2,210	2,241
2	1,207	1,252
3	416	777
TOTAL	**6,186**	**11,323**

Evaluation: I provide TREC-style query-relevance files with graded relevance judgments (0-3) for entity and document evaluation. The official measures for both tasks include MAP and Recall@1000 with binary relevance above 1 (i.e. relevance mappings: *0=0.0, 1=0.0, 2=1.0,* *3=1.0*), and NDCG@10 with custom weighted relevance judgments (i.e. relevance mappings: *0=0.0, 1=0.0, 2=1.0, 3=2.0*).

I deliberately gear measures toward the most key documents and entities (i.e. relevance scores of 2 or 3) to prioritise systems ranking these higher vs more tangential but on-topic information (i.e. relevance score of 1). MAP assumes the user wants to find many relevant documents or entities, exposing ranking order throughout the run. On the other hand, NDCG20, with custom scaling to overweight critical information, aim to provide a clear signal of whether systems highly rank the essential documents and entities. Due to recall being important for research-based tasks, Recall@1000 shows missed information.

4.3.2 Research Questions

In these research questions, I assess how effective current systems are on CODEC and analyse areas for improvements. I experiment with simple "ground truth" retrieval methods leveraging entity and text-based expansion methods. These questions are intended to show the headroom and motivate the research directions for future query expansion models. I follow the experimental setup outlined in the Methodology section.

- **RQ3.1: Is CODEC datasets complex?** This research question looks to independently verify that CODEC meets my complex criteria.

- **RQ3.2: How do current systems perform on complex topics?** This research question explores the performance of baseline retrieval and re-ranking systems on CODEC, highlighting any headroom for improvement and conducting query analysis.

- **RQ3.3: How does ground-truth entity-based query expansion help?** I develop an entity-based query expansion for document retrieval by leveraging the entity judgments to explore the effectiveness on complex queries.

- **RQ3.4: How does ground-truth text-based query expansion help?** I use the query reformations as golden text-based expansion content to evaluate the effectiveness on CODEC.

4.3.3 RQ3.1: Is CODEC dataset complex?

Table 4.3.3 shows aggregated annotation for the 42 queries present in TREC Deep Learning 2019. This shows the four complex criterion (i.e., Multifaceted, Entities, Comprehension, Knowledge) and the overall complex classification, and the number of and percentage of queries that fall within each classification (i.e., does not meet, partially meets, strongly meets)

These results show that all the CODEC queries meet the complex criteria, and almost every individual criterion is strongly met. This is not necessarily surprising, given I developed the query set with these requirements in mind. Nonetheless, these findings support the fact that CODEC queries are complex and suitable as a core dataset for this thesis.

Table 4.18: Aggregated annotation of complex criteria on TREC Deep Learning 2019. Number of queries and percentage of query set within each criterion and annotation class.

	Multifaceted	**Entities**	**Comprehension**	**Knowledge**	**Complex**
Does not meet (0)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Partially meets (1)	1 (2.4%)	4 (9.5%)	1 (2.4%)	0 (0.0%)	0 (0.0%)
Strongly meets (2)	**41 (97.6%)**	**38 (90.5%)**	**41 (97.6%)**	**42 (100.0%)**	**42 (100.0%)**

4.3.4 RQ3.2: How do current systems perform on complex topics?

Table 4.19 shows the effectiveness of current sparse, dense, entity-centric and PRF initial retrieval methods on CODEC. These results largely align with those of TREC Robust, including that dense retrieval methods have considerably worse recall effectiveness than sparse and entity-centric methods. For example, ColBERT-TCT with PRF has 8% lower R@1000 compared to a tuned BM25 with RM3 expansion. In fact, no method significantly outperforms BM25 with RM3 in recall on these datasets. Interestingly, on CODEC queries, the entity-centric RM3 expansion method (ENT+RM3) has the best R@1000, highlighting the usefulness of entities, particularly within long queries with multiple entities mentioned. This aligns with similar results on Robust description queries.

Table 4.19: The effectiveness of initial retrieval methods and then combined with PLM re-ranking (T5-3b). For first-pass retrieval significance testing is against BM25+RM3. For PLM re-ranking it is BM25+RM3 ⇒ PLM. "$^+$" significantly better and "$^-$" significantly worse, *bold* depicts best system.

		CODEC		
		NDCG@20	MAP	R@1000
Retrieval	TCT	0.357	0.211	0.740$^-$
	TCT+PRF	0.357	0.246	0.754$^-$
	ENT	0.310	0.209	0.809
	ENT+RM3	0.326	0.232	0.833
	BM25	0.317	0.214$^-$	0.783$^-$
	BM25+RM3	0.330	0.239	0.816
Re-Ranking	TCT ⇒ PLM	0.482	0.348$^-$	0.740$^-$
	TCT+PRF ⇒ PLM	0.478	0.351$^-$	0.754$^-$
	ENT ⇒ PLM	0.490	0.372	0.809
	ENT+RM3 ⇒ PLM	0.490	0.373	**0.833**
	BM25 ⇒ PLM	**0.498**	0.372	0.783$^-$
	BM25+RM3 ⇒ PLM	0.497	**0.377**	0.816

I see large absolute improvements in NDCG@20 when combined with PLM re-ranking. Specifically, increasing on average by +0.15 across my first pass retrieval using CODEC natural language queries. This highlights that PLM re-ranking models offer an opportunity to improve the precision in early ranks, which is the basis for effective query expansion methods.

Overall, these results show that current sparse and entity-centric expansion methods are still more effective than newer dense or learned sparse. In addition, the gains from PLM re-ranking passages provide significant gains in effectiveness, particularly in early ranks. These findings mirror those of baseline systems on TREC Robust 2004.

4.3.5 RQ3.3: How does simple entity-based query expansion help?

CODEC allows researchers to explore the role of entities in document ranking using the provided document judgment, entity judgments, and entity links that connect documents and entities. To understand the relationship between document and entity relevance, I take the 6,186 document judgments and map the relevance of entities mentioned in each document using the entity links and entity judgments. I assume entities without judgments are *Not Relevant*. I analyse both top-1 predicted entity and top-k candidate entities for document ranking.

Figure 4.5 shows the Pearson Correlation Coefficient between document relevance and the percentage of entities in the document grouped by relevance, i.e. *Not Relevant*, *Not Valuable*, *Somewhat Valuable*,*Very Valuable*. Both predicted entity (+0.19) and candidate entities (+0.22) support that documents with higher proportions of *Very Valuable* entities are positively correlated with document relevance.

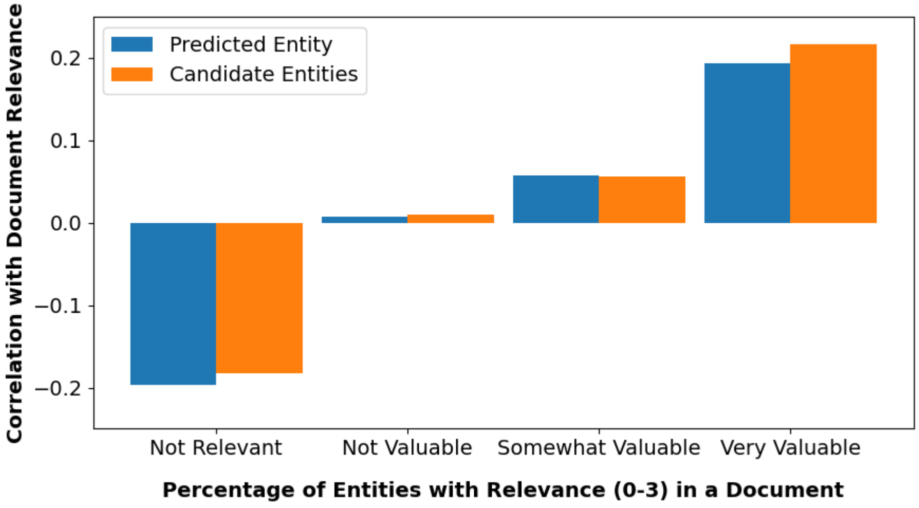


Figure 4.5: Correlation of document and entity relevance on CODEC.

I develop an entity expansion method to build on these findings. Prior work has shown enriching queries with entity-based information improves ad hoc ranking performance [42]. **Entity-QE** is an oracle entity expansion method for document retrieval that enriches the query with names of relevant entities taken from entity judgments. I use BM25 for initial retrieval, removing stopwords, using Porter stemming, and using CODEC tuned BM25 fold parameters.

With CODEC standard four-folds, I cross-validate the (1) weighting of original query terms, (2) weighting of *Very Valuable* entity terms, and (3) weighting of *Somewhat Valuable* terms. Across the four folds, Entity-QE term weighting are: original queries average 9.2 terms with 80% weighting, with *Very Valuable* entities adding 42.6 terms on average with 16% weighting, and *Somewhat Valuable* entities adding 77.8 terms on average with 4% weighting. These weightings highlight how the most useful entities are weighted higher.

Table 4.20: Entity-QE Document ranking. *Bold* indicates best system and ($^{\triangle}$) indicates 5% paired-t-test significance against BM25+RM3.

	NDCG@10	**MAP**	**Recall@1000**
BM25+RM3	0.327	0.239	0.816
Entity-QE	**0.405**$^{\triangle}$	**0.287**$^{\triangle}$	**0.857**$^{\triangle}$

Table 4.20 shows Entity-QE improves Recall@1000 to 0.857, which is statistically significant when compared to the best initial retrieval systems BM25+RM3. I also see significant improvements on both NDCG@10 and MAP. These findings support that entity-centric ranking methods benefit complex topics. This motivates this research direction which I explore in further in Chapter 5 with my Latent Entity Expansion (LEE) model. Furthermore, CODEC having aligned document and entity judgments enables new classes of entity-centric ranking models to be developed and evaluated.

4.3.6 RQ3.3: How does ground-truth text-based query expansion help?

In this research question, I study the utility of the manually reformulated queries used by the experts, which I treat as ground-truth expansion text. I show that the best manual reformulation outperforms the original query on document and entity ranking. I also develop a query expansion method that uses all query reformulations that improve over the strong baselines.

Best Reformulation vs Original Query

I use a tuned BM25 and RM3 expansion model to analyse the performance of query reformulations against the original query, as this is a strong system across document and entity ranking. Figure 4.6 shows the distribution of the best query reformulation against the original query across document and entity ranking for MAP, NDCG@10, and Recall@1000.

The best query reformulation improves Recall@1000 of document ranking to 0.845, a statistically significant difference. As is depicted in the boxplot, the best reformulation leads to almost 75% of topics having Recall@1000 over 0.80. However, the best query reformulation has a smaller relative improvement on Recall@1000 for entity ranking (0.712) and is not statistically significant. This suggests that several query reformulations are required for a robust initial entity ranking.

Analysing history-6 topic, *What were the lasting social changes brought about by the Black Death?*, the original query performs poorly with a Recall@1000 of 0.428 on document ranking. However, seven of thirteen query reformulations have Recall@1000 between 0.810 and 0.905, i.e *The Black Death (Bubonic Plague) and end of Feudalism*, *The Black Death (Bubonic Plague) and the Renaissance*, and *Bubonic Plague / Black Death and Roman Catholic Church*. The

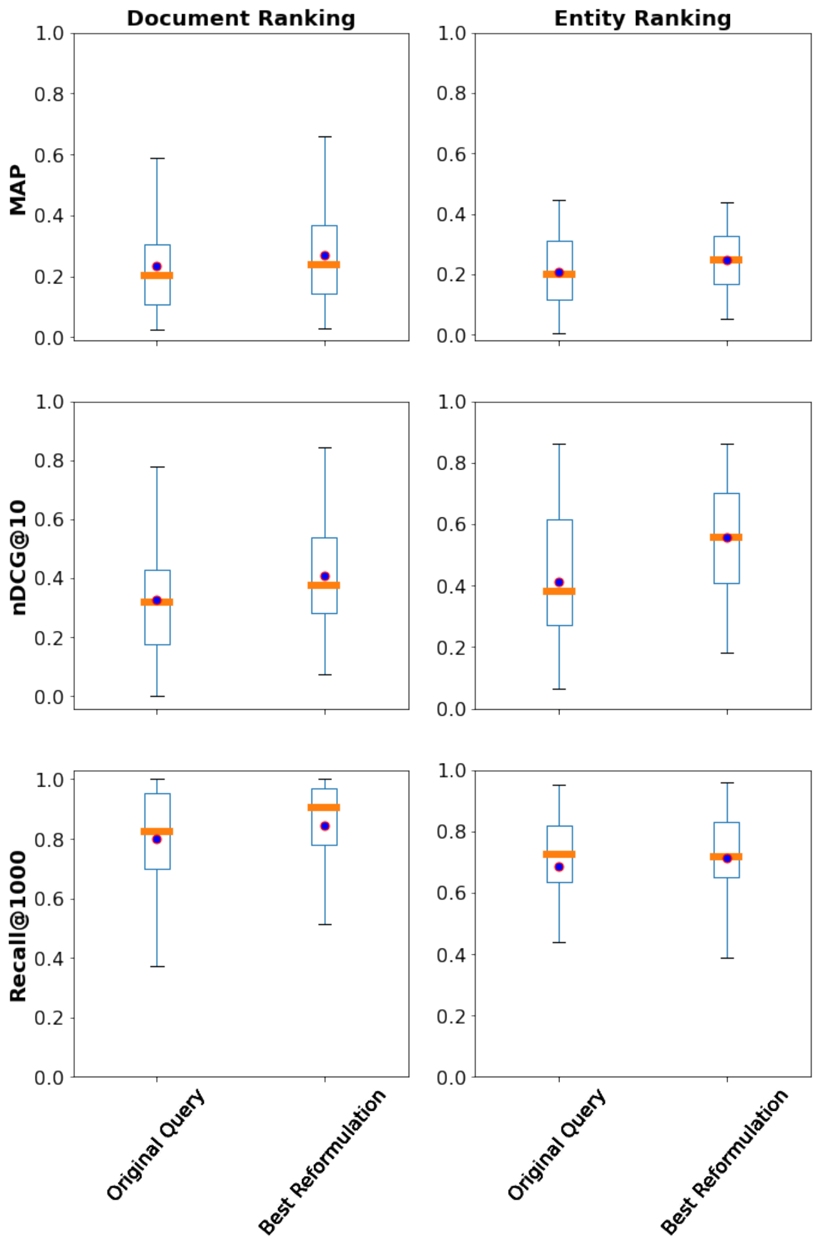


Figure 4.6: Boxplot BM25+RM3 topic performance of (1) original query, and (2) best manual query reformulation. Blue dot indicates means and orange line median across topics.

researcher is iterating on entity names and synonym expansion to identify missing documents and entities.

The best reformulation significantly improves document and entity ranking compared to the original query on MAP and NDCG@10 measures. Document ranking MAP improves from 0.233 to 0.270, and NDCG@10 from 0.327 to 0.407. NDCG@10 saw the largest relative improvement, with around 75% topics having an NDCG@20 over 0.3 (i.e. proportionally fewer failing topics than the original query). Similarly, the best query reformulation significantly improves entity ranking, with MAP improving from 0.209 to 0.248 and NDCG@10 from 0.412 to 0.557.

Entity ranking NDCG@10 saw a 35% improvement due to the best query reformulation, the largest relative improvement of any measure across either task. Analysing the runs, this was driven by query reformulations accessing specific clusters of highly relevant entities within the top ranks. For example, the original query for topic history-15, *Why did Winston Churchill lose the 1945 General Election after winning World War II?*, had an NDCG@10 of 0.323. The best query reformulation, *Appeasement and Great Depression cost Conservatives in 1945 General Election*, improves NDCG@10 to 0.609. The improvement of top-ranked entities is due to the introduction of key events (i.e. [Appeasement] and [Great Depression]) and entities (i.e. [Conservatives]) being part of the query reformulation.

Query expansion using query reformulations

I develop a query expansion method **Reform-QE**, which uses both the original and query reformulation terms. I use BM25 in a similar setup to Entity-QE, cross-validating the weighting of the original query terms against the weighting of the aggregate query reformulation terms. The original queries average 9.2 terms, and the aggregate query reformulation averages 42.8 terms. For document ranking, the original queries average 66.7% weighting and aggregate query reformulation averages 33.3% weighting across the folds. For entity ranking, the original queries average 60% weighting and aggregate query reformulation averages 40% weighting across the folds.

Table 4.21 depicts document ranking results for the query expansion methods that use the query reformulations. Reform-QE significantly improves over the best initial retrieval system, BM25+RM3, achieving Recall@1000 of 0.864. Reform-QE also has the statistically improvement in NDCG@10 and MAP when compared with BM25+RM3. This highlights how strong text-based expansion using text covering different dimensions of the topic is highly effective, more so than any individual query.

Table 4.21: Reform-QE Document ranking. *Bold* indicates best system and ($^{\triangle}$) indicates 5% paired-t-test significance against BM25+RM3+T5.

	NDCG@10	MAP	Recall@1000
BM25+RM3	0.327	0.239	0.816
Reform-QE	**0.384**	**0.275**	**0.864**$^{\triangle}$

Table 4.22 shows entity ranking results of the query expansion methods that use the query reformulations. Reform-QE significantly outperforms the best entity system, BM25+RM3, across MAP, NDCG@10, and Recall@1000. There is a larger relative improvement when using query reformulations compared to document ranking, highlighting how several queries are needed to expose the full range of relevant entities.

Overall, using query reformulations, text content covering different aspects of the topic offers systems a chance to explore complex topics and access information not explicitly expressed

Table 4.22: Reform-QE Entity ranking. *Bold* indicates best system and ($^{\triangle}$) indicates 5% paired-t-test significance against BM25+RM3.

	NDCG@10	**MAP**	**Recall@1000**
BM25+RM3	0.412	0.209	0.685
Reform-QE	**0.525**$^{\triangle}$	**0.253**$^{\triangle}$	**0.738**$^{\triangle}$

in the query. Based on these findings, I develop novel PLM-based expansion methods in Chapter 6 where I use PLMs to generate text content based on different facets of the topic.

4.3.7 Conclusion

In this section, I introduce CODEC, a document and entity ranking resource that focuses on complex research topics. Social science researchers produce 42 topics spanning history, economics, and politics. To support open research, I create a new semantically annotated and focused collection derived from subsets of the Common Crawl. CODEC is grounded to the KILT's Wikipedia knowledge base for entity linking and retrieval. I provide 17,509 document and entity judgments (416.9 per topic) by assessing the pooled initial runs and manual exploration of the topics using interactive search systems, adding 387 manual query reformulations (9.2 per topic). I also verify CODEC is a complex dataset based on my criteria, finding that 100% of queries are complex.

CODEC system analysis demonstrates topics are challenging for state-of-the-art traditional models and neural rankers. Failures demonstrate encoding entities and relationships is challenging for both document and entity ranking. Specifically, queries with large amounts of latent knowledge, where new expansion techniques are a promising research direction. I find that document relevance is positively correlated with the occurrence of relevant entities. I leverage this relationship with ground-truth entity query expansion method that outperforms strong baseline systems on document ranking. I also demonstrate that query reformulation can play an important role in accessing latent dimensions within complex topics. Both individual query reformulations and aggregated reformulations improve document and entity ranking.

Overall, this resource represents an important step toward developing and evaluating entity-centric search models on complex topics. This dataset complements TREC Robust 2004 as the core datasets for this thesis, allowing me to develop and evaluate new query expansion models using PLMs. I also developed several threads of entity-centric and facet-based expansion that we'll build on during later chapters.

4.4 Chapter Conclusion

In this chapter, I focus on analysing current and developing new datasets focused on complex queries that we'll use to create and evaluate my new query expansion methods. I start by building an understanding of what makes queries "complex" based on system effectiveness analysis and detailed annotations of the TREC Deep Learning [33,34]. Based on this analysis, I establish my complex query criteria to classify topics to the scope of this thesis. Specifically, these topics are open-ended with multiple facets, concerning many central entities, and require reasoning and research to understand. I show that the current dataset, TREC Robust 2004 [198], and I develop a new dataset, CODEC [131], that contains complex queries that can be leveraged to build new query expansions models leveraging PLMs.

In my resource, DL-HARD [130], I augment TREC Deep Learning [33, 34] datasets with extensive query annotations to understand what makes queries challenging. For example, these annotations include question intent categories, answer types, wikified entities, and topic categories. This analysis shows that current ranking models are highly effective QA-style and short factoid-based questions. However, I identify substantial room for improvement on queries that contain multiple facets, require deep research and reasoning, and would elicit a long-form response. For example, a query such as "can fever cause miscarriage in early pregnancy?". This analysis, as part of DL-HARD, supports the development of my complex query criteria.

I develop my complex query criterion and annotation process to ensure that this thesis focuses on the category of queries that can most benefit from improved contextualisation from new query expansion pipelines leveraging PLMs. Specifically, I include four criterion: (1) *Multifaceted*: open-ended with multiple facets, (2) *Entities*: requires understanding across multiple central entities and concepts, (3) *Comprehension*: complicated and requires reasoning across the information, and (4) *Knowledge*: needs significant research to understand fully. I also outline a detailed annotation process to assess whether topics are complex, partially complex, or not complex.

I use my complex criteria to demonstrate that TREC Deep Learning 2019, 2020, and DL-HARD contain few complex or partially complex queries. However, I show that TREC Robust 2004 contains many complex and partially complex queries and is thus a suitable dataset for this thesis. I also run baseline systems to show that current sparse and entity-centric expansion methods are more effective at initial retrieval than newer dense or learned sparse methods. I also show the effectiveness of PLM re-ranking, particularly in improving precision in top ranks.

Nonetheless, I need a second complex dataset for this thesis, and I prioritise building such a dataset from the ground up to support the development and evaluation of new query expansion models. Complex Document and Entity Collection (CODEC) focuses on complex information needs from social science domains, such as history, finance, and politics. This dataset includes essay-style queries, golden "narratives" encapsulating information needs, query facets, and dense annotations of relevant documents and entities. The resource builds complex topics, verified using my complex criteria.

My analysis of systems on CODEC highlights the difficulty that both state-of-the-art traditional models and neural rankers face when handling complex queries. The failures I observed point to challenges in encoding entities and their relationships, affecting both document and entity ranking performance. This is especially evident in queries containing significant amounts of latent knowledge, where new expansion techniques offer a promising area for future research. I found a positive correlation between document relevance and the presence of relevant entities. By leveraging this relationship, I introduced a ground-truth entity query expansion method that surpasses strong baseline systems in document ranking tasks. Additionally, I demonstrate that query reformulation plays a key role in uncovering latent dimensions within complex topics. I expand on these initial directions in Chapter 5, where I use PLM-focused feedback and entity representations, and in Chapter 6, where I generate text content for expansion, including using query sub-topics.

Overall, in this chapter, I finalise the complex datasets that will be used to develop and evaluate new query expansion models. Specifically, TREC Robust 2004 and my new dataset, CODEC, will serve as the primary datasets for this thesis. These datasets predominantly contain queries with multiple facets, which require contextualisation from multiple entities and concepts and deep research and reasoning to understand. Additionally, I have begun developing several research threads that will be expanded in subsequent chapters, including using PLMs for precise feedback, entity-centric expansion models, and rich text-based and subtopic-driven context for expansion models.

Chapter 5

Query Expansion with PLM Ranking

In the previous chapter, I show that sparse pseudo-relevance feedback (PRF) systems have the potential to overcome some limitations in the recall of complex topics. However, all PRF models suffer from the same problem: if the initial query is challenging, the candidate set is unlikely to contain relevant documents in the top ranks, which will cause PRF models to fail. In this chapter, I use PLM pipelines to precisely rank the most relevant text context and build a joint entity and term expansion model to improve retrieval effectiveness on complex queries. This research supports my hypothesis that new PLM capabilities, such as precise ranking, can be incorporated into query expansion pipelines to achieve state-of-the-art results.

I tackle the problem of search over complex queries using three complementary techniques. First, I demonstrate that applying a strong neural re-ranker before sparse or dense PRF can improve the retrieval effectiveness by 5–8%. Second, I propose an enhanced expansion model, Latent Entity Expansion (LEE), which applies fine-grained word and entity-based relevance modelling incorporating localized features. Specifically, I find that by including both words and entities for expansion achieve a further 2–8% improvement in NDCG. My analysis also demonstrates that LEE is largely robust to its parameters across datasets and performs well on entity-centric queries. And third, I include an “adaptive” component in the retrieval process, which iteratively refines the re-ranking pool during scoring using the expansion model and avoids re-ranking additional documents. I find that this combination of techniques achieves the best NDCG, MAP and R@1000 results on the TREC Robust 2004 and CODEC document datasets.

In this chapter, I split this complementary research into two sections: Adaptive Expansion and Latent Entity Expansion. In Section 5.1, I define Adaptive Expansion and show the retrieval benefits on current PRF systems by using PLM re-ranking before query expansion. Then, in Section 5.2, I define and show the effectiveness benefits of Latent Entity Expansion, with and without adaptive expansion.

5.1 Adaptive Expansion

A fundamental problem in information retrieval is query-document lexical mismatch [18]. A common approach to address this issue is pseudo-relevance feedback, where a first-pass top-*k* candidate set of documents is retrieved, and these feedback signals can augment the query for a second-pass retrieval. Early work on PRF focused on term-based query expansion [1, 95, 136, 137], with later work showing entity-based representations can offer improvements on the hardest topics [42]. Recently, this PRF paradigm has also leveraged dense vectors [145, 201, 219]. However, all these models suffer from the same problem: if the initial query is complex, the candidate set is unlikely to contain relevant documents in the top ranks, which will cause PRF models to fail.

Meanwhile, PLMs for re-ranking [111] have led to significant advances in effectiveness, particularly precision in the top ranks. I observe this in my baseline experiments on CODEC and Robust (Section 4), where NDCG@20 increases by 25-50% with re-ranking. In this work, I pull together these research threads on neural re-ranking and query expansion methods to improve the core task of document retrieval. Figure 5.1 shows how I address the problem of poor pseudo-relevance feedback by applying re-ranking prior to query expansion and re-executing this query. I find that expansion with PLM feedback improves the recall-oriented effectiveness of sparse and dense PRF approaches.

Nonetheless, after my expansion with neural feedback, I find that a second round of neural re-ranking is required to maximize precision. Thus, I draw inspiration from recent adaptive re-ranking work [124] and propose my "Adaptive Expansion" framework. Specifically, Figure 5.1 shows how I dynamically refine the re-ranking pool during scoring using the expansion model. This allows me to use PLM feedback for expansion and re-ranking in a single pass and reduces the number of documents scored by around 35%. Thus, this new query expansion paradigm leverages PLM capabilities to precisely rank within minimal additional computational overhead.

I summarise my contributions below:

- I show that by changing the traditional expansion pipelines to have a PLM re-ranker prior to expansion, increases recall of sparse and dense expansion methods by 5–8%. Thus, demonstrating that PLM-focused feedback for expansion models can improve effectiveness on complex queries.

- I propose Adaptive Expansion, which iteratively refines the re-ranking pool during scoring using the expansion model and avoids re-ranking additional documents (saving around 35% compute based on documents to re-rank). Therefore improving retrieval effectiveness of complex topics without signifiant increase in computational requirements.

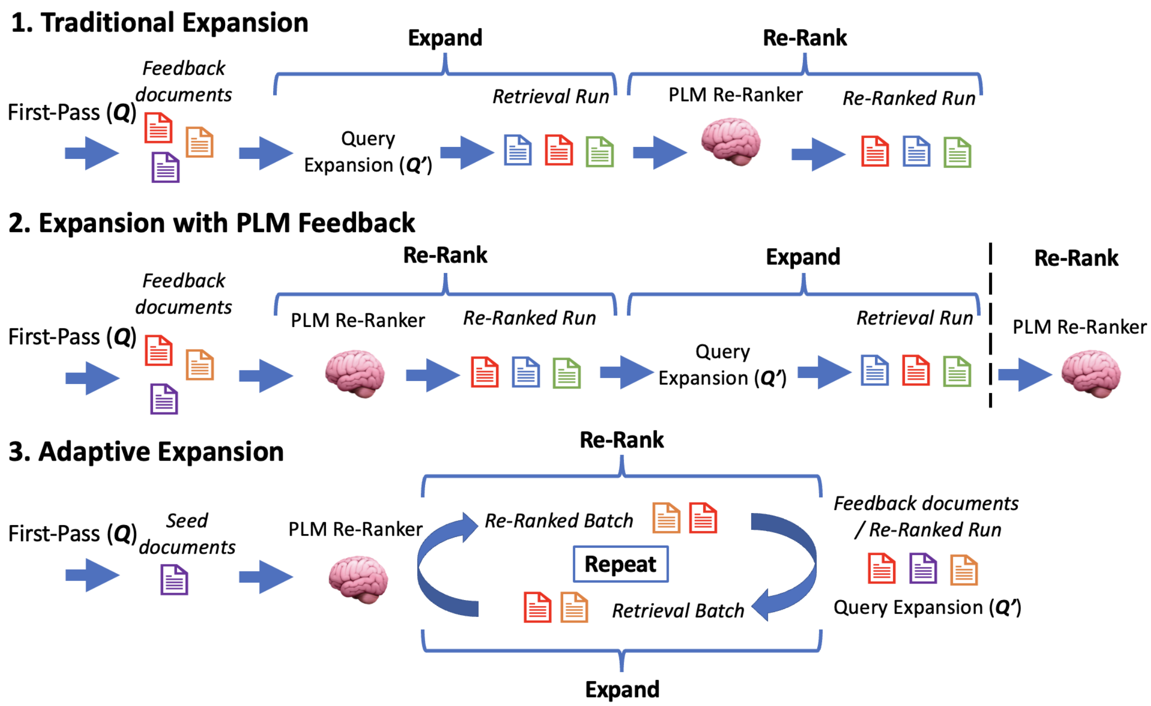


Figure 5.1: Rethinking query expansion pipelines leveraging PLM feedback. *1) Traditional Expansion:* standard first-pass retrieval and expand approach, often with neural re-ranking after. *2) Expansion with PLM Feedback:* a re-ranked run provides more accurate feedback for expansion. *3) Adaptive Expansion:* iteratively re-rank batches of documents before issuing expanded query to retrieve the next batch.

5.1.1 Methodology

Based on the analysis of current retrieval models, I rethink the standard query expansion pipeline drawing on several research threads. Specifically, PLM re-ranking models [39, 149, 150] offer an opportunity to improve the precision of document feedback to form more effective expansion models. I also draw from recent work on adaptive re-ranking [124] to allow my expansion model to use PLM feedback without incurring additional re-ranking cost.

As defined in Chapter 3, given an information need (query) q, I want to return a ranked list of documents $R_k^D = [d_1, d_2, ..., d_k]$ relevant to the query q from a collection D. For generality, documents, d, may also refer to other retrieval units, such as passages. I abstract a document ranking pipeline, and focus on changing the ordering of *query expansion* and *neural re-ranking* components. Figure 5.1 shows the three expansion pipelines I explore:

- **Traditional expansion**: The standard document ranking pipeline with expansion [58, 104, 126]. Specifically, retrieving an initial set of documents using PRF retrieval models [1, 145, 201], before using a neural re-ranker to create a final re-ranked list of documents. The issue with this pipeline is that signals from advanced neural re-rankers are not used to improve initial recall.

- **Expansion with PLM feedback**: I move PLM re-ranking before the expansion model in the pipeline to improve the precision of the feedback set; thus, improving expansion effectiveness. Additionally, a second re-ranking pass could further improve the precision; however, this would also increase computational expense due to extra document scoring.

- **Adaptive expansion**: Instead of having a static run that I re-rank, I propose dynamically updating my document frontier as more documents are scored using my query expansion model. Specifically, I alternate my re-ranking of documents between the initial retrieval documents and the dynamic frontier based on the expansion model, which uses all currently re-ranked documents as its feedback set. This iterative batch process of re-ranking and expansion continues, with a batch size, until I reach my intended number of documents. Intuitively, updating my query expansion model as more documents are scored is similar to a manual researcher building their understanding of a topic through reading information. Additionally, unlike expansion with PLM feedback with a second re-ranking pass, adaptive expansion does not require additional computation from document re-ranking. Algorithm 1 shows the adaptive re-ranking framework that we extend, replacing the query-independent corpus graph (G) with our novel query expansion models that leverage PLM feedback.

5.1.2 Experimental Setup

I follow the same experimental setup as [124] to allow a fair comparison with GAR adaptive re-ranking baselines. Specifically, I can take an initial BM25 run (R_0) and use a batch size b of 16 to alternate between the initial BM25 run and LEE retrieval with a total re-ranking budget of 1,000 documents. I use the same experimental setup for the adaptive re-ranking experiments as part of my LEE research in Section 5.2. Moreover, I outline in detail all my various query expansion pipeline implementations in Chapter 3.

5.1.3 Research Questions

In these research questions, I assess the benefits of using a PLM before query expansion to rank the more relevant feedback content. I go on to explore the benefits of adaptive expansion. Specifically:

- **RQ4.1: RQ4.1: Does re-ranking before query expansion improve retrieval effectiveness?** I explore sparse and dense PRF methods with PLM re-ranking before the query expansion stage. This research question explores recall gains with enhanced precision in the feedback.

- **RQ4.2: Does adaptive expansion have effectiveness gains without the same computation overhead?** I explore adaptive expansions' efficiency and effectiveness gains on sparse and dense methods. I compare to adaptive GAR [125] baselines and benchmark the efficient gains based on number of documents re-ranked.

5.1.4 RQ4.1: Does re-ranking prior to query expansion improve retrieval effectiveness?

Table 5.7 shows current state-of-the-art neural and traditional models (with a second re-ranking phase [150]) on the TREC Robust 2004. This includes CEDR [126] and PARADE [104], which are strong systems fine-tuned on Robust04. I also compare my system to comparable neural pseudo-relevance feedback techniques that leverage multiple rounds of re-ranking. I include the CEQE, which used CEDR's initial re-ranked run. I also implement a ColBERT-TCT-PRF and RM3 expansion run on top of a PLM (T5-3b) run, with all systems' PLM re-ranked for a fair comparison. I conduct significance testing against BM25 with RM3 expansion PLM re-ranking system.

Table 5.1: Expansion with PLM feedback and second-pass re-ranking; "+" significant improvement over BM25 ⇒ RM3 ⇒ PLM.

		Robust04 - Title			Robust04 - Description		
		NDCG	MAP	R@1000	NDCG	MAP	R@1000
1x Re-Rank	SPLADE [55] ⇒ PLM	0.539	0.309	0.597	0.590	0.357	0.617
	EQFE [42]	0.601	0.328	0.806	-	-	-
	BM25 ⇒ CEQE [145] ⇒ PLM	0.626	0.373	0.764	-	-	-
	BM25 ⇒ RM3 ⇒ CEDR [126]	0.632	0.370	0.776	0.645	0.400	0.758
	BM25 ⇒ RM3 ⇒ PARADE [104]	0.642	0.380	0.776	0.650	0.408	0.758
	ENT ⇒ RM3 ⇒ PLM	0.615	0.366	0.745	0.658	0.407	0.759
	TCT ⇒ PRF ⇒ PLM	0.584	0.345	0.681	0.572	0.364	0.619
	BM25 ⇒ RM3 ⇒ PLM	0.634	0.377	0.777	0.652	0.406	0.750
2x Re-Rank	CEDR ⇒ CEQE [145] ⇒ PLM	0.644	0.384	0.787	-	-	-
	TCT ⇒ PLM ⇒ PRF ⇒ PLM	0.592	0.349	0.697	0.630	0.390	0.702
	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	**0.656**$^{+}$	**0.390**$^{+}$	**0.813**$^{+}$	**0.674**$^{+}$	**0.416**$^{+}$	**0.780**$^{+}$

These results show that PLM re-ranking before query expansion significantly improves recall on RM3 expansion on TREC Robust, including a +5.6% relative increase in title queries and a +4.0% relative increase in description queries. With a second-pass PLM re-ranking, there's significant gains in NDCG and MAP on both title and description queries. Furthermore, there are effectiveness gain on ColBERT-TCT-PRF with two passes of PLM re-ranking, with the largest increases on description queries where re-ranking effectiveness is strongest.

Table 5.8 shows the PLM query expansion pipelines on the CODEC dataset. Similar to the prior table, I include ColBERT-TCT-PRF and RM3 expansion with PLM re-ranking before and after expansion models. These results show significant improvement in sparse recall from PLM-focused expansion, with a +6.0% relative improvement in R@1000. Similarly, dense expansion improves by +7.0% over traditional expansion pipelines. I also observe marginal improvement in NDCG and MAP across both dense and sparse expansion.

Table 5.2: Expansion with PLM feedback and second-pass re-ranking; "+" significant improvement over BM25 ⇒ RM3 ⇒ PLM.

		CODEC		
		NDCG	MAP	R@1000
1x Re-Rank	ENT ⇒ RM3 ⇒ PLM	0.490	0.373	0.833
	TCT ⇒ PRF ⇒ PLM	0.606	0.351	0.754
	BM25 ⇒ RM3 ⇒ PLM	0.644	0.377	0.816
2x Re-Rank	TCT ⇒ PLM ⇒ PRF ⇒ PLM	0.636	0.369	0.808
	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	**0.659**	**0.379**	**0.865**$^{+}$

These results highlight how large gains in effectiveness can be achieved with PLM feedback across standard sparse and dense PRF retrieval models. Moreover, a second pass neural re-ranker over the initial retrieval runs further improves NDCG and MAP. This leads to R@1000 being significantly improved compared to the state-of-the-art baseline, with NDCG and MAP significantly better on Robust04 titles and descriptions. These results show the potential for using PLMs to rank text content as part of the query expansion pipeline for complex queries.

5.1.5 RQ4.2: Does adaptive expansion have effectiveness gains without the same computation overhead?

In this research question, I explore adaptive expansion to improve effectiveness without a second pass re-ranking.

What are the effectiveness benefits of adaptive expansion?

Table 5.3 shows RM3 adaptive expansion on TREC Robust 2004 against the traditional RM3 expansion pipeline, RM3 with two PLM passes (before and after expansion), and the adaptive "GAR" systems [124]. I conduct significance testing against BM25 with RM3 expansion with PLM re-ranking.

Table 5.3: Adaptive re-ranking effectiveness ("⇔"), with significance testing ("+") against BM25 ⇒ RM3 ⇒ PLM.

		Robust04 - Title			Robust04 - Description		
		NDCG	MAP	R@1000	NDCG	MAP	R@1000
2x Re-Rank	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	**0.656**$^{+}$	**0.390**$^{+}$	**0.813**$^{+}$	0.674^{+}	0.416^{+}	0.780^{+}
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.634	0.377	0.777	0.652	0.406	0.750
1x Re-Rank (Adaptive)	BM25 ⇒ GAR-BM25 ⇔ PLM	0.629	0.372	0.768	0.652	0.402	0.747
	BM25 ⇒ GAR-ColBERT ⇔ PLM	0.630	0.374	0.769	0.649	0.402	0.739
	BM25 ⇒ GAR-ENT ⇔ PLM	0.637	0.377	0.781	0.661	0.408	0.758
	BM25 ⇒ RM3 ⇔ PLM	0.655^{+}	0.387^{+}	**0.813**$^{+}$	**0.675**$^{+}$	**0.418**$^{+}$	**0.783**$^{+}$

I find that GAR-based methods that use words (GAR-BM25), entities (GAR-ENT), and dense representation (GAR-ColBERT) are not significantly better than a standard PLM re-ranking pipeline. Given GAR's strong effectiveness in passage ranking, these results suggest that document ranking over long documents requires relevance modelling across multiple documents.

However, my adaptive expansion pipeline with RM3 is consistently more effective than all GAR systems, being significantly better on Robust04 over the traditional PLM re-ranking pipeline. This shows that using adaptive re-ranking, with a re-ranking budget of 1000 documents, can increase R@1000 by 4.4-4.6%, MAP by 2.7-3.0%, and NDCG by 3.3-3.5%. In fact, these results almost precisely align to the pipeline with PLM re-ranking before and after expansion, without requiring additional document re-ranking.

Table 5.4 shows RM3 adaptive expansion on CODEC against the traditional RM3 expansion pipeline and other baseline systems, conducting significance testing against BM25 with RM3 expansion with PLM re-ranking. Much like the TREC Robust results, the GAR baselines do not show meaningful improvement over my traditional expansion pipeline. However, I observe meaningful but not significant improvement in recall with adaptive expansion with RM3, increasing R@1000 by a relative +3.8%. The adaptive expansion pipeline, with one pass of re-ranking, is only marginally less effective than two passes of PLM re-ranking.

Table 5.4: Adaptive re-ranking effectiveness (“⇔”), with significance testing (“+”) against BM25 ⇒ RM3 ⇒ PLM.

		CODEC		
		NDCG	MAP	R@1000
2x Re-Rank	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	**0.659**	**0.379**	**0.865**$^+$
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.644	0.377	0.816
1x Re-Rank (Adaptive)	BM25 ⇒ GAR-BM25 ⇔ PLM	0.634	0.362	0.797
	BM25 ⇒ GAR-ColBERT ⇔ PLM	0.645	0.368	0.822
	BM25 ⇒ GAR-ENT ⇔ PLM	0.644	0.366	0.821
	BM25 ⇒ RM3 ⇔ PLM	0.653	0.373	0.847

What are the computational benefits of adaptive expansion?

For simplicity, I measure computational expense by the number of documents that require PLM re-ranking, which should be a strong proxy across implementations and hardware. Thus, the computations benefits of adaptive expansion are due to the document set differences between the initial run (i.e., BM25) and the RM3 with PLM feedback (i.e., BM25 ⇒ PLM ⇒ RM3). For example, on Robust04 titles, two passes of PLM results in 1,503 unique documents being scored per query, compared to only 1,000 for adaptive re-ranking (i.e., saves 33% scoring cost). I find similar trends in Robust04 descriptions (637 fewer documents to re-score) and CODEC (525 fewer documents to re-score). Overall, this shows my adaptive expansion framework can significantly improve recall-oriented effectiveness without requiring a second pass of neural re-ranking. This is important given the latency constraints of production search systems and energy consumption concerns of modern AI [188].

5.1.6 Conclusion

In the section, I show that PLMs can be integrated as part of expansion pipelines to improve the precision of feedback documents, and thus, improve expansion models. Specifically, I find that by adding a PLM re-ranking stage before sparse and dense expansion increases recall by 5–8% on complex queries. These are significant effectiveness gains through incorporating PLMs; however, two PLM re-ranking stages are required to maximise precision.

As a solution to the additional re-ranking requirements, I propose a new expansion paradigm, "Adaptive Expansion", that dynamically refines the re-ranking pool during scoring using an expansion model. This allows me to use PLM feedback for expansion and re-ranking in a single pass and reduces the number of documents scored by around 35%. I show that adaptive expansion leads to significant effectiveness gains on complex topics without two re-ranking stages, thus greatly reducing computational cost. In the next section, I will build on these findings by developing new query expansion models that are specifically designed to leverage PLM-focused feedback.

5.2 Latent Entity Expansion

Armed with the effectiveness gains I observe through putting a PLM re-ranker before expansion, and leveraging the hypothesis that entities play a crucial role in complex topics, I propose a new expansion model operating over PLM feedback: Latent Entity Expansion (LEE). LEE is a joint probabilistic term and entity-based expansion model. In contrast with prior work in Latent Concept Expansion [137], I show that a hybrid expansion model with terms and entities is more effective than comparable individual expansion models. I also demonstrate improved effectiveness from passage feedback based on PLM re-ranking that provide a more fine-grained hybrid relevance model. Furthermore, unlike prior work [137], I find that using dependencies from entity co-occurrence improves effectiveness with passage feedback but can be harmful with document feedback.

Through extensive experiments under various conditions, to my knowledge, LEE produces the highest recall ever achieved on TREC Robust 2004 and CODEC datasets by 6-12%. Query analysis shows that LEE's hybrid expansion model with terms and entities improves the hardest entity-centric queries, where a fine-grained relevance model and entity dependencies are particularly useful. Furthermore, LEE with adaptive expansion sets new state-of-the-art results for MAP and NDCG without requiring a second round of neural re-ranking, and I show that hyperparameters are robust across datasets. Overall, this work demonstrates the potential of probabilistic term-entity expansion models when combined with PLM re-ranking to the improve effectiveness of complex queries.

I summarise my contributions below:

- I provide a detailed study of existing probabilistic word and entity expansion models with document and passage feedback from neural re-ranking. These results show the different expansion model setups when using PLM feedback on complex topics and motivate my proposed hybrid model.

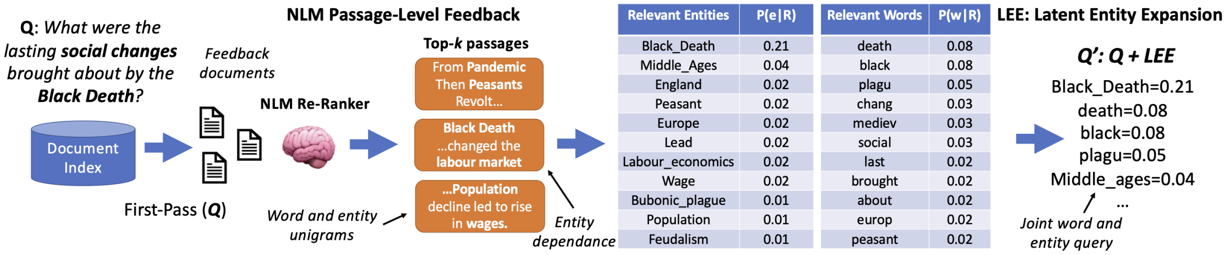
- I propose a new hybrid relevance model, LEE, for query expansion that incorporates entity dependencies. This expansion model is designed to improve retrieval effectiveness on complex topics leveraging PLM feedback.

- I show the LEE expansion model is state-of-the-art by 6-12% on recall, and when combined with additional neural re-ranking, results in 2-8% improvement on NDCG and MAP. These results highlight the meaningful contribution to improving retrieval effectiveness on complex topics.

- I show that LEE with adaptive expansion achieves similar effectiveness to additional PLM re-ranking without required two passes of PLM re-ranking. These results show we can improve retrieval effectiveness on complex topics without additional computation cost.

5.2.1 Methodology

Figure 5.2 depicts my Latent Entity Expansion model that incorporates words and entities and is designed to improve effectiveness of complex, knowledge-heavy queries. Specifically, this query expansion approach uses a strong PLM re-ranked list of documents (or passages), which benefits precision in the top ranks (making my feedback more accurate). Thus, I assume top-k documents to be query-relevant feedback R, which I use to construct Latent Entity Expansion (LEE) based on a hybrid relevance model of words ($\{w_1, w_2, ..., w_i\} \in d$) and entities ($\{e_1, e_2, ..., e_N\} \in d$). I use LEE to expand the initial, $q^o \rightarrow q^{LEE}$, and retrieve my final set of documents, $[d_1, d_2, ..., d_k]$. It should be highlighted that I use "documents" in the methodology, however, this generalises to "passages", which I find is more effective in my experiments.


Q: What were the lasting social changes brought about by the Black Death?
Document Index
First-Pass (Q)
Feedback documents
NLM Passage-Level Feedback
NLM Re-Ranker
Top-k passages
From Pandemic Then Peasants Revolt...
Black Death ...changed the labour market
...Population decline led to rise in wages.
Word and entity unigrams
Entity dependance
LEE: Latent Entity Expansion
Q': Q + LEE
Black_Death=0.21
death=0.08
black=0.08
plagu=0.05
Middle_ages=0.04
...
Joint word and entity query

Relevant Entities	P(e|R)
Black_Death	0.21
Middle_Ages	0.04
England	0.02
Peasant	0.02
Europe	0.02
Lead	0.02
Labour_economics	0.02
Wage	0.02
Bubonic_plague	0.01
Population	0.01
Feudalism	0.01

Relevant Words	P(w|R)
death	0.08
black	0.08
plagu	0.05
chang	0.03
mediev	0.03
social	0.03
last	0.02
brought	0.02
about	0.02
europ	0.02
peasant	0.02

Figure 5.2: Overview of LEE: A PLM re-ranker supplies a strong passage ranking to provide fine-grained feedback. LEE uses word and entity unigrams and entity dependencies to construct a hybrid word and entity-based probability distribution.

Deriving Expansion Words

Equation 5.1 show how I estimate the probability of a word $P(w|R)$ given the assumed relevant documents R. $P(q|d)$ is obtained by normalizing the PLM re-ranking scores [150], before I turn it into a probability by dividing the sum of all the normalized scores, $\sum_{d' \in R} P(q|d')$. The probability of a word given a document, $P(w|d)$, is the term frequency divided by the document length. Following LCE [137], I normalize the distribution using $P(w|D)$ (that I approximate for convenience with $\text{IDF}(w, D)$). Later, I show that this feature is important for modelling the relevance of documents.

$$P(w|R) = \sum_{d \in R} P(q|d)P(w|d)P(w|C) \tag{5.1}$$

Deriving Expansion Entities

Analogously, I estimate the query-relevance of a document based on the entities contained within that document ($e \in d$). Prior work [137] only uses unigram representations because word dependencies do not improve results. In contrast, LEE incorporates both entity unigrams and dependencies and finds meaningful improvement in effectiveness with passage PLM feedback. The base formulation for entity terms follows how I model word unigrams, providing a unigram estimate of $P(e|R)$. However, I also include entity dependence terms based on co-occurrence to model the relationship between entities.

Estimating relevance of entity dependence, I estimate this probability as follows:

$$P([e_1, e_2]|R) = \sum_{d \in R} P(q|d)P([e_1, e_2]|d)P([e_1, e_2]|D) \tag{5.2}$$

where $P(q|d)$ is the normalised PLM score and $P([e_1, e_2]|d)$ is the sum of both entity frequencies divided by the document length. I approximate $P([e_1, e_2]|C)$ as the product of entity IDFs, $\text{IDF}(e_1) \cdot \text{IDF}(e_2)$. Incorporating entity co-occurrence increases the weighting of entities that co-occur with many entities in relevant documents. This helps prioritise the "central entities" that are particularly useful for identifying relevant documents for complex topics. Unlike [137], results show this entity dependence feature is particularly beneficial with passage feedback, although not meaningful at a document level.

I then combine the unigram and entity dependence models as follows. Mathematically, I model $P(e|R)$ as follows:

$$P(e|R) = \beta \sum_{e_i \in R} P([e, e_i]|R) + (1 - \beta)P_{\text{unigram}}(e|R) \tag{5.3}$$

where $P([e, e_i]|R)$ is the probability of the entity pair (e, e_i) being in a relevant document, and $P_{\text{unigram}}(e|R)$ is probability of entity e, obtained using a unigram language model.

LEE Duet Representation

The final score of a document $d \in R$ is derived from an interpolation of the term-based and entity-based query expansion retrieval scores:

$$\text{Score}(d,q) = \lambda \cdot \text{Score}_{\text{word}}(d,q) + (1-\lambda) \cdot \text{Score}_{\text{entity}}(d,q) \tag{5.4}$$

where $\text{Score}_{\text{word}}(d,q)$ is the document score based on my word query expansion model and $\text{Score}_{\text{entity}}(d,q)$ is the document score based on my entity query expansion model. For simplicity to execute over large collections, I use BM25 [172] to execute my probabilistic queries over separate document and entity indexes. Furthermore, following work by RM3 [1], I also include the probabilistic interpolations between the terms in the original query and my probability distribution. I normalise these scores and interpolate using $\lambda \in [0,1]$. In practice, $\lambda = 0.5$ is reasonable across my complex datasets, which I show in my results section later.

Adaptive Expansion with LEE

I formalise adaptive expansion with LEE following my proposed expansion pipeline in Section 5.1. Specifically, given the original query q^o and the current re-ranked documents, D_{PLM}, I produce my duet representation, q^{LEE}, to retrieve the next batch of unscored, b-sized documents to be re-score. Thus, as more documents are scored, and the D_{PLM} set increases in size, my word and entity-based probabilistic query is updated and becomes more representative.

5.2.2 Experimental Setup

Retrieval and Expansion To avoid query drift, all LEE runs in the paper use the tuned BM25 system based on the input initial run [172]. I tune LEE hyperparameters using a grid search and cross-validation to optimise R@1000. Specifically, for term and entity system weighting, I tune feedback passages (i.e., *fb_docs*: 10 to 100 with a step size of 10), the number of feedback terms (i.e., *fb_terms*: 10 to 100 with a step size of 10), the interpolation between the original terms and expansion terms (i.e., *original_query_weight*: 0.1 to 0.9 with a step of 0.1). For the entity component, I tune the co-occurrence weighting (β: 0.1 to 0.9 with a step of 0.1), and lastly, the hybrid weighting between word and entity (λ: 0.1 to 0.9 with a step of 0.1) and the run depth (k_{LEE}: 1000, 2000, 3000, and 4000). All hyperparameters are released for reproducibility: *link*. Furthermore, I outline all pipeline implementations in Chapter 3.

Adaptive Expansion I follow the same experimental setup as Section 5.1 for fair comparison to baseline adaptive expansion methods, only changing the LEE expansion model.

5.2.3 Research Questions

These research questions focus on how PLMs integrated into query expansion pipelines can develop new query expansion models that utilise this precise PLM-focused feedback. My experimentation assesses the effectiveness gains of my proposed Latent Entity Expansion method. I build this formulation from the ground up, benchmarking entities vs words and passages vs documents. I then compare LEE to strong standalone and adaptive expiation baselines. My research questions are as follows:

- **RQ4.3: What are the optimal expansion models with PLM feedback?** I explore sparse PRF methods with PLM re-ranking before the query expansion stage and vary the feedback unit (passage and documents), expansion model (RM3, LCE, and LEE), and vocabulary (words, entities, and hybrid). This research question explores recall gains associated with these methods on complex topics.
- **RQ4.4: How does LEE effectiveness compare to baseline systems?** I compare LEE to strong sparse and dense expansion baselines with PLM re-ranking before and after query expansion. This research question assesses whether our method has significantly improved effectiveness on complex topics.
- **RQ4.5: Can LEE and adaptive expansion be combined?** I explore LEE's efficiency and effectiveness gains when combined with adaptive expansion. Given LEE is based on re-ranked documents, this shows the upper bound of effectiveness when I only re-rank 1,000 documents. In essence, can we gain the desired effectiveness on complex topics without two phases of PLM re-ranking.

5.2.4 RQ4.3: What are the optimal expansion models with PLM feedback?

Table 5.5 and Table 5.6 compare the effectiveness of expansion models on top of a PLM re-ranked run, building upon the most effective sparse approaches identified previously. Specifically, comparing BM25 with RM3 expansion and neural re-ranking to my expansion models with PLM [150] feedback. I vary the expansion models (RM3, LCE), the unit of feedback (documents and passages), and vocabulary (words and entities). My proposed LEE hybrid model that combines word and entity vocabularies is the last two rows in the table.

RQ4.3a: Are passages or documents more effective for neural expansion?

Across both datasets, I see the relative improvement of passages (rows without ^{D}) to particularly improve NDCG (i.e., Robust04 titles +1.8%, descriptions +2.4%, and CODEC +7.2%) and MAP

Table 5.5: Query expansion varying expansion (e.g. RM3, LCE, and LEE), PLM feedback (e.g. documents (^{D}) or passages), and vocabulary (e.g. "Entity" or words). Significance testing against BM25 ⇒ RM3 ⇒ PLM; significantly better ("$^{+}$") and worse ("$^{-}$").

		Robust04 - Title			Robust04 - Description		
		NDCG	MAP	R@1000	NDCG	MAP	R@1000
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.634	**0.377**	0.777	0.652	**0.406**	0.750
	BM25 ⇒ PLM ⇒ RM3-Entity^{D}	0.600^{-}	0.322^{-}	0.779	0.619^{-}	0.343^{-}	0.781^{+}
	BM25 ⇒ PLM ⇒ RM3-Entity	0.612^{-}	0.331^{-}	0.776	0.643	0.364^{-}	0.792^{+}
	BM25 ⇒ PLM ⇒ RM3^{D}	0.630	0.350^{-}	0.813^{+}	0.616^{-}	0.334^{-}	0.780^{+}
	BM25 ⇒ PLM ⇒ RM3	0.638	0.353^{-}	0.812^{+}	0.625^{-}	0.339^{-}	0.797^{+}
	BM25 ⇒ PLM ⇒ LCE-Entity^{D}	0.614^{-}	0.335^{-}	0.797	0.640	0.360^{-}	0.806^{+}
	BM25 ⇒ PLM ⇒ LCE-Entity	0.626	0.343^{-}	0.793	0.659	0.377^{-}	0.810^{+}
	BM25 ⇒ PLM ⇒ LCE^{D}	0.636	0.353^{-}	0.824^{+}	0.659	0.375^{-}	0.829^{+}
	BM25 ⇒ PLM ⇒ LCE	0.647	0.360^{-}	0.825^{+}	0.668	0.377^{-}	0.843^{+}
	BM25 ⇒ PLM ⇒ LEE^{D}	0.648	0.366	0.834^{+}	0.673^{+}	0.388	0.845^{+}
	BM25 ⇒ PLM ⇒ LEE	**0.660**$^{+}$	0.376	**0.837**$^{+}$	**0.687**$^{+}$	0.401	**0.855**$^{+}$

Table 5.6: Query expansion varying expansion (e.g. RM3, LCE, and LEE), PLM feedback (e.g. documents (^{D}) or passages), and vocabulary (e.g. "Entity" or words). Significance testing against BM25 ⇒ RM3 ⇒ PLM; significantly better ("$^{+}$") and worse ("$^{-}$").

		CODEC		
		NDCG	MAP	R@1000
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.644	**0.377**	0.816
	BM25 ⇒ PLM ⇒ RM3-Entity^{D}	0.590^{-}	0.292^{-}	0.851^{+}
	BM25 ⇒ PLM ⇒ RM3-Entity	0.645	0.331^{-}	0.854^{+}
	BM25 ⇒ PLM ⇒ RM3^{D}	0.615	0.312^{-}	0.865^{+}
	BM25 ⇒ PLM ⇒ RM3	0.641	0.335	0.874^{+}
	BM25 ⇒ PLM ⇒ LCE-Entity^{D}	0.578^{-}	0.283^{-}	0.849
	BM25 ⇒ PLM ⇒ LCE-Entity	0.643	0.325^{-}	0.857^{+}
	BM25 ⇒ PLM ⇒ LCE^{D}	0.606^{-}	0.313^{-}	0.872^{+}
	BM25 ⇒ PLM ⇒ LCE	0.632	0.326^{-}	0.877^{+}
	BM25 ⇒ PLM ⇒ LEE^{D}	0.619	0.321^{-}	0.879^{+}
	BM25 ⇒ PLM ⇒ LEE	**0.663**	0.357	**0.883**$^{+}$

(i.e., Robust04 titles +2.0%, descriptions +3.2%, and CODEC +10.2%), with less relative improvement at R@1000. This shows that passages with PLM scoring provide a more fine-grained relevance signal for my query expansion, potentially reducing noise from long documents with less relevant passages. I note that CODEC contains many long, domain-specific documents with over 450 non-stop words on average (roughly 50% more than Robust04), and passages have a larger relative improvement.

Looking at specific instances, passage feedback is vital for the CODEC query "What were the lasting social changes brought about by the Black Death?", where using LEE query with PLM passage feedback increases NDCG from 0.483 to 0.736. Both passage and document feedback methods correctly identify the central entity as [Black_Death], which is contained within 91% of judged relevant documents. However, the document-level feedback only has general topical entities within the higher ranks, such as [Pandemic], [Infection], and [Bubonic_plague]. On the other hand, passage-level relevance signals can identify other important entities, such as

[Peasant] (i.e., [Black_Death] leads to [Peasant]s' revolts), and entity co-occurrences such as [Population] and [Wage], (i.e., [Black_Death] leads to changes in workers [Wage] due to [Population] decline). Thus, highlighting the importance of passage feedback for localized features.

RQ4.3b: How does LEE's word-entity expansion compare with existing expansion approaches?

Across all datasets, LEE has the best R@1000 and NDCG of any expansion method. Specifically, it has significantly better R@1000 compared to the PLM re-ranking baseline, with between 7.5-13.7% relative improvement. NDCG significantly improves on Robust04 titles and descriptions and shows relative improvements on CODEC of 2.9% (although not significant). LEE is the only query expansion technique where MAP across all datasets is not significantly worse than the base neural re-ranking pipeline. To the best of my knowledge, LEE results are the best reported R@1000 results across all datasets. It highlights the strong recall-oriented effectiveness of word-entity hybrid models that build on neural passage re-ranking. Further, it shows the results don't hurt and sometimes help the precision; thus, additional re-ranking is not necessarily required.

I conduct a query-by-query analysis to understand why LEE has such significant improvements in R@1000. Focusing on Robust04 and comparing to BM25, I find that LEE helps 166 and hurts 33 title queries, compared to RM3, which helps 139 and hurts 47 queries. These findings are even more evident in description queries, where LEE helps 181 and hurts 30 queries, compared to RM3, which helps 156 and hurts 45 queries. This supports that LEE query expansion that leverages a combination of both words and entities is more robust than simply using words alone.

Analysing specific queries, I see that joint modelling of words and entities can be beneficial for retrieving relevant documents, when one probability distribution fails. For example, when term-based approaches fail, such as the Robust04 description query, "What impact has the Chunnel had on the British economy and/or the life style of the British?", where BM25 and BM25 with RM3 expansions both have an R@1000 of 0.061. This failure is driven by vocabulary miss match, where "chunnel" is a less common colloquialism for "Channel Tunnel". However, LEE achieves R@1000 of 0.862 through strong relevance signals from re-ranked passages due to avoiding vocabulary miss match by weighting [Channel_Tunnel] as the most relevant entity, thus capturing multiple lexical variations, i.e., "Chunnel", "Channel Tunnel", "Eurostar", or "Eurotunnel".

Figure 6.3 shows that the largest relative gains are on the hardest queries (ordered based on original BM25 retrieval effectiveness on Robust description queries). Specifically, LEE improves R@1000 of the hardest 5% of queries by around 0.6 compared to BM25 and 0.55 compared to BM25 with RM3 expansion. Furthermore, substantial gains over RM3 are also observed in 5-25% (+0.2) and 25-50% (+0.1) buckets. This highlights that using top-*k* passages and joint modelling of terms and entities effectively improves the hardest queries, with minimal drop in effectiveness on the easy queries (i.e., only slightly reducing effectiveness on 75%-95% band).

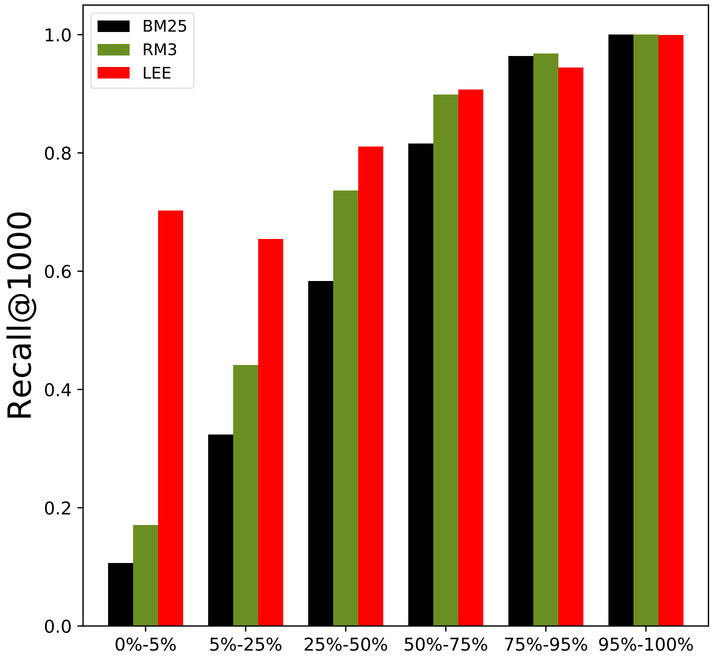


Figure 5.3: Query difficulty plot stratified by original BM25 score on Robust04 descriptions with RM3 and PLM ⇒ LEE.

RQ4.3c: Does entity dependencies help the query expansion model?

These results also show that entity-based and hybrid expansion models benefit more from passage feedback. On average, across the datasets, such models see around 100% greater improvement on NDCG and 150% on MAP from using passages relative to word-based representations. Furthermore, I find that fine-grained passage signals are important for leveraging entity information, especially when using dependencies to infer relationships between entities. I find that including the entity co-occurrences improves effectiveness versus simply modelling entities based on unigrams; they provide consistent improvements across the datasets increasing MAP by 3.3% on average, NDCG 0.9% and R@1000 0.4%, with no system being negatively affected. However, I find that entity co-occurrence at a document level is less effective, with MAP reducing on average by 0.2%, with small gains in NDCG and R@100 of 0.3%, and Robust04 systems being negatively affected.

Overall, LEE improves recall-oriented effectiveness by leveraging PLM re-ranked passages to infer a strong hydrid word and entity relevance model. I show that it is more effective than either RM3 and LCE on words and entities individually, and that entity dependencies help with passage feedback. My unsupervised LEE queries are significantly better across all datasets on

R@1000, either substantially improving or not significantly hurting MAP and NDCG. These are the best-reported R@1000 results across these datasets and highlights hybrid word-entity expansion models in combination with PLM passage ranking can improve the effectiveness on complex queries.

5.2.5 RQ4.4: How does LEE effectiveness compare to baseline systems?

This research question addresses how the LEE expansion model with passage feedback compares to sparse and dense systems with an additional round of neural re-ranking. Specifically, Table 5.7 shows LEE (with a second re-ranking phase [150]) compared to current state-of-the-art neural and traditional models on the target datasets. This includes CEDR [126] and PARADE [104], which are strong systems fine-tuned on Robust04. Table 5.8 shows similar experiments on CODEC dataset.

Table 5.7: Expansion with PLM feedback and second-pass re-ranking; "+" significant improvement over BM25 ⇒ RM3 ⇒ PLM.

		Robust04 - Title			Robust04 - Description		
		NDCG	MAP	R@1000	NDCG	MAP	R@1000
1x Re-Rank	EQFE [42]	0.601	0.328	0.806	-	-	-
	BM25 ⇒ RM3 ⇒ PARADE [104]	0.642	0.380	0.776	0.650	0.408	0.758
	TCT ⇒ PRF ⇒ PLM	0.584	0.345	0.681	0.572	0.364	0.619
	BM25 ⇒ RM3 ⇒ PLM	0.634	0.377	0.777	0.652	0.406	0.750
	BM25 ⇒ PLM ⇒ LEE	0.660$^+$	0.376	**0.837**$^+$	0.687$^+$	0.401	**0.855**$^+$
2x Re-Rank	CEDR ⇒ CEQE [145] ⇒ PLM	0.644	0.384	0.787	-	-	-
	TCT ⇒ PLM ⇒ PRF ⇒ PLM	0.592	0.349	0.697	0.630	0.390	0.702
	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	0.656$^+$	0.390$^+$	0.813$^+$	0.674$^+$	0.416$^+$	0.780$^+$
	BM25 ⇒ PLM ⇒ LEE ⇒ PLM	**0.667**$^+$	**0.393**$^+$	**0.837**$^+$	**0.715**$^+$	**0.438**$^+$	**0.855**$^+$

Table 5.8: Expansion with PLM feedback and second-pass re-ranking; "+" significant improvement over BM25 ⇒ RM3 ⇒ PLM.

		CODEC		
		NDCG	MAP	R@1000
1x Re-Rank	TCT ⇒ PRF ⇒ PLM	0.606	0.351	0.754
	BM25 ⇒ RM3 ⇒ PLM	0.644	0.377	0.816
	BM25 ⇒ PLM ⇒ LEE	0.663	0.357	**0.883**$^+$
2x Re-Rank	TCT ⇒ PLM ⇒ PRF ⇒ PLM	0.636	0.369	0.808
	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	0.659	0.379	0.865$^+$
	BM25 ⇒ PLM ⇒ LEE ⇒ PLM	**0.664**$^+$	**0.380**	**0.883**$^+$

I also compare my systems to comparable neural pseudo-relevance feedback techniques that leverage multiple rounds of re-ranking. I include the CEQE, which used CEDR's initial re-ranked run, and I also implement a ColBERT-TCT-PRF and RM3 expansion runs on top of a PLM run, with all systems PLM re-ranked for a fair comparison. I conduct significance testing against BM25 with RM3 expansion PLM re-ranking system.

Although neural re-ranking improves MAP and NDCG results, it is worth highlighting how competitive the LEE unsupervised expansion method (without re-ranking) is when compared to prior work. Specifically, neural re-ranking only increases NDCG between 0.002-0.028 and MAP 0.018-0.037 across the datasets. This improvement from PLM re-ranking is roughly five times less from neural re-ranking when compared to a BM25 with RM3 expansion run. For example, using the unsupervised LEE query on Robust04 titles, I achieve NDCG@10 of 0.561, which is higher than reported SPLADE [55] results 0.485 (a comparable unsupervised method), better than T5-3b [150] 0.545, and approaching fine-tune PARADE [104] 0.591.

Robust04 description queries show the largest relative improvement from a second-pass re-ranking; but even this dataset shows 88 queries reduce effectiveness in MAP, with 160 queries improving. An example of where neural re-ranking helps Robust04 description topic 433, “Is there contemporary interest in the Greek philosophy of stoicism?”. Interestingly, R@1000 of BM25 is 0.308, and traditional RM3 expansion completely fails with R@1000 of 0.000, while LEE improves R@1000 to 0.615, by expanding with useful words [vlasto, philosophy, socrates, greek, stoicism, contemporary] and entities [Gregory_Vlastos, Socrates, Greek_language, Vlastos, Philosophy, Stoicism]. However, without neural re-ranking, MAP is only 0.017 compared to 0.626 with re-ranking.

Query analysis shows LEE outperforms on entity-centric queries, where the topical focus is a specific concept or named entity, where dense models struggle with these query types [179]. For example, Robust04 query 376, “World Court” where the user refers to “International Court of Justice”, ColBERT-TCT and ColBERT-TCT-PRF systems only achieve R@1000 of 0.137 and 0.147. Sparse methods also do not perform well, with R@1000 for BM25 of 0.225 and RM3 expansions of 0.235. The sparse models struggle as “world” and “court” are common words with many meanings and instances. However, LEE can use entity mentions to infer the specific instance of the “world court” and model the probability of [International_Court_of_Justice] entity the highest, increasing R@1000 to 0.735. I also find LEE without re-ranking performs better, with a MAP of 0.338 versus 0.250 when re-ranked again, which highlights the benefits of explicit entity modelling with LEE.

LEE’s effectiveness within entity-centric information needs is not surprising when I analyse the implicit entity ranking from LEE using CODEC’s entity judgements. I compare this to the best baseline systems provided with CODEC, where systems score entity relevance based on the relevance score of Wikipedia pages representing each entity [161]. I can see that LEE is very effective in the top ranks, achieving NDCG@3 of 0.767 and NDCG@10 of 0.554 (much higher than all dataset baselines). This highlights the accuracy of the LEE entity model and explains the improvement of queries requiring explicit entity modelling.

Overall, these results highlight the potential of hybrid word-entity expansion models with PLM feedback. I show significant improvements in recall on complex queries compared to state-of-the-art baselines. Furthermore, I show how a second pass re-ranking can further improve LEE’s precision and that the hybrid word-entity modelling helps entity-centric queries. Nonetheless, doing a second-pass PLM scoring increases the computational re-ranking expense, which I seek to address in the following research question by combining with adaptive expansion.

5.2.6 RQ4.5: Can LEE and adaptive expansion be combined?

In this research question, I explore combining LEE with adaptive expansion to improve effectiveness without a second pass re-ranking. Table 5.9 and Table 5.10 shows adaptive LEE expansion against LEE with two PLM passes, the adaptive "GAR" systems [124], and an adaptive RM3 expansion system for comparison. I conduct significance testing against BM25 with RM3 expansion with PLM re-ranking on Robust and CODEC datasets.

Table 5.9: Adaptive re-ranking effectiveness ("⇔"), with significance testing ("+") against BM25 ⇒ RM3 ⇒ PLM.

		Robust04 - Title			Robust04 - Description		
		NDCG	MAP	R@1000	NDCG	MAP	R@1000
2x Re-Rank	BM25 ⇒ PLM ⇒ RM3 ⇒ PLM	0.656$^+$	0.390$^+$	0.813$^+$	0.674$^+$	0.416$^+$	0.780$^+$
	BM25 ⇒ PLM ⇒ LEE ⇒ PLM	0.667$^+$	**0.393**$^+$	0.837$^+$	**0.715**$^+$	**0.438**$^+$	**0.855**$^+$
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.634	0.377	0.777	0.652	0.406	0.750
	BM25 ⇒ PLM ⇒ LEE	0.660$^+$	0.376	0.837$^+$	0.687$^+$	0.401	**0.855**$^+$
1x Re-Rank (Adaptive)	BM25 ⇒ RM3 ⇔ PLM	0.655$^+$	0.387$^+$	0.813$^+$	0.675$^+$	0.418$^+$	0.783$^+$
	BM25 ⇒ LEE ⇔ PLM	**0.668**$^+$	0.392$^+$	**0.838**$^+$	0.704$^+$	0.435$^+$	0.834$^+$

Table 5.10: Adaptive re-ranking effectiveness ("⇔"), with significance testing ("+") against BM25 ⇒ RM3 ⇒ PLM.

		CODEC		
		NDCG	MAP	R@1000
2x Re-Rank	BM25 ⇒ NLM ⇒ RM3 ⇒ NLM	0.659	0.379	0.865$^+$
	BM25 ⇒ PLM ⇒ LEE ⇒ PLM	0.664$^+$	0.380	0.883$^+$
1x Re-Rank	BM25 ⇒ RM3 ⇒ PLM	0.644	0.377	0.816
	BM25 ⇒ PLM ⇒ LEE	0.663	0.357	0.883$^+$
1x Re-Rank (Adaptive)	BM25 ⇒ RM3 ⇔ PLM	0.653	0.373	0.847
	BM25 ⇒ LEE ⇔ PLM	**0.669**$^+$	**0.382**	**0.887**$^+$

RQ4.5a: Is adaptive expansion with LEE the most effective adaptive re-ranking method?

These results support LEE with adaptive expansion as the most effective adaptive method across all datasets and measures. The significance testing aligns with two re-ranking phases, i.e., being significantly better on Robust04 across all measures and CODEC on R@1000 and NDCG. In fact, LEE with adaptive expansion is nominally better than two re-ranking passes on CODEC across all measures and Robust04 titles on NDCG and R@1000. Analysis shows that iteratively sampling different documents after each batch can lead to a better language model.

Although, R@1000 reduces on Robust04 descriptions from 0.853 to 0.834 when comparing adaptive re-ranking and two PLM passes. On inspection, this reduction in recall effectiveness is driven by alternating batches between BM25 and LEE following [124] algorithm. Because

LEE is much more relatively effective than BM25, there is a reduction in overall effectiveness. Nonetheless, adaptive re-ranking does improve NDCG by 2.5% and MAP by 6.6% over any state-of-the-art methods on description queries, and future work could vary how I sample from these different pools.

Lastly, Figure 5.4 shows the LEE query effectiveness as I progress along the adaptive batches on Robust04, where each batch is 16 scored documents. For comparison, I plot my RM3 adaptive expansion baseline. What is evident is that LEE query effectiveness is consistently higher throughout the adaptive expansion process, i.e., I only need a few documents to outperform RM3. Additionally, long description queries on Robust04 provide a strong re-ranking signal to keep improving my query expansion throughout the run.

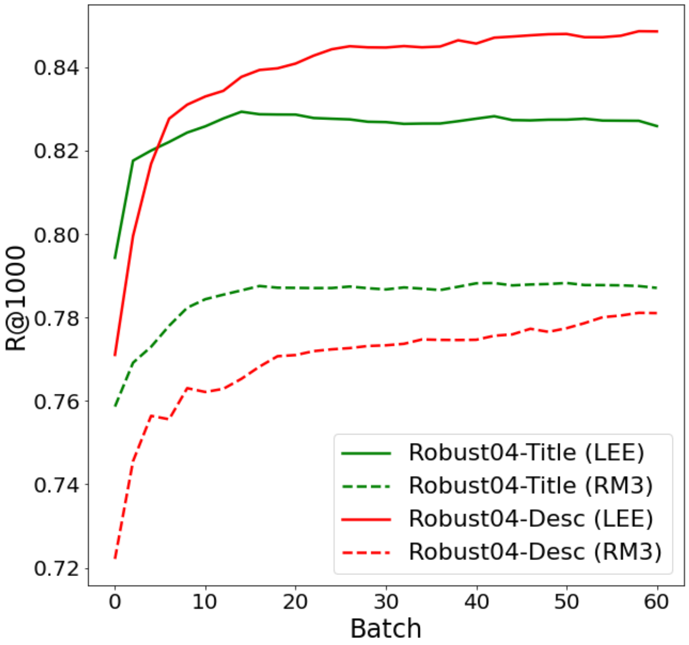


Figure 5.4: Adaptive expansion with RM3 and LEE.

Overall, I show that LEE, as part of an adaptive expansion framework, can significantly improve recall-oriented effectiveness without requiring a second pass of neural re-ranking. I achieve a new state-of-the-art effectiveness on complex topics, while incurring no additional computational expensive from PLM re-ranking.

RQ4.5b: Impact of hyperparameters

As outlined in Experimental Setup, I tune my LEE model following the official cross-validation setup outlined for target datasets. However, here I analyse: (1) how effective my method is zero-shot and (2) the impact of λ, i.e., the relative weighting of words and entities. This research question is intended to assess how robust LEE is to parameter shifts.

To assess LEE expansion in a zero-shot scenario, I use LEE parameters tuned on the CODEC dataset zero-shot on Robust04 titles and descriptions (the exact parameters can be found as part

Table 5.11: Tuned LEE model vs zero-shot LEE model (CODEC parameters); "+" and "-" are significance testing against tuned.

	Robust04 - Titles			Robust04 - Descriptions		
1x Re-Rank	NDCG	MAP	R@1000	NDCG	MAP	R@1000
Tuned	0.660	0.376	0.837	0.687	0.401	0.855
Zero-Shot	0.660	0.374	0.836	0.688	0.401	0.846
2x Re-Rank						
Tuned	0.667	0.393	0.837	0.715	0.438	0.855
Zero-Shot	0.667	0.394	0.836	0.710	0.435	0.846
Adaptive						
Tuned	0.668	0.392	0.838	0.704	0.435	0.834
Zero-Shot	0.668	0.393	0.837	0.701	0.432	0.829

of the released metadata). Table 5.11 shows the "Tuned" LEE expansion model against the "Zero-shot" parameters for my unsupervised LEE query, two rounds of PLM re-ranking, and adaptive expansion. I observed no significant differences between the tuned and zero-shot LEE run under these different scenarios, and in some cases zero-shot is the same or marginally better on Robust04 titles. Therefore, this highlights that my proposed method of using PLM passage feedback and combining words and entities with dependencies is robust to its parameter across datasets.

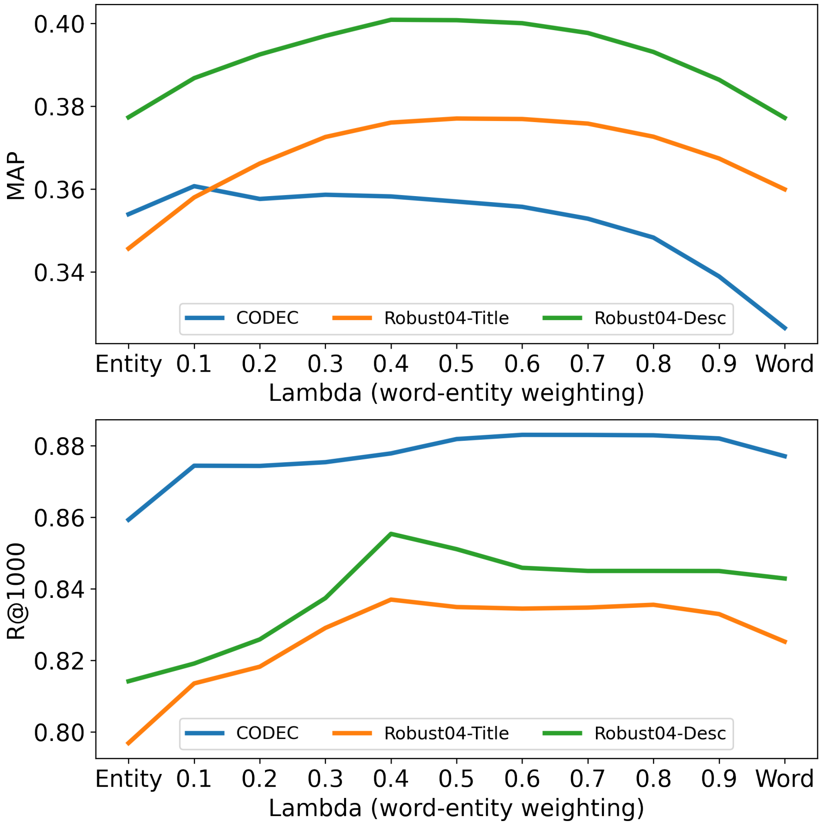


Figure 5.5: Lambda (i.e., relative word-to-entity weighting) impact for LEE expansion on CODEC and Robust04 datasets.

Figure 6.4 shows the impact of lambda (i.e., relative word-to-entity weighting) on the effectiveness of LEE unsupervised query across my target datasets. Specifically, I observe that for R@1000 and MAP, the best weighting is a combination of words and entities. For Robust04 datasets, MAP maximizes around 0.5, which weights word and entity expansions equally. However, for CODEC, precision is maximized around 0.1, favouring weighting entities and showing their precision benefits on domain-specific essay questions. On the other hand, Robust04 shows optimal recall with a relatively even combination of words and entities. However, unlike MAP, R@1000 for CODEC is maximized through a high weighting of words. Overall, this should show the precision-recall trade-offs for different datasets and confirms that both words and entities are required for robust effectiveness.

5.2.7 Conclusion

In this section, I show that the LEE word-entity hybrid expansion using fine-grain passage feedback from PLM re-ranking significantly improves R@1000 across target complex datasets, with between 8-14% improvement over RM3 expansion. I show that the joint modelling of words and entities at a passage level allows improved relevance modelling, including incorporating entity dependencies. These findings highlight that combining PLM ranking capabilities with entity-centric query expansion models can significantly improve effectiveness on complex queries.

I conduct in-depth query analysis and show my method is robust regarding query-level hurts vs helps and improves recall of the hardest 5% by queries by 0.6 on Robust04. Additionally, I find that my implicit entity ranking is highly effective within the top ranks using CODEC, which helps improve entity-centric queries. These qualitative findings highlight the real-world benefits of using PLMs within new query expansion pipelines.

In fact, I find that a second-pass PLM re-ranking on top of LEE, achieve the best NDCG, MAP, and R@1000 on the target datasets, and set a new state-of-the-art for recall-oriented effectiveness. Lastly, I combine LEE with adaptive re-ranking to avoid two PLM passes and achieve state-of-the-art effectiveness without additional document re-ranking. Overall, this work supports my hypothesis that novel query expansion pipelines leveraging PLMs can improve retrieval effectiveness on complex queries.

5.3 Chapter Conclusion

In this chapter, I address a fundamental issue common to all PRF models: when dealing with complex queries, the candidate set often lacks relevant documents in the top ranks, leading to PRF failure. To mitigate this, I leverage the ranking capabilities of pre-trained language models to enhance the precision of feedback, ultimately improving query expansion models. I also introduce an adaptive expansion technique that enables the use of PLM-based feedback without the need for a second-pass re-ranking. Additionally, I demonstrate that my Latent Entity Expansion model, which combines joint word and entity representations, is a new state-of-the-art retrieval model for complex topics.

It has been demonstrated that PLM-based re-ranking significantly improves effectiveness, particularly by enhancing precision at the top ranks. I observe this in my baseline experiments on CODEC and TREC Robust, where NDCG@20 increases by 25-50% with re-ranking. In this work, I pull together these research threads on neural re-ranking and query expansion. My findings show that modifying traditional expansion pipelines to include a PLM re-ranker before query expansion increases the recall of both sparse and dense expansion methods by 5-8%.

However, after applying neural feedback for query expansion, I found that a second round of neural re-ranking is still necessary to maximize precision. To address this, I propose the "Adaptive Expansion" framework, which dynamically refines the re-ranking pool during PLM scoring using the expansion model. This approach enables the use of PLM feedback for both expansion and re-ranking in a single pass, reducing the number of documents that need to be scored by approximately 35%. As a result, this new query expansion paradigm harnesses PLM capabilities to deliver precise expansion with minimal additional computational overhead.

I go on to evaluate the key components that contribute to the effectiveness of query expansion models when utilising PLM-ranked feedback, comparing factors such as documents vs. passages, entities vs. words, and various expansion models. My results show that the Latent Entity Expansion (LEE) model, which combines word-entity hybrid expansion with fine-grained passage-level feedback from PLM re-ranking, significantly boosts R@1000 across target datasets, achieving an 8-14% improvement over RM3 expansion. I demonstrate that jointly modelling words and entities at the passage level enhances relevance by capturing entity dependencies. These findings underscore that integrating PLM ranking with entity-centric query expansion models can greatly improve effectiveness, particularly for complex queries.

I perform an in-depth query analysis, demonstrating that my method is robust in balancing query-level gains and losses, with a notable 0.6 improvement in recall for the most challenging 5% of queries on the Robust04 dataset. Additionally, I find that my implicit entity ranking method is highly effective at improving top-rank results in CODEC, particularly for entity-centric queries. These qualitative insights emphasise the real-world advantages of incorporating PLMs into new query expansion pipelines. Furthermore, a second-pass PLM re-ranking applied on top of my LEE model achieves the highest NDCG, MAP, and R@1000 scores on target datasets, setting a new state-of-the-art for recall-oriented effectiveness on complex topics. By combining LEE with adaptive re-ranking, I eliminate the need for a second PLM pass while still achieving state-of-the-art results without additional document re-ranking.

Overall, this chapter addresses the issue of ineffective candidate feedback in PRF models for complex queries by leveraging PLM ranking capabilities to enhance feedback precision. I show that novel query expansion pipelines —incorporating PLM-based feedback and adaptive re-ranking — along with new models like Latent Entity Expansion, can dramatically improve retrieval effectiveness on complex queries without the need for additional re-ranking passes.

In the next chapter, we shift focus to leveraging PLM's generative capabilities to improve query expansion models. Furthermore, we'll explore pipelines that include both PLMs for ranking and PLMs for generating within a single pipeline for complex topics.

Chapter 6

Query Expansion with PLM Generation

In the previous chapter, I show how PLM capabilities of ranking relevant content can be leveraged to build new state-of-the-art query expansion pipelines to make significant gains on complex queries. In this chapter, I will focus on utilising PLM's generative capabilities to build new expansion pipelines. Specifically, developing new pipelines leveraging PLMs to generate relevant text content to help contextualise complex queries and improve retrieval effectiveness.

I propose a new query expansion paradigm, Generative Relevance Feedback (GRF), which instead of using feedback documents from the index to build an expansion model, GRF directly generates a variety of synthetic documents. I explore different types of documents to generate and GRF models across sparse, dense, and learned sparse paradigms. My results show that combining multiple types of generated documents is a more effective strategy than any individual document. Furthermore, I show that GRF is consistently and often significantly better than a compared pseudo-relevant feedback method, RM3. Lastly, I show that PRF and GRF methods have complementary ranking signals and can achieve state-of-the-art effectiveness on complex queries.

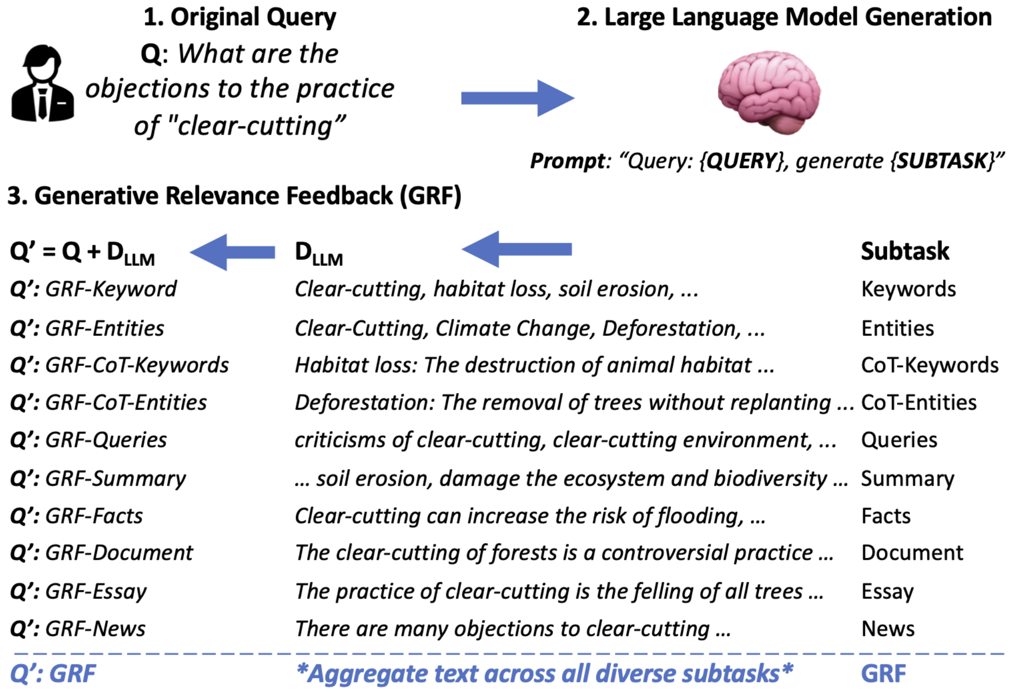
Drawing inspiration from the initial CODEC experiments where I show that multiple, diverse subtopics are effective content for query expansion for complex topics. Combining this insight with my GRF work, which shows that using GRF with PRF helps to lessen the impact of GRF hallucinations, I look to build a new query expansion model. Specifically, I develop Generative Relevance Modelling (GRM) that can weight the relevance of generated documents differently based on a relevance measure. I find that using PLMs to directly score the relevance of the generated content is ineffective. However, weighting my generated documents using ranking signals from "real" documents from the corpus that are semantically similar can be a highly effective proxy. Furthermore, I show that leveraging a PLM re-ranker for relevance scoring significantly improves retrieval effectiveness of GRF, allowing incorporation of subtopic-specific context.

This chapter has a section for both Generative Relevance Feedback and Generative Relevance Modelling. These results and analysis show the significant gains that can be achieved on complex topics by leveraging PLM generation capabilities as part of new query expansion pipelines.

6.1 Generative Relevance Feedback

Recent advances in pre-trained language models such as GPT-3, PaLM, ChatGPT demonstrate the capability to generate long-form fluent text. They are being combined with search engines, including BingGPT, to generate summaries of search results in interactive forms. In this work, I leverage these models not to generate end user responses, but to PLMs as input to the core retrieval algorithm.

Current query expansion models generally use pseudo-relevance feedback to improve first-pass retrieval effectiveness; however, this fails when the initial results are not relevant. Instead of building a language model from retrieved results, I propose Generative Relevance Feedback (GRF)(Figure 6.1) that builds query expansion models from long-form text generated from pre-trained language models. The hypothesis is that PLMs contain useful world knowledge [162] to help contextualise complex queries.


1. Original Query
Q: What are the objections to the practice of "clear-cutting"
2. Large Language Model Generation
Prompt: "Query: {QUERY}, generate {SUBTASK}"
3. Generative Relevance Feedback (GRF)

Q′ = Q + D_{LLM}	D_{LLM}	Subtask
Q′: GRF-Keyword	Clear-cutting, habitat loss, soil erosion, ...	Keywords
Q′: GRF-Entities	Clear-Cutting, Climate Change, Deforestation, ...	Entities
Q′: GRF-CoT-Keywords	Habitat loss: The destruction of animal habitat ...	CoT-Keywords
Q′: GRF-CoT-Entities	Deforestation: The removal of trees without replanting ...	CoT-Entities
Q′: GRF-Queries	criticisms of clear-cutting, clear-cutting environment, ...	Queries
Q′: GRF-Summary	... soil erosion, damage the ecosystem and biodiversity ...	Summary
Q′: GRF-Facts	Clear-cutting can increase the risk of flooding, ...	Facts
Q′: GRF-Document	The clear-cutting of forests is a controversial practice ...	Document
Q′: GRF-Essay	The practice of clear-cutting is the felling of all trees ...	Essay
Q′: GRF-News	There are many objections to clear-cutting ...	News
Q′: GRF	*Aggregate text across all diverse subtasks*	GRF

Figure 6.1: GRF uses PLM-generated text for query expansion.

I start by focusing on sparse GRF, where I build a probabilistic term-based expansion model combining the original query and synthetic documents generated by a PLM. I study the effective methods (prompts) for generating text of varying genres including: keyword queries, entities, facts, news articles, documents, and essays. I evaluate GRF on the target document retrieval benchmarks and show that GRF methods significantly outperform previous PRF methods on complex queries.

I study how PLMs can generate relevant text across diverse generation subtasks, including keywords, entities, chain-of-thought reasoning, facts, news articles, documents, and essays. I find that long-form text generation (e.g., news articles, documents, and essays) is 7-14% more effective as a feedback set compared to shorter texts (e.g., entities and keywords). Furthermore, the closer the generation subtask is to the style of the target corpus (e.g., news generation for newswire corpus or document generation for web document corpus), the more effective GRF is. Nonetheless, I find that aggregating the text generated across all subtasks is more effective than any standalone subtask, showing that these subtasks are complementary in extracting relevant knowledge contained in PLMs. The combined GRF method (which I refer to as "GRF"), improves MAP between 5-19% and NDCG@10 17-24% compared to RM3, and achieves the best R@1000 effectiveness on all datasets compared to state-of-the-art expansion baselines.

I go on to show that GRF can be easily extended to dense and learned sparse retrieval paradigms. In fact, I find that dense and learned sparse retrieval with GRF improves over comparable PRF techniques on nDCG@20 by 9%, while recall improves for learned sparse by 8% and dense by 11%. These results highlight how PLMs used as part of query expansion pipelines generalise across retrieval paradigms to improve retrieval effectiveness on complex topics.

Nonetheless, I conduct query analysis and find that generative and pseudo-relevance feedback have contrasting merits. For example, GRF provides external context not present in first-pass retrieval, i.e., PLMs can explain how the practice of "clear-cutting" relates to [habitat], [climate] and [deforestation]. Conversely, PLMs can hallucinate [85] and generate content not present in relevant documents. For example, PRF performs better on topics that need to be grounded to the corpus, i.e., "human stampede" where PRF correctly identifies events contained in relevant documents [Saudi] and [China], versus PLM-generated content that discusses stampedes in [India] and [Kerala] that are not present in relevant documents. Based on this analysis, Figure 6.2 shows the proposed weighted fusion method (PRF+GRF) that combines ranking signals of PRF and GRF to improve recall-oriented effectiveness.

Overall, this section presents Generative Relevance Feedback, showing how PLM's generative capabilities can be harnessed to develop novel expansion pipelines to improve the retrieval effectiveness of complex queries. The contributions of this work:

- I propose GRF, a generative relevance feedback approach which builds a relevance model using text generated from a PLM. This is the first published work leveraging larger generative models for query expansion.

- I demonstrate that PLMs can generate diverse and relevant text content for zero-shot relevance feedback to improve document retrieval, showing that long-form text in the style of the target dataset is the most effective.

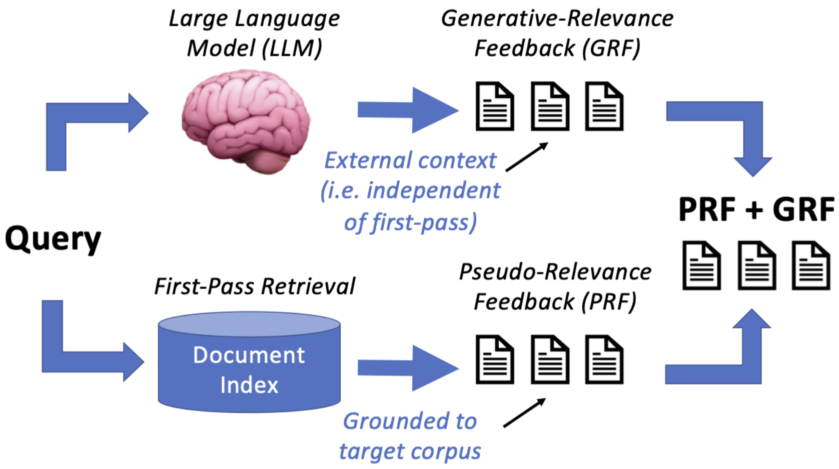

Large Language Model (LLM)
Generative-Relevance Feedback (GRF)
External context (i.e. independent of first-pass)
Query
PRF + GRF
First-Pass Retrieval
Document Index
Pseudo-Relevance Feedback (PRF)
Grounded to target corpus

Figure 6.2: PRF+GRF: Combing generative and pseudo-relevant feedback.

- I demonstrate that GRF is highly effective across retrieval paradigms (sparse, dense, and learned sparse), improving precision and recall over comparable PRF models by around 10% on complex queries.

- I conduct qualitative and quantitative query analysis on PRF and GRF to show these techniques have contrasting benefits. Furthermore, the fusion of GRF and PRF runs significantly improves R@1000 across all experiments, motivating our GRM work in the next section.

6.1.1 Methodology

I propose Generative Relevance Feedback (GRF) that tackles query-document lexical mismatch using text generation for zero-shot query expansion. Unlike traditional PRF approaches for query expansion [1, 136, 137], GRF is not reliant on first-pass retrieval effectiveness to find useful terms for expansion. Instead, I leverage PLMs [22] to generate zero-shot relevant text content.

I build upon prior work on Relevance Models [1] to incorporate the probability distribution of the terms generated by a PLM. This approach enriches the original query with useful terms from diverse generation subtasks, including keywords, entities, chain-of-thought reasoning, facts, news articles, documents, and essays. I find that the most effective query expansions leverage synthetic documents from: (1) long-form text generation subtasks and (2) text content closer in style to the target document corpus. In essence, I show that PLMs can effectively generate text context close to the target relevant documents.

Furthermore, I propose a GRF-Combined method that combines text content across all generation subtasks to construct a term-based probability distribution. The intuition behind this approach is that if the terms used are consistently generated across subtasks (e.g., within the entity, fact, and news generations), then these terms are likely useful for expansion. Additionally, multiple diverse subtasks also help expose tail knowledge or uncommon synonyms that are helpful for retrieval. I find this approach is more effective than any stand-alone generation subtask.

I also extend generative relevance feedback to dense [105] and learned sparse [55] retrieval paradigms. This shows that GRF generalises across retrieval paradigms to improve retrieval effectiveness on complex topics. Furthermore, based on my query analysis, I propose combining GRF and PRF ranking signals to further improve retrieval effectiveness on complex queries. Specifically, I show that GRF is useful in providing external context and not dependent on first-pass retrieval effectiveness, while PRF grounds the query in corpus-specific information.

Generation Subtasks

I study how PLMs can generate relevant text, d^{PLM}, across diverse generation subtasks for GRF expansion. This allows analysis of how GRF[1] effectiveness varies across the different generation subtasks:

- **Keywords (64 tokens)**: Generates a list of keywords. This subtask targets the important words or phrases for the topic, similar to subtopic generation [121, 176].
- **Entities (64 tokens)**: Generates a list of entities. This subtask targets relevant concepts, similar to KG-based approaches [42].
- **CoT-Keywords (256 tokens)**: Chain-of-thought (CoT) keywords including an explanation of "why" they are relevant. This forces the PLM to explain the reasoning selection.
- **CoT-Entities (256 tokens)**: Chain-of-thought (CoT) with relevant entities including an explanation of "why" they are relevant. This forces the PLM to explain the reasoning behind selection.
- **Queries (256 tokens)**: Generate a list of queries based on the original query. This is a type of PLM-based query reformulation.
- **Summary (256 tokens)**: Generate a summary based on the query. This forces the PLM to concisely write a summary (or answer) to satisfy the query.
- **Facts**: Generate a list of facts based on the query. The PLM is prompted to generate a knowledge-intensive list of sentences on the topic, close to [114].
- **Document (512 tokens)**: Generate a relevant document based on the query. This subtask is closest to generating a long-form pseudo-relevant web document.

[1] **Generated data for reproducibility: link**

- **Essay (512 tokens)**: Generate an essay based on the query. This is a long-form essay-style response to the query.
- **News (512 tokens)**: Generate a news article based on the query. This tasks the PLM to generate text in the style of a news article.

Table 6.1.1 shows how I prompt a PLM to generate query-specific text through 10 generation subtasks.

Table 6.1: GRF prompts and max token limits for generation subtasks

Generation Subtask	Max Tokens	Prompt Template
Keywords	64	f’Query: {query} Based on the query, generate a bullet-point list of relevant keywords present in relevant documents:’
Entities	64	f’Query: {query} Based on the query, generate a bullet-point list of relevant entities present in relevant documents:’
COT-Keywords	256	f’Query: {query} Based on the query, generate a bullet-point list of relevant keywords present in relevant documents. Next to each point, breifly explain why:’
COT-Entities	256	f’Query: {query} Based on the query, generate a bullet-point list of relevant entities present in relevant documents. Next to each point, breifly explain why:’
Queries	256	f’Query: {query} Based on the query, generate a bullet-point list of diverse keyword queries that will find relevant documents:’
Summary	256	f’Query: {query} Based on the query, write an summary:’
Facts	256	f’Query: {query} Based on the query, generate a bullet-point list of relevant facts present in relevant documents:’
Document	512	f’Query: {query} Based on the query, generate a relevant document:’
Essay	512	f’Query: {query} Based on the query, write an essay:’
News	512	f’Query: {query} Based on the query, write an news article:’

GRF Query Expansion

I formalise generative relevance feedback methods for sparse, dense, and learned sparse retrieval. This includes where a pre-trained language model, PLM, generates a single document for contextualisation, d^{PLM}, as well my “Combined” methods leveraging k number of generated documents $[d_1^{PLM}, d_2^{PLM}, ..., d_k^{PLM}]$.

Sparse GRF: Equation 6.1 shows how I estimate the probability of a term being relevant, $P_{PLM}(w|R)$, given the PLM-generated text document, d^{PLM}. I assume the generated document is relevant; thus, I set $P(q|d^{PLM}) = 1$. Therefore, $P_{PLM}(w|R)$ is equal to $P(w|d^{PLM})$, which is the term frequency divided by the document length.

$$P_{PLM}(w|R) = P(q|d^{PLM})P(w|d^{PLM}) \tag{6.1}$$

Similar to RM3 [1] expansion, Equation 6.4 shows GRF expansion, which combines the probability of a term given the original query $P(w|q)$ and probability derived from the PLM generation, $P_{PLM}(w|R)$. β (original query weight) is a hyperparameter to weight the relative importance of my generative expansion terms. Additionally, θ (number of expansion terms) is a hyperparameter with W_θ being the set of most probable PLM generated terms.

$$P_{GRF}(w|R) = \beta P(w|q) + \begin{cases} (1-\beta)P_{PLM}(w|R), & \text{if } w \in W_\theta. \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

The *Combined* sparse GRF expansion model concatenates text generated across all subtasks to produce d^{PLM}_{combo}. I then calculate $P(w|d^{PLM}_{combo})$ using this aggregated text, as outlined above. I find that the combination using the text across all types is most effective.

Dense GRF: I adopt the Rocchio PRF approach [106] for dense GRF to allow different weighting of the query vector and the feedback vector. This allows embeddings of PLM-generated text to contextualise the query vector in a controllable way. Specifically, Equation 6.3 shows that the new vector, $\vec{GRF}$, is the combination of the original query vector, $\vec{q}$, generated document vector, $\vec{d^{PLM}}$. I include α and β to weight the relative importance of the query and GRF vectors.

$$\vec{GRF} = \alpha\vec{q} + \beta\vec{d^{PLM}} \tag{6.3}$$

For the *Combined* dense GRF method, $\vec{GRF}$, is the combination of the original query vector, $\vec{q}$, and mean of the generated document vectors, $\vec{D^{PLM}} = 1/k \times (\vec{d_1^{PLM}} + \vec{d_2^{PLM}} + ... + \vec{d_k^{PLM}})$.

Learned Sparse GRF: I draw on prior work combining pseudo-relevance feedback with learned sparse representations [96]. Specifically, I combine the normalised learned sparse representations of the query, $LS(w|q)$, with the representation of a generated document, $LS(w|d^{PLM})$. β (original query weight) is a hyperparameter to weight the relative importance of my generative learned expansion terms, and θ (number of expansion terms) limits the most probable PLM-generated learned sparse representations. This results in the weightings of the query learned sparse to be re-weighted and new terms added based on the generated documents.

$$LS_{GRF}(w|R) = \beta LS(w|q) + \begin{cases} (1-\beta)LS(w|d^{PLM}), & \text{if } w \in W_\theta. \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

The *Combined* learned sparse GRF model is, for $LS(w|D^{PLM})$, similar to [146], I normalise each generated document's sparse representation and aggregate them together before normalising again.

Fusion of GRF and PRF

Query analysis shows that generative and pseudo-relevance feedback have different retrieval benefits and drawbacks. Thus, I propose combining these document-scoring signals. Specifically, Equation 6.5 shows my weighted reciprocal rank fusion method (WRRF) (adapted from [31]) that combines my GRF and PRF runs (PRF+GRF). Here, WRRF uses a scoring formula, $r(d)$, based on the document's rank in a specific run. There is a set D of documents to be ranked, a set of rankings R, and k parameter is included so low-rank document signals do not disappear (default usual 60). I add a hyperparameter λ, which weights the relative importance of pseudo-relevant document rankings, $r \in R_{PRF}$, and $(1-\lambda)$ for generative document rankings, $r \in R_{PRF}$. This formulation allows me to tune the relative weighting of GRF and PRF across models and datasets.

$$WRRF(d \in D) = \sum_{r \in R} 1/(k + r(d)) \times \begin{cases} \lambda, & \text{if } r \in R_{PRF}. \\ (1-\lambda), & \text{if } r \in R_{GRF}. \end{cases} \tag{6.5}$$

6.1.2 Experimental Setup

PLM Generation: For my text generation I use the GPT3 API [22]. Specifically, I use the text-davinci-002 model with parameters: temperature of 0.7, top_p of 1.0, frequency_penalty of 0.0, and presence_penalty of 0.0. I release all generation subtask prompts, generated text content and runs for reproducibility [2].

Retrieval and Expansion: I outline the different sparse, dense and learned sparse implementation details for GRF:

- **BM25+GRF**: To avoid query drift, all GRF runs in the paper use a tuned BM25 system for the input initial run [172]. I tune GRF hyperparameters: BM25 parameters ($k1$, b), the number of feedback terms (θ), and the interpolation between the original terms, and generative expansion terms (β). The tuning methodology is the same as BM25 and BM25 with RM3 expansion to make the GRF directly comparable.

- **TCT+GRF**: I use TCT-ColBERT-v2-HNP's [105] model trained on MS MARCO [147] and a max-passage approach. I shard documents into passages of 10 sentences, encoding the document title within each passage, and use a stride length of 5. I use the ColBERT-TCT encoder to create the GRF vectors, and tune Rocchio PRF α (between 0.1 and 0.9 with a step of 0.1), and β (between 0.1 and 0.9 with a step of 0.1).

- **SPLADE+GRF**: I use the SPLADE [55] *naver/splade-cocondenser-ensembledistil* checkpoint to create a passage index using the same processing as TCT+GRF. I use Pyserini [110] and their "impact" searcher for max-passage aggregation. When combining query or document vectors, I normalise the weights of each term. I tune fb_terms (20,40,60,80,100) and *original_query_weight* (between 0.1 and 0.9 with a step of 0.1).

See Chapter 3 for more details on baselines and experimental setup.

6.1.3 Research Questions

These research questions focus on how PLMs integrated into query expansion pipelines can develop new query expansion models that utilise this precise PLM-generated text content. My

[2] https://github.com/grill-lab/GRF

experimentation assesses the effectiveness gains of my proposed GRF and what types of documents help contextualise specific datasets. I compare my GRF methods to traditional PRF methods, and explore combining GRF and PRF ranking signals. My research questions are:

- **RQ5.1: What generative content is most effective for query expansion?** This research question focuses on understanding what standalone and combined GRF methodologies help on TREC Robust and CODEC datasets.
- **RQ5.2: How does GRF compare to state-of-the-art PRF models?** I compare my sparse, dense, and learned sparse GRF methods to comparable PRF methods to assess improvement on complex topics.
- **RQ5.3: What queries does GRF help vs PRF?** I conduct a deep query analysis to understand the benefits and drawbacks of PRF and GRF.
- **RRQ5.4: Do GRF and PRF have complementary ranking signals?** Based on my query analysis, I explore fusing GRF and PRF runs to assess whether they are complementary.

6.1.4 RQ5.1: What generative content is most effective for query expansion?

Table 6.2 and Table 6.3 show the effectiveness of generative feedback with varying units of text (Keywords-News) and my Combined method that uses text from all subtasks. I primarily focus on sparse GRF for this research question, and include results from TREC Deep Learning 2019 and 2020 to provide more diversity or corpora to base my conclusions.

Table 6.2: Effectiveness of sparse generative relevance feedback based on different generation subtasks, and *bold* depicts best system.

	Robust04 -Title			CODEC		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.435	0.252	0.717	0.327	0.218	0.748
Entities	0.452	0.252	0.698	0.341	0.216	0.750
CoT-Keywords	0.436	0.248	0.704	0.327	0.239	0.774
CoT-Entities	0.450	0.252	0.714	0.355	0.243	0.789
Queries	0.450	0.257	0.710	0.347	0.233	0.773
Summary	0.491	0.277	0.730	0.398	0.260	0.796
Facts	0.501	0.284	0.744	0.353	0.255	0.795
Document	0.480	0.276	0.728	0.376	0.265	0.795
Essay	0.494	0.284	0.736	**0.405**	0.270	0.803
News	0.501	0.287	0.745	0.398	0.270	0.828
Combined	**0.528**	**0.307**	**0.788**	**0.405**	**0.285**	**0.830**

Generation subtasks that target short text spans or lists (Keywords, Entities, Keywords-COT, Entities-COT, and Queries) are less effective. In fact, I find that subtasks that target longer text generation (Summary, Facts, Document, Essay, News) are more effective across all datasets.

Table 6.3: Effectiveness of sparse generative relevance feedback based on different generation subtasks, *bold* depicts best system.

	DL-19			DL-20		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.565	0.377	0.749	0.554	0.435	0.822
Entities	0.531	0.363	0.741	0.544	0.414	0.824
CoT-Keywords	0.550	0.382	0.748	0.542	0.423	0.817
CoT-Entities	0.563	0.389	0.757	0.552	0.430	0.832
Queries	0.551	0.367	0.760	0.568	0.439	0.851
Summary	0.577	0.414	0.761	0.585	0.472	0.865
Facts	0.569	0.401	0.769	0.583	0.459	0.871
Document	0.618	0.428	0.787	0.589	0.476	0.872
Essay	0.609	0.421	0.779	0.551	0.440	0.859
News	0.609	0.409	0.777	0.578	0.457	0.853
Combined	**0.620**	**0.441**	**0.797**	**0.607**	**0.486**	**0.879**

This indicates that more terms generated from the PLM provide a better relevance model, and increases MAP between 7-14 % when I compare these two categories.

Furthermore, I find the most effective generation subtasks are aligned with the style of the target dataset. For example, Facts and News are the best standalone generation methods across all measures on Robust04, where the dataset contains fact-heavy topics and a newswire corpus. Additionally, Essay and News are the best generation subtasks on CODEC across all measures, which aligns with its essay-style queries over news (BBC, Reuters, CNBC, etc.) and essay-style (Brookings, Forbes, eHistory, etc.) corpus. Lastly, Document is the best generation subtask across DL-19 and DL-20, aligning with MS Marcos web document collection. Overall, these finding support that PLM generative content in the styles of the target task is the most effective.

Nonetheless, the Combined GRF method is consistently as good if not better than any standalone subtask, improving NDCG by 0.0-5.4%, MAP by 2.1-7.0% and R@1000 by 0.2-5.8% across the datasets. This shows that combining PLM generated text from various generation subtasks is a robust and effective method of relevance modelling.

I also include dense and learned sparse results in Appendix B. These results align with and further support my findings, that long-form text in the style of the target dataset is the most effective individual expansion context. However, the Combined method, leveraging PLM generated text from our diverse subtasks, consistently outperforms any stand-alone method.

6.1.5 RQ5.2: How does GRF compare to state-of-the-art PRF models?

Table 6.4 shows GRF (Combined) against state-of-the-art sparse, dense, and PRF methods across our target complex datasets. Specifically, I compare my sparse GRF model to BM25 [172] with RM3 expansion [1], my dense GRF model to Colbert TCT [105] with Roccio feedback [112], and learned sparse GRF with SPLADE [55] with RM3 expansion. These research questions focus on complex queries, and I solely report results on Robust and CODEC, conducting significance testing against the comparable PRF baselines for each GRF paradigm.

Table 6.4: GRF against state-of-the-art sparse, dense and learned sparse PRF models. '+" indicates significance improvements against comparable PRF method for each paradigm and *bold* depicts best system.

	Robust04 - Title			Robust04 - Description			CODEC		
Sparse	NDCG@10	MAP	R@1000	NDCG@10	MAP	R@1000	NDCG@10	MAP	R@1k
BM25	0.445	0.252	0.705	0.415	0.227	0.664	0.316	0.214	0.783
BM25+RM3	0.451	0.292	0.777	0.445	0.278	0.750	0.326	0.239	0.816
BM25+GRF	**0.528**$^{+}$	**0.307**	**0.788**	**0.550**$^{+}$	**0.318**$^{+}$	**0.776**$^{+}$	**0.405**$^{+}$	**0.285**$^{+}$	**0.830**
Dense									
TCT	0.466	0.233	0.637	0.424	0.214	0.595	0.351	0.211	0.740
TCT+PRF	0.493	0.274	0.684	0.452	0.245	0.628	0.358	0.239	0.757
TCT+GRF	**0.517**$^{+}$	**0.276**	**0.700**$^{+}$	**0.571**$^{+}$	**0.289**$^{+}$	**0.708**$^{+}$	**0.385**	**0.261**	**0.821**$^{+}$
Learned Sparse									
SPLADE	0.387	0.206	0.660	0.426	0.230	0.672	0.309	0.183	0.726
SPLADE+RM3	0.418	0.248	0.703	0.448	0.268	0.715	0.311	0.216	0.770
SPLADE+GRF	**0.462**$^{+}$	**0.265**$^{+}$	**0.730**$^{+}$	**0.493**$^{+}$	**0.276**	**0.732**$^{+}$	**0.337**	**0.222**	**0.785**

These results show that GRF has the best nDCG@10, MAP, and R@1000 across all datasets and retrieval paradigms. GRF is consistently, often significantly better, across the board and highlights the effectiveness gains on complex topics from including PLM-generated text in expansion models. For sparse GRF, when compared to RM3, I observe nDCG@10 significantly improves on all datasets with 17-14% relative improvement, highlighting the considerable precision benefits of generative feedback. Furthermore, I see 5-19% relative gains in MAP, and smaller but consistent gains in R@1000 of 1-4%.

These results strongly support that PLM generation is an effective method for query expansion without relying on first-pass retrieval effectiveness. For example, I look at the hardest 20% of Robust04 ordered by NDCG@10 of BM25; I find that RM3 offers minimal uplift improvements with NDCG@10 by +0.006, MAP by +0.008, R@1000 by +0.052. In contrast, because GRF is not reliant on first-pass retrieval effectiveness, and GRF improves NDCG@10 by +0.145, MAP by +0.068, and R@1000 by +0.165 (a relative improvement of 100-200% on NDCG@10 and MAP).

When I compare TCT with GRF to PRF, I observe GRF outperforms dense retrieval with PRF on these complex datasets, with a performance gap of 5-24% on NDCG@10, 1-18% MAP, and 2-13%. There is a significant improvement on 3/3 datasets for R@1000, 2/3 nDCG@10, and MAP just on Robust Descriptions. Similarly to sparse GRF, I observe the large relative gains on Robust Descriptions and CODEC, where the longer queries provide better context for the PLM to generate relevant synthetic documents. Similarly, SPLADE with GRF improves nDCG@10 compared to RM3 between 8-11%, with significant improvements in 2/3 datasets. Additionally, there is significant improvement in 1/3 datasets for MAP (3-7% relative improvement) and 2/3 datasets for R@1000 (3-7% relative improvement).

These results support GRF as an effective and robust query augmentation approach across sparse, dense, and learned sparse retrieval paradigms. This supports the fact that PLM-generated content is highly effective at contextualising complex queries. In the following research question, I conduct query analysis to understand the different behaviours of generated and pseudo-relevant feedback.

6.1.6 RQ5.3: What queries does GRF help vs PRF?

In this section, I analyse the sparse, dense and learned sparse GRF runs. Figure 6.3 shows the query difficulty plot stratified by nDCG@10 effectiveness of BM25 on Robust04 titles. I report MAP and also include BM25 with RM3 and GRF expansion. Specifically, this shows the hardest (0-25%) and easiest (75-100%) first-pass queries based on precision. It is noticeable that GRF is better than PRF on the hardest 75% of first-pass queries (0-75% in my chart), with MAP consistently above RM3 expansion. However, on the easiest first-pass queries (75-100% in my chart), RM3 is more effective than GRF. In essence, I show that pseudo-relevance feedback is more effective than generative-relevance feedback when first-pass precision is very high.

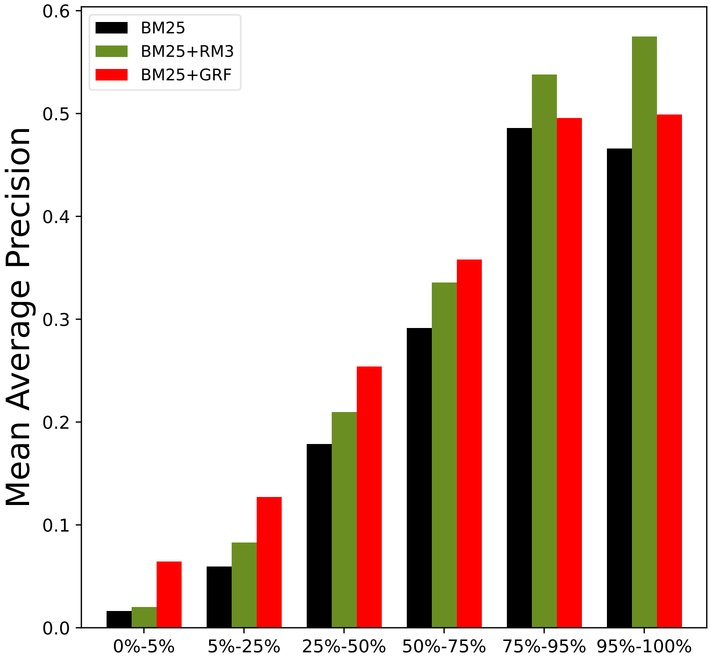


Figure 6.3: Query difficulty plot stratified by nDCG@10 of BM25 on Robust04 titles. I show MAP effectiveness of BM25 and BM25 with RM3 and GRF expansion.

Furthermore, I analyse the query helps vs hurts for Robust04 titles, comparing BM25 query effectiveness to RM3 and GRF expansion. For R@1000, RM3 hurts 47 queries and helps 139 queries, while GRF impacts more queries, hurting 53 and helping 150. Interestingly, of the 47 queries that RM3 hurts, 33 (70%) are either improved or unaffected by GRF. Conversely, of the 53 queries where GRF hurts effectiveness, 40 (75%) of queries are helped or unaffected by RM3 expansion. Again, this highlights that generative and pseudo-relevance feedback affect different queries, suggesting these could have complementary ranking signals.

Specifically, a hard first-pass topic is 691 on Robust04 descriptions, *What are the objections to the practice of "clear-cutting"*. BM25 has a R@1000 of 0.333 and nDCG@10 of 0.000, and

RM3 expansion further reduces R@1000 to 0.286 with nDCG@10 unchanged at 0.000. Reviewing the pseudo-relevant documents used for RM3 expansion, these are general documents around the forest industry and expand with terms: [forest], [timber], [logging], and [industry]. On the other hand, GRF expansion uses generated content that directly addresses the questions and injects terms about the "objections" to clear-cutting. Analysing the different generated documents, CoT-Keywords expands with [habitat], [climate] and [deforestation], Facts expands with [flooding], [water], [risk], and News expands with [soil], [erosion], and [environment]. The term expansions can be seen in Table 6.1.6. This results in BM25 with GRF expansion increasing R@1000 to 0.810 and nDCG@10 to 0.454.

Table 6.5: RM3 vs GRF expansion terms on Robust query "What are the objections to the practice of "clear-cutting"?"

	nDCG@10	R@1000	Top 20 Expansion Terms (highest weighted left, lowest weighted right)
BM25	0.000	0.333	-
BM25+RM3	0.000	0.286	forest forests nthe would clear from has logging n cutting timber trees have industry said more about new forestry its
Keywords	0.073	0.333	can clear cutting practice objection relevant some potential objections lead soil erosion destroy habitats make difficult regrowth
Entities	0.000	0.333	clear cutting objections practice
COT-Keywords	0.158	0.667	habitat loss climate deforestation change soil erosion removal trees without replanting them otherwise restoring land its previous tree cover destruction
COT-Entities	0.220	0.545	loss deforestation habitat climate change soil erosion water pollution biodiversity
Queries	0.369	0.476	clear cutting objections criticisms effects environment solutions
Summary	0.000	0.429	can lead main objections practice clear cutting soil erosion damage local ecosystem loss biodiversity
Facts	0.234	0.619	clear cutting can damage local water sources increase risk flooding wildlife habitats
Document	0.440	0.619	can clear cutting have lead loss forests highly controversial practice due negative environmental social impacts involves removal all trees area
Essay	0.249	0.514	can clear cutting trees all area removed practice objections have impact well species number negative environment loss when animals left
News	0.314	0.667	can clear cutting lead increase many objections practice type logging where all trees area cut down extremely harmful environment soil
Combined	**0.454**	**0.810**	can clear cutting loss practice lead soil objections erosion trees habitat all area climate local have impact change water environment

In contrast, an easy query for first-pass retrieval is topic 626 on Robust04 titles, *human stampede*, where BM25 achieves nDCG@10 of 0.287 and R@1000 of 1.000, and RM3 expansion improves nDCG@10 to 0.517 while retaining perfect recall. Reviewing the RM3 expansion terms and the relevant documents, I see specific terms that are useful and refer to collection-mention stampedes, i.e., [Saudi], [China], [pilgrimage], and [Calgary]. Conversely, without being grounded in the events covered in the collection, GRF expands with general terms, i.e. [human], [death], [crowd], [panic], [tragedy], or terms relating to events not converted in the collection, i.e. [India], [Kerala], [2011], etc. In this case, the PLM-generated content refers to the Sabarimala Temple stampede, which occurred seven years after the Robust04 corpus, i.e., PLM hallucinations and off-topic content can harm specific queries.

Overall, this analysis highlights that generative and pseudo-relevance feedback help different profiles of queries, which suggests they are complementary. Thus, in the next research question, I explore combing the ranking signals of generative and pseudo-relevant feedback to further improve retrieval effectiveness on complex queries.

6.1.7 RQ5.4: Do GRF and PRF have complementary ranking signals?

Table 6.6 shows the effectiveness of PRF, GRF and my weighted reciprocal rank fusion (PRF+GRF), across my three search paradigms (sparse, dense, learned sparse). For sparse, I use BM25 with RM3 expansion as the baseline for significance testing; for dense, I use ColBERT-TCT with PRF; for learned sparse, I use SPLADE with RM3 expansion.

Table 6.6: The effectiveness of fusing PRF and GRF runs. "+" indicates significant improvements against PRF from the respective search paradigm (i.e., BM25+RM3 for sparse, etc.), and *bold* depicts best system.

	Robust04 - Title		Robust04 - Desc		CODEC	
Sparse	MAP	R@1000	MAP	R@1000	MAP	R@1000
BM25+RM3	0.292	0.777	0.278	0.750	0.239	0.816
BM25+GRF	0.307	0.788	0.318^{+}	0.776^{+}	$\mathbf{0.285}^{+}$	0.830
BM25+(PRF+GRF)	$\mathbf{0.323}^{+}$	$\mathbf{0.817}^{+}$	$\mathbf{0.331}^{+}$	$\mathbf{0.823}^{+}$	0.275^{+}	$\mathbf{0.853}^{+}$
Dense						
TCT+PRF	0.274	0.684	0.245	0.628	0.239	0.757
TCT+GRF	0.276	0.700^{+}	0.289^{+}	0.708^{+}	**0.261**	$\mathbf{0.821}^{+}$
TCT+(PRF+GRF)	$\mathbf{0.287}^{+}$	$\mathbf{0.707}^{+}$	$\mathbf{0.303}^{+}$	$\mathbf{0.727}^{+}$	**0.261**	$\mathbf{0.821}^{+}$
Learned Sparse						
SPLADE+RM3	0.248	0.703	0.268	0.715	0.216	0.770
SPLADE+GRF	$\mathbf{0.265}^{+}$	0.730^{+}	0.276	0.732^{+}	0.222	0.785
SPLADE+(PRF+GRF)	$\mathbf{0.265}^{+}$	$\mathbf{0.743}^{+}$	$\mathbf{0.276}^{+}$	$\mathbf{0.757}^{+}$	**0.225**	**0.790**

I find that combining PRF and GRF consistently, often significantly, improves recall across datasets and search paradigms. For example, fusion has the best R@1000 across 12/12 and significantly improves over PRF on 11/12 experiments (2 more than GRF alone). I find consistent improvements in R@1000 across the search paradigms over PRF, with fusion increasing by 6.5% on sparse, by 6.9% on dense, and by 3.5% on learned sparse on average. Additionally, I see on MAP increases by 11.2% on sparse, by 12.5% on dense, and by 3.4% learned sparse on average. Fusion also shows small but consistent improvements over GRF alone on R@1000 and MAP.

To understand the effect of hyperparameter λ, Figure 6.4 plots R@1000 of my weighted reciprocal rank fusion method (BM25+(PRF+GRF)) varying λ, i.e., when λ is 0.0 this is the equivalent of GRF, 0.5 equates to RRF [31], and 1.0 equates to PRF. This shows that generative and pseudo-relevant feedback methods are complementary. For example, R@1000 increases as λ approaches 0.3-0.6, highlighting the benefits of combined ranking signals.

Overall, I show that I can further improve recall on complex queries by combining the ranking signals of generative and pseudo-relevant feedback models. I show that GRF and PRF are

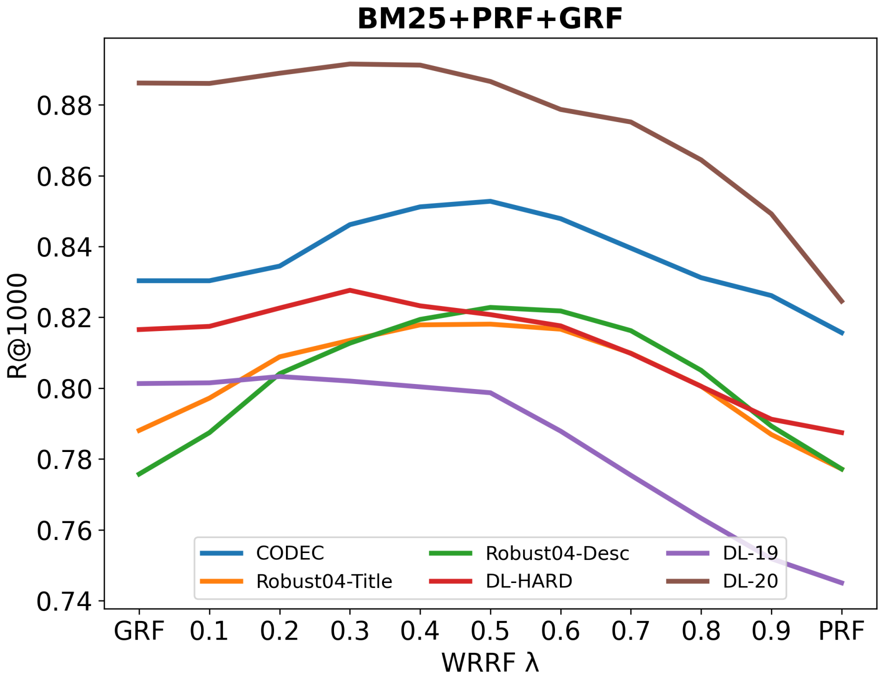


Figure 6.4: λ on R@1000 of weighted reciprocal rank fusion (WRRF). Where 0.0 is GRF and 1.0 is PRF.

complementary and explore the impact of weighting generative and pseudo-relevant feedback signals across datasets and models.

6.1.8 Conclusion

In this section, I demonstrate that I can improve retrieval effectiveness on complex queries by utilising PLM's generative capabilities to build new expansion pipelines. Specifically, I propose Generative Relevance Feedback (GRF), which means that instead of using feedback documents, I directly generate a variety of synthetic documents using PLMs for my expansion model. This approach leverages world knowledge contained in PLMs to help contextualise complex queries.

I explore different types of documents to generate useful text content. Specifically, I study how PLMs can generate relevant text across diverse generation subtasks, including keywords, entities, chain-of-thought reasoning, facts, news articles, documents, and essays. I find that long-form text generation (e.g., news articles, documents, and essays) where the generation subtask is to the style of the target corpus (e.g., news generation for the newswire corpus or document generation for the web document corpus) is the most effective context for GRF. Furthermore, these results show that combining multiple types of generated documents is a more effective strategy than any individual document.

I develop GRF expansion models across sparse, dense, and learned sparse paradigms. I demonstrate that GRF is highly effective in each retrieval paradigm, improving, often significantly, precision and recall over comparable PRF models by around 10%. These are some of the best initial retrieval results across these complex datasets and highlight how PLM generative content can be used to improve retrieval effectiveness.

I conduct detailed query analysis on GRF and PRF systems, finding that generative and pseudo-relevance feedback have contrasting merits. Specifically, GRF provides external context that is not present in first-pass retrieval, but PLMs can hallucinate and generate content that is not present in relevant documents. On the other hand, PRF performs better on topics that need to be grounded to the corpus and where first-pass retrieval is effective. Based on this analysis, I proposed a weighted fusion method that combine the ranking signals of PRF and GRF to further improve recall-oriented effectiveness. In the next section, I further develop this thread by leveraging the corpus to construct query expansion models based on PLM-generated content more effectively.

6.2 Generative Relevance Modelling with Relevance Aware Sample Estimation

In the previous section, my work on Generative Relevance Feedback (GRF) demonstrates that using text produced by PLMs for query expansion can enhance the initial search results. However, a challenge with PLMs is their tendency to “hallucinate”, which can negatively impact the performance of specific queries. Therefore, although some generated documents may accurately match the user’s information need, any generated content that deviates from this need can lead to suboptimal query expansion terms, ultimately compromising retrieval effectiveness.

To illustrate, consider the complex query, “What technological challenges does Bitcoin face to becoming a widely used currency?”. Providing an answer goes beyond merely listing Bitcoin facts. It requires a deep dive into Bitcoin’s technological framework, an overview of the larger financial landscape, insights into potential scalability hurdles, security risks, and an analysis of the socioeconomic elements that drive the uptake of new technologies. Each subtopic of this question is intricate, and when combined, they present a layered query that demands a detailed answer. We saw this in the CODEC reformulation results where "golden" reformulations were extremely effective query content. However, if the PLM were to generate subtopics that were not relevant, like the lack of privacy, this would cause the query expansion model to fail.

The proposed Generative Relevance Model (GRM) is depicted in Figure 6.5 and designed for complex queries containing multiple diverse subtopic. This expands work on sub-topic based query enrichment methods using topic models [56]. Specifically, I use a PLM in this pipeline to produce synthetic documents that address subtopics linked to the original query. These generated documents are intended for query expansion, and my GRM allows individual relevance weighting.

However, my findings indicate that ranking these generated documents directly for query expansion is not optimal. This is mainly because the ranking model is not attuned to the specifics of

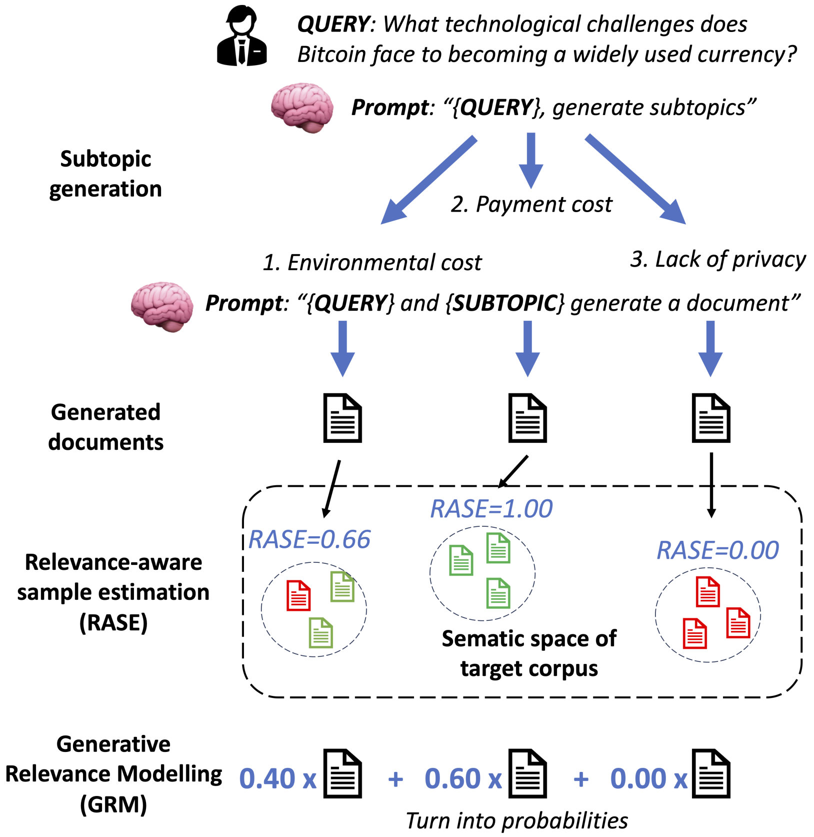

QUERY: What technological challenges does Bitcoin face to becoming a widely used currency?
Prompt: "{QUERY}, generate subtopics"
Subtopic generation
2. Payment cost
1. Environmental cost
3. Lack of privacy
Prompt: "{QUERY} and {SUBTOPIC} generate a document"
Generated documents
RASE=0.66
RASE=1.00
RASE=0.00
Relevance-aware sample estimation (RASE)
Sematic space of target corpus
Generative Relevance Modelling (GRM)
0.40 x
+ 0.60 x
+ 0.00 x
Turn into probabilities

Figure 6.5: My approach allows a diverse range of generated documents to be weighted within an expansion model based on the estimated relevance of semantically similar documents from the collection.

the target collection. To address this issue, I introduce my Relevance-Aware Sample Estimation (RASE) method, which identifies documents from the collection that resemble each generated document, before employing an external scoring function to gauge their relevance. Doing so gives more weight to generated documents that closely align with relevant documents from the target collection. This allows us to select useful not useful subtopics, thus enabling diversity in the document synthesis process.

I conduct experiments using my complex datasets. I generate 50 documents for each topic within these datasets. However, I observe that the effectiveness of these generated documents for query expansion varies significantly. Some greatly improve search results, achieving a high Recall@1000 score of 0.83, while others were less effective, scoring only 0.59. I show that using GRM with RASE allows me to more effectively weight my query expansion models and leads to large and significant gains across my datasets. Specifically, I show that using a neural re-ranker [150] as my relevance estimation improves MAP by 6-9% and Recall@1k by 2-4% over state-of-the-art GRF on complex topics. My findings unveil a promising novel direction for query expansion models based on PLM-generated documents but grounded in target collection. Furthermore, an efficiency study shows that I achieve these improvements with minimal increases in computation by utilising score caching.

My contributions are:

- I present GRM with RASE, an approach for modeling the relevance of generated documents based on the estimated relevance of similar documents from the target collection. This method is designed to contextualise complex topics that contain multiple subtopics.
- I demonstrate how PLMs can generate documents addressing subtopics of complex information needs. However, my analysis shows that expansion effectiveness varies dramatically depending on which document is used for feedback.
- I show that RASE, using a PLM re-ranker as a scoring function, significantly improves over expansion baselines. This highlights that PLM generative and ranking capabilities can be combined to create state-of-the-art expansion effectiveness on complex queries.
- An efficiency study shows computational trade-offs when using RASE as part of a retrieve and re-ranking pipeline. Highlighting we can improve effectiveness on complex queries with minimal computational cost.

6.2.1 Methodology

I formally define my Generative Relevance Modeling (GRM) approach using Relevance-Aware Sample Estimation (RASE). I first outline my approach to generating synthetic documents based on subtopics to ensure diversity, which is critical for complex topics. Yet, my analysis shows a significant variance in query expansion effectiveness based on which generated documents are used within a relevance model. Thus, I define Generative Relevance Modeling (GRM), including a query likelihood score, to provide my expansion model with a means of reducing the impact of hallucinations and increasing important knowledge for contextualisation. Lastly, I formally define Relevance-Aware Sample Estimation, which is aimed at estimating the relevance of generated synthetic documents to the initial query, grounded in a target collection. My intuition is that generated documents similar to the collection's relevant documents should provide better feedback signal.

Document Generation

As I outlined in my complex criteria, complex topics have multiple subtopic or subtopics. To provide a more comprehensive view of the information related to the query, I generate documents that cover these different subtopics. Specifically, given the initial query q, I utilize PLM [22] with chain-of-thought reasoning [204] to first generate i subtopics for the query q. Then, I prompt the PLM to generate documents based on these subtopics, $D^{PLM} = \{d_1^{PLM}, d_2^{PLM}, \ldots, d_i^{PLM}$ where d_i^{PLM}, the i-th PLM-generated document, covers the i-th subtopic for the query. This strategy helps avoid missing relevant documents that might be overlooked if one only focused on a single aspect of the query. To provide a balance between depth (exploring a subtopic in more detail through multiple documents) and breadth (covering multiple subtopics), I perform this generation G times, giving me N ($K \times G$) diverse query-specific generated documents.

Generative Relevance Model

GRM builds upon prior work on Relevance Modelling [1] but specifically focuses on generative expansion. In this framework, I assume the generated documents, denoted as D^{PLM}, are relevant and thereby define them as a relevant set, R. The critical task is to estimate the probability of observing a word w given the relevance set, i.e., $P(w|R)$. Formally:

$$P(w|R) = \sum_{d^{PLM} \in R} P(w|d^{PLM})P(q|d^{PLM}) \tag{6.6}$$

In this equation, $P(q|d^{PLM})$ is the query likelihood score, which quantifies the relevance of a document d^{PLM} to the original query q. If a score-based method is used we apply a normalisation as is standard practice [110]. My results show that directly scoring the generated documents is not effective; thus, I developed the RASE method, which will be discussed next.

Relevance-Aware Sample Estimation

RASE focuses on estimating the query likelihood score, $P(q|d^{PLM})$, by modelling the relevance of documents in a collection that are similar to the generated document d^{PLM}. Specifically, for a given document collection D, I identify a subset of documents $[D_1, D_2, \ldots, D_T]$ that are closest to d^{PLM} according to a similarity function $\psi(d^{PLM}, D)$. Inspired by recent work [125], I simply use BM25 [172] to retrieve my list of similar documents from the collection. Next, I compute $P(q|d^{PLM})$ using the relevance signals from $P(q|d^i)$, operationalised via a Discounted Cumulative Gain (DCG) approach [83]:

$$P(q|D^{PLM}) = P(q|d^i) + \sum_{i=2}^{K} \frac{P(q|d_i)}{log_2(i)} \tag{6.7}$$

The DCG technique enables me to integrate a variety of relevance estimation models, from simpler methods like BM25 to more complex neural re-rankers like T5-3B [150]. Additionally, I can estimate an upper bound on my relevance estimation by leveraging query relevance judgments ("gold estimation").

6.2.2 Experimental Setup

Document Generation

GRM is a method to weight the diverse generative content during query expansion. Similar to GRF, I use GPT-3 API [22] for my document generation. Specifically, I use the `gpt-3.5-turbo` model on Chat mode with parameters: $temperature = 0.7$, $top\_p = 1.0$, $frequency\_penalty = 0.0$, $presence\_penalty = 0.0$, and maximum length of 512. I 1-shot prompt ChatGPT to create five subtopics ($K = 5$) before I generate documents based on these subtopics. I repeat this process $G = 10$ times to create a reasonable set of 50 diverse generated documents per topic. I release all prompts and generated documents for reproducibility: ***link***

Parameter tuning

I tune the BM25 model [172] for each dataset. Specifically, I tune the k_1 parameter within the range of 0.1 to 5.0, with a step size of 0.2, and the b parameter within the range of 0.1 to 1.0, with a step size of 0.1. For RASE, inspired by prior work on document-to-document similarities [125], I apply the tuned BM25 model for the document similarity function ψ, treating the generated document as a query. Additionally, I tune the T number of documents retrieved from the target collection, ranging from 10 to 100 with a step of 10.

I tune the remaining GRM hyperparameters: the number of feedback docs (fb_docs: 5 to 95 with a step of 10), the number of feedback terms (fb_terms: 5 to 50 with a step of 10), the interpolation between the original terms and generative expansion terms ($original\_query\_weight$: 0.1 to 0.9 with a step of 0.1). The tuning methodology is the same as BM25, BM25 with RM3 expansion and GRF to make them directly comparable.

Relevance Estimate Functions

For my relevance estimate function, $P(q|d^C)$, I use the following formulations and normalise scores:

- **RASE-Uniform**: I set the relevance estimation to 1.0 to show the impact of other methods. This should be similar to GRF [127] excluding differences in document lengths.

- **RASE-BM25**: A tuned BM25 [172] model provides an efficient term-based scoring function.

- **RASE-T5**: I use the T5-3B [150] re-ranker to maximise effectiveness, leveraging max-passage aggregate to calculate the document scores.

- **RASE-Gold**: I use scaled relevance judgments to show Oracle performance, i.e., upper bound in terms of effectiveness for RASE with a perfect ranker (relevance judgments).

I also compare to **GRM-T5**, where a T5-3b re-ranker [150] directly scores my generated documents based on the original query (without RASE grounding to target collection). These scores build my GRM query expansion model, tuning fb_docs, fb_terms and $original\_query\_weight$.

See Chapter 3 for more details on baselines and experimental setup.

6.2.3 Research Questions

These research questions focus on developing and understanding novel query expansion pipelines that leverage PLM-generated text content. I conduct analysis to understand the variance of retrieval effectiveness based on different PLM-generated documents. This motivates my GRM expansion models and uses RASE as a method of weighting generated content. Lastly, I study the computational requirements to understand efficiency trade-offs.

- **RQ6.1: Does the selection of generated documents impact query expansion effectiveness?** In this research question, I focus on understanding how effectiveness vary with PLM-generated documents for query expansion.
- **RQ6.2: Does Relevance-Aware Sample Estimation improve the effectiveness of Generative Relevance Modeling?** In the research question, I compare my GRM methods to state-of-the-art PRF and GRF baselines, showing the effectivenss gains on complex topics.
- **RQ6.3: What about the computational requirements?** Lastly, I understand the computational resources required for these novel PLM-enabled query expansion pipelines.

6.2.4 RQ6.1: Does the selection of generated documents impact query expansion effectiveness?

Figure 6.6 displays how the choice of generated documents used for expansion impacts retrieval effectiveness on Robust04 title queries. The boxplot represents the query effectiveness distribution if I select documents for expansion from the worst to the best (from left to right). Thus, the far right shows the distribution of 250 queries if I always picked the "best" generated document, and the far left shows the distribution of 250 queries if I always picked the "worst" generated document.

Effectiveness based on selection quality: The selection of generated documents significantly impacts the overall effectiveness. For instance, MAP ranges from 0.21 for the worst-generated documents per query to 0.34 for the Oracle-generated document. Similarly, the worst possible document results in an R@1000 of 0.59, whereas the best-generated document increase recall by 0.24 to 0.83.

Looking at a specific query, Figure 6.7 show generated documents with the highest and lowest recall for CODEC topic `economics-8` (How has the push toward electric cars impacting the demand for raw materials). This example highlights how specific generations can result in off-topic hallucinations that do not align with the target collection or information need, i.e., focusing on ethical considerations of mining. On the other hand, I can see how certain generated documents align with the underlying information need, i.e., focusing on the raw materials (cobalt, nickel, lithium, etc.) and countries that make up the supply chain (DRC, Indonesia, etc.).

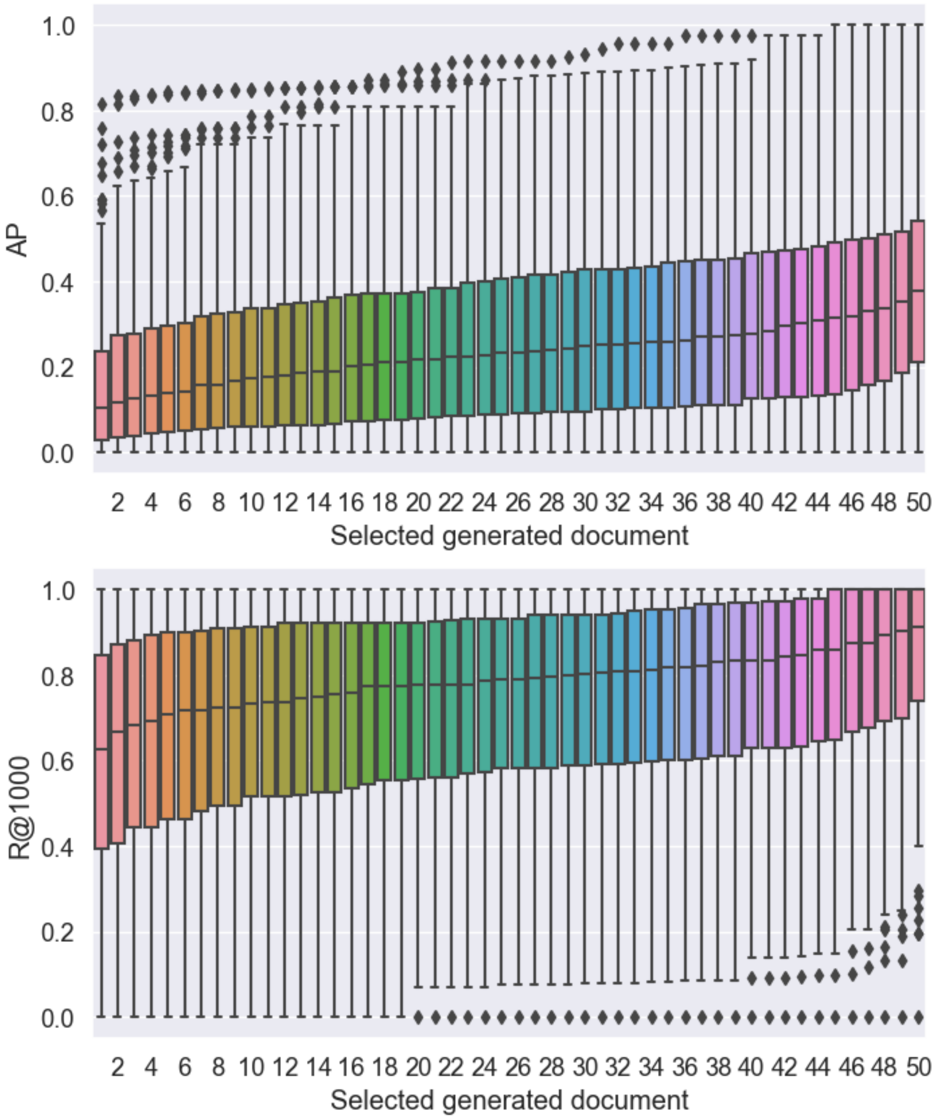
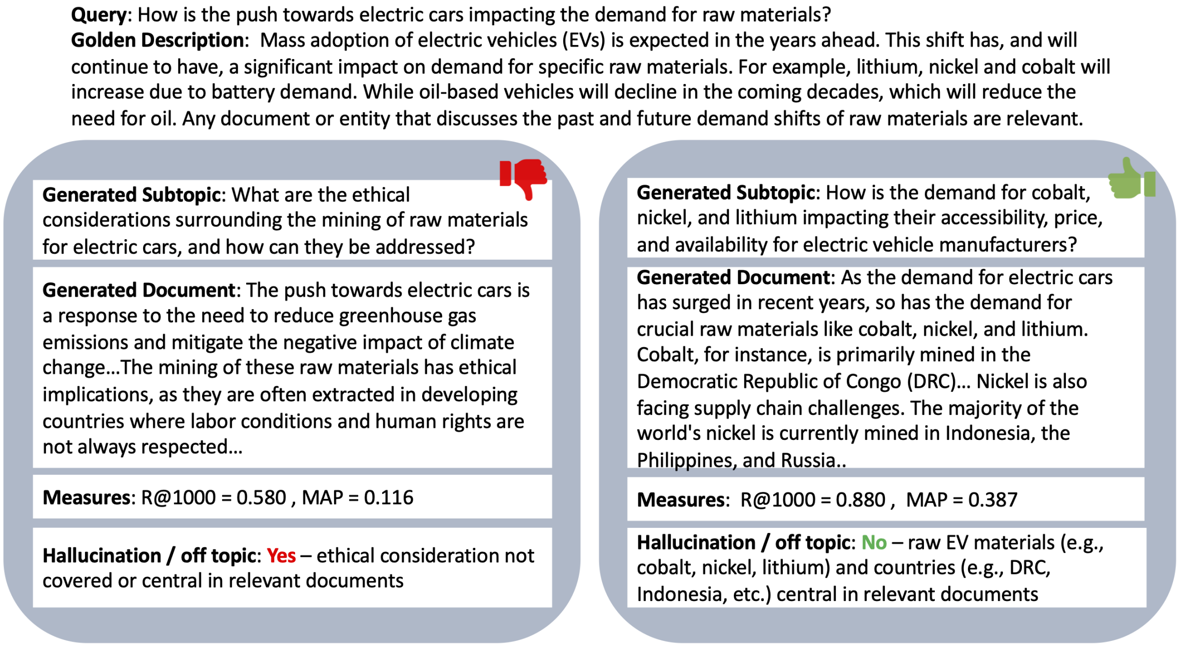


Figure 6.6: MAP and R@1000 boxplot of varying generated documents ordered by effectiveness on Robust04 titles, i.e. the worst generated document for expansion for each topic to the left (1) and best generated document to the right (50).

Query variance: The boxplots on Figure 6.6 show that the effectiveness differs greatly based on the query, suggesting high variance even with constant selection quality. For example, some queries achieve a MAP of 0.0 (even with the Oracle document), indicating no relevant documents were retrieved. Conversely, some queries achieve a MAP of 0.81 and R@1000 of 1.0, even with the worst possible document. This demonstrates that generating multiple documents doesn't necessarily alleviate the difficulty of hard queries.

These findings strongly suggest the potential of PLMs for generating documents for query expansion, especially for complex topics. However, they also highlight an issue: the effectiveness of different generated documents used for query expansion can vary dramatically. Some documents align with the content in relevant documents from the target collection, others do



Query: How is the push towards electric cars impacting the demand for raw materials?
Golden Description: Mass adoption of electric vehicles (EVs) is expected in the years ahead. This shift has, and will continue to have, a significant impact on demand for specific raw materials. For example, lithium, nickel and cobalt will increase due to battery demand. While oil-based vehicles will decline in the coming decades, which will reduce the need for oil. Any document or entity that discusses the past and future demand shifts of raw materials are relevant.

Generated Subtopic: What are the ethical considerations surrounding the mining of raw materials for electric cars, and how can they be addressed?

Generated Document: The push towards electric cars is a response to the need to reduce greenhouse gas emissions and mitigate the negative impact of climate change…The mining of these raw materials has ethical implications, as they are often extracted in developing countries where labor conditions and human rights are not always respected…

Measures: R@1000 = 0.580 , MAP = 0.116

Hallucination / off topic: Yes – ethical consideration not covered or central in relevant documents

Generated Subtopic: How is the demand for cobalt, nickel, and lithium impacting their accessibility, price, and availability for electric vehicle manufacturers?

Generated Document: As the demand for electric cars has surged in recent years, so has the demand for crucial raw materials like cobalt, nickel, and lithium. Cobalt, for instance, is primarily mined in the Democratic Republic of Congo (DRC)… Nickel is also facing supply chain challenges. The majority of the world's nickel is currently mined in Indonesia, the Philippines, and Russia..

Measures: R@1000 = 0.880 , MAP = 0.387

Hallucination / off topic: No – raw EV materials (e.g., cobalt, nickel, lithium) and countries (e.g., DRC, Indonesia, etc.) central in relevant documents

Figure 6.7: The worst (left) and best (right) generated documents for CODEC topic economics-8.

not due to hallucinations. In the following section, I propose a solution to weight generated documents more effectively based on their similarity to the collection's relevant documents.

6.2.5 RQ6.2: Does Relevance-Aware Sample Estimation improve the effectiveness of Generative Relevance Modeling?

Table 6.7 presents document retrieval effectiveness on Robust04 title and description datasets using different relevance estimate functions. Interestingly, RASE-Uniform performs similarly to GRF on Robust04, suggesting that blindly generating diverse subtopics does not consistently improve query expansion for specific topics. Similarly, GRM with T5 to rank the generated documents directly (without RASE) offers no effectiveness improvements over RASE-Uniform. This highlights that scoring the generated documents independently of the target collection is ineffective. Specifically, the ranking function cannot assess the semantic overlaps between the generated documents and the target collection. However, using BM25 for relevance estimation based on sampled documents leads to a small but consistent improvement, although not significant gains over GRF.

Utilising a PLM re-ranker [150], such as in RASE-T5, reveals that better RASE can more effectively weight beneficial generated documents for query expansion. Specifically, this results in significant gains across all measures on Robust04 descriptions and on MAP and nDCG on Robust04 titles over GRF. In fact, RASE-T5 outperforms all PRF models. My results also indicate that an Oracle relevance estimation could further boost performance for MAP by 13-19% and for R@1000 by 2-3%. These results highlight the possible effectiveness gains on

Table 6.7: Retrieval effectiveness of GRM with different RASE methods on Robust04. "+" indicates significant improvements against GRF and *bold* depicts the best system (excluding our oracle *RASE-Gold* method).

		Robust04 - Titles			Robust04 - Descriptions		
		MAP	nDCG	R@1000	MAP	nDCG	R@1000
PRF	TCT+PRF	0.274	0.541	0.684	0.245	0.493	0.628
	SPLADE+RM3	0.248	0.518	0.703	0.268	0.535	0.715
	CEQE	0.310	0.579	0.764	-	-	-
	BM25+RM3	0.292	0.571	0.777	0.278	0.551	0.750
GRF	GRF	0.307	0.603	0.788	0.318	0.605	0.776
GRM	T5	0.304	0.594	0.778	0.313	0.605	0.778
	RASE-Uniform	0.306	0.594	0.778	0.313	0.605	0.779
	RASE-BM25	0.312	0.599	0.781	0.315	0.607	0.779
	RASE-T5	**0.327**$^{+}$	**0.615**$^{+}$	**0.796**	**0.342**$^{+}$	**0.631**$^{+}$	**0.805**$^{+}$
	RASE-Gold	*0.388*$^{+}$	*0.672*$^{+}$	*0.819*$^{+}$	*0.387*$^{+}$	*0.675*$^{+}$	*0.823*$^{+}$

Table 6.8: Retrieval effectiveness of GRM with different RASE methods on CODEC. "+" indicates significant improvements against GRF and *bold* depicts the best system (excluding our oracle *RASE-Gold* method).

		CODEC		
		MAP	nDCG	R@1000
PRF	TCT+PRF	0.239	0.532	0.757
	SPLADE+RM3	0.216	0.506	0.770
	BM25+RM3	0.239	0.530	0.816
GRF	GRF	0.285	0.585	0.830
GRM	T5	0.302^{+}	0.609^{+}	0.845
	RASE-Uniform	0.306^{+}	**0.611**$^{+}$	**0.850**$^{+}$
	RASE-BM25	0.303^{+}	0.608^{+}	0.843
	RASE-T5	**0.309**$^{+}$	**0.611**$^{+}$	0.848^{+}
	RASE-Gold	*0.336*$^{+}$	*0.642*$^{+}$	*0.855*$^{+}$

complex topics by combining PLM's generation and ranking capabilities in query expansion pipelines.

Table 6.8 shows retrieval effectiveness on CODEC, which illustrates this method on both complex essay-style queries [131]. On CODEC, all GRM methods with RASE significantly outperform GRF. Since CODEC's complex topics often encompass multiple subtopics, my diverse subtopic-driven prompting approach greatly enhances the effectiveness of RASE-Uniform. Relevance estimation via BM25 offers a slight improvement in R@1000, but T5 significantly improves all measures. Interestingly, while RASE-Gold boosts precision in MAP and nDCG, it shows less improvement in recall. These results, combined with TREC Robust 2004, support the benefit of generative relevance modelling on complex topics.

Figure 6.8 shows Robust04 titles R@1000 effectiveness of GRM with RASE-Uniform and RASE-T5 when I vary the number of subtopic documents (K) and generations (G). I observe that increasing K or G beyond a certain point using RASE-Uniform leads to very little or no increase

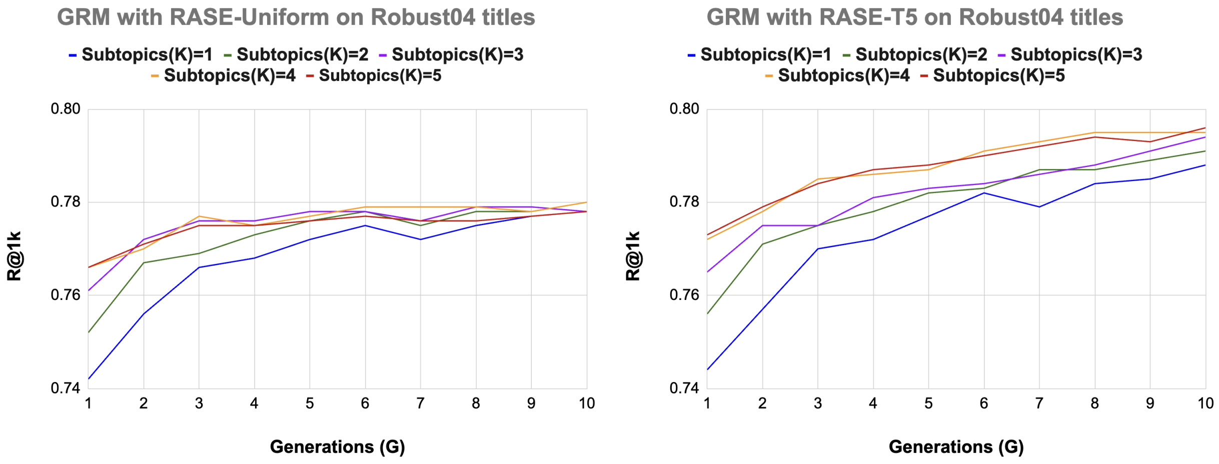


Figure 6.8: Robust04 titles R@1000 of GRM with RASE-Uniform (left) and RASE-T5 (right) varying number of subtopic documents (K) and generations (G).

in effectiveness. However, RASE-T5 almost linearly increases effectiveness with more generated documents. Specifically, this highlights that RASE-T5 can discriminate between useful generated documents and those that are off-topic and hallucinations.

Overall, my results demonstrate that using a PLM re-ranking for RASE in GRM significantly improves effectiveness over GRF methods, leading to an increase of 6-9% in MAP, 2-4% in nDCG, and 2-4% in R@1000. This highlights the potential of GRM for state-of-the-art document retrieval across multiple complex datasets.

6.2.6 RQ6.3: What about the computational requirements?

The usage of PLMs within query expansion pipelines requires additional computational resources [225], whether they are used for re-ranking or query expansions. As such, I note that my method requires PLM generation of K documents G times for each query (5×10). However, as Figure 6.8 shows, the number of generated documents used for query expansion could be reduced, and RASE-T5 would still achieve state-of-the-art recall.

Table 6.9: Computational analysis of number of unique documents scored [150] per query (D/Q) and effectiveness. "+" indicates significant improvements against BM25+RM3 ⇒ T5 and *bold* best system.

	Robust04 - Titles				Robust04 - Descriptions			
	D/Q	MAP	nDCG	R@1k	D/Q	MAP	nDCG	R@1k
BM25+RM3⇒T5	1,000	0.377	0.634	0.777	1,000	0.406	0.652	0.750
RASE-T5	787	0.327	0.615	**0.796**	394	0.342	0.631	**0.805**$^{+}$
RASE-T5⇒T5	1,402	**0.382**	**0.646**	**0.796**	1,169	**0.423**$^{+}$	**0.688**$^{+}$	**0.805**$^{+}$

Moreover, RASE computes the relevance of each of the ($K \times G$) generated documents based on sampled documents from that target collection, up to the depth of T. Thus, a single query will require computational requirements based on my ranking function [150] ($K \times G \times T$). Across the datasets, my tuned T averages around 50, however, by caching the already scored documents, I can reduce computation to the average number of unique documents per query (D/Q). The D/Q ratio and RASE-T5 effectiveness are shown in Table 6.9. I find that RASE-T5 improves recall over a BM25+RM3 re-ranked with T5 and requires fewer unique documents to score. Nonetheless, I require a second pass of re-ranking to increase MAP and nDCG over a standard retrieve and re-rank pipeline. Specifically, this improves MAP and nDCG between 1-6% and significantly improves Robust04 descriptions, but requires additional scoring of a few hundred more documents.

In this work, I primarily focus on the effectiveness of my methods rather than efficiency. Recent research efforts have been directed towards smaller, more efficient PLMs [177], carbon footprint-friendly methods [119], and sustainable approaches to natural language processing [143]. Therefore, I leave efficiency improvements for future work by utilizing smaller yet effective models or optimizing the hyperparameters K, G, and T. Instead, I demonstrate that significant improvements in retrieval effectiveness can be made by using generative PLMs for query expansion on complex topics.

6.2.7 Conclusion

This section extends my work using PLMs' generative capabilities to improve query expansion models for complex topics. Generative Relevance Feedback demonstrates that using text produced by PLMs for query expansion can improve the initial search results. However, a challenge with PLMs is their tendency to hallucinate and generate non-relevant text, which can negatively impact query effectiveness. I explore this hypothesis further through subtopic document generation, and confirm a large variance in the effectiveness of generated documents as query expansion context. For example, some generated documents dramatically improved search results, achieving a high Recall@1000 score of 0.83, while others made effectiveness much worse, scoring only 0.59.

Therefore, I propose the Generative Relevance Model to weight generated documents differently based on contextualisation benefit (relevance). However, my results show that directly scoring the relevance of generated documents does not work, as it requires understanding of the target corpus to know what will or won't be useful. To address this issue, I introduce my Relevance-Aware Sample Estimation method, which identifies documents from the collection that resemble each generated document, before employing an external scoring function to gauge their relevance. Doing so gives more weight to generated documents that closely align with relevant documents from the target collection.

I show that using GRM with RASE allows weighting of query expansion models more effectively and leads to large and significant gains across my complex datasets. Specifically, I show that using a PLM re-ranker as my relevance estimation improves MAP by 6-9% and Recall@1k by 2-4% over state-of-the-art GRF, and achieving the best generative expansion results on my target datasets. Thus, this novel expansion pipeline combined both PLM's generative and ranking capabilities to improve the retrieval effectiveness of complex queries, and supports the hypothesis of this thesis.

6.3 Chapter Conclusion

In this chapter, I explore the use of pre-trained language models and their generative capabilities to create innovative expansion pipelines. My focus is on leveraging PLMs to generate relevant text content that contextualizes complex queries and enhances retrieval effectiveness. Specifically, I introduce Generative Relevance Feedback, which generates a variety of synthetic documents rather than relying on feedback documents from the index to build an expansion model. Additionally, I develop Generative Relevance Modelling, which encodes a query likelihood score based on grounding synthetic documents to the corpus and leveraging PLM ranking capabilities.

I propose a new query expansion paradigm, Generative Relevance Feedback (GRF), that instead of using feedback documents from the index to build an expansion model, I directly generate a variety of synthetic documents. I find that long-form text generation (e.g., news articles, documents, and essays) where the generation subtask is to the style of the target corpus (e.g., news generation for the newswire corpus or document generation for the web document corpus) yields the most effective context for GRF. Furthermore, I find that combining text across various generation tasks proves to be more effective than focusing on any single style. I propose GRF models tailored for sparse, dense, and learned sparse retrieval paradigms, demonstrating that GRF significantly improves precision and recall approximately 10% across all retrieval approaches when compared to traditional PRF models.

I conduct qualitative and quantitative query analysis comparing GRF and PRF systems, revealing that generative and pseudo-relevance feedback offer distinct advantages. Specifically, GRF provides external context absent in the first-pass retrieval, though it risks hallucination, by generating content not found in relevant documents. In contrast, PRF excels in scenarios where topics need grounding in the corpus and where first-pass retrieval is effective. Building on this analysis, I propose a weighted fusion method that integrates the ranking signals from both approaches to enhance recall-oriented effectiveness. My findings indicate that PRF and GRF methods possess complementary ranking signals, enabling me to achieve state-of-the-art results on complex queries. I further develop this concept by leveraging the corpus to construct more effective query expansion models within my GRM work.

Specifically, I develop Generative Relevance Modelling (GRM) that can weight the relevance of generated documents differently based on a relevance measure. To further investigate this hypothesis, I delve into subtopic document generation and observe significant variability in the effectiveness of generated documents as contextual expansion for queries. For instance, some generated documents substantially enhance search results, achieving an Recall@1000 40% higher than documents containing a high proportion of hallucinations. Consequently, Generative Relevance Model can assign different weights to generated documents based on their contextual relevance. I find that directly scoring the relevance of generated documents is ineffective without grounding them in the target corpus.

To tackle this challenge, I present my Relevance-Aware Sample Estimation method, which identifies documents in the collection that resemble each generated document and applies an external scoring function to assess their relevance. This approach allows me to prioritise generated documents that closely align with relevant documents from the target collection. I demonstrate that combining GRM with RASE enhances the effectiveness of my query expansion models, resulting in substantial improvements across my complex datasets. Specifically, utilising a PLM re-ranker for relevance estimation yields increases in MAP of 6-9% and Recall@1000 of 2-4% compared to state-of-the-art GRF, achieving the best generative expansion results on my target datasets. Thus, this innovative expansion pipeline effectively integrates both the generative and ranking capabilities of PLMs to enhance the retrieval effectiveness of complex queries, reinforcing the hypothesis of this thesis.

Chapter 7

Conclusion and Future Work

7.1 Contributions and Conclusions

In this thesis, I have explored the integration of pre-trained language models (PLMs) into multi-stage query expansion pipelines, focusing on enhancing document retrieval effectiveness for complex queries. My research is driven by the hypothesis that PLMs can significantly improve retrieval performance over traditional sparse and dense models by leveraging their capabilities to both rank and generate relevant content. This premise guides my development of novel query expansion models, which are particularly advantageous for handling complex queries that require additional reasoning and contextualisation across multiple documents and entities.

7.1.1 Contributions

This thesis makes significant contributions to the information retrieval community and the field query expansion. These include:

- **Complex Query Criteria**: In Chapter 4, I conduct an in-depth analysis of hard queries from past datasets and establish a framework for categorising complex queries. This classification is grounded in thoroughly examining queries and model effectiveness from TREC Deep Learning [130]. Based on this, I define complex queries as multifaceted, involving multiple entities and requiring deep comprehension and knowledge. Additionally, I develop an annotation process to systematically identify these query types, ensuring applicability across different datasets. This complex criteria can be utilised by the broader IR community, as demonstrated in this thesis, to prioritize model research on high-impact query types.
- **Analysis of Prior Datasets**: In Chapter 4, I leverage the complex query criteria and deeply annotate multiple datasets, releasing these findings. For example, I demonstrate that standard datasets, such as TREC Deep Learning, predominantly consist of non-complex

queries. However, I find that most TREC Robust 2004 queries exhibit complexity. Additionally, I conduct an in-depth system analysis to highlight where state-of-the-art retrieval and re-ranking systems struggle, identifying key areas for improvement. This analysis reinforces TREC Robust as a suitable dataset for evaluating complex information needs while underscoring the need for new dataset development.

- **Construction of the CODEC Dataset**: My analysis of complex queries drives the creation of a new dataset to support fundamental research on these topics [131]. Introduced in Chapter 4, CODEC provides complex essay-style queries, gold-standard "narratives" capturing the underlying information need, query facets, and dense annotations of relevant documents and entities. This novel resource facilitates developing and evaluating retrieval models designed for complex queries. Additionally, I analyse state-of-the-art retrieval and re-ranking systems to identify areas for improvement in query expansion models. I find that even simple expansion techniques incorporating entities and query facets enhance ranking effectiveness. These findings serve as a foundation for future research in retrieval and ranking. Furthermore, CODEC is can be leveraged by the wider IR community as a unique resource focusing on complex queries.

- **PLM Re-Ranking for Expansion Feedback**: In Chapter 5, I reimagine query expansion pipelines by leveraging the ranking capabilities of pre-trained language models to construct a more targeted expansion model [128]. This shift improves performance across sparse and dense retrieval by 5–8%. Furthermore, I introduce the "adaptive expansion" concept, inspired by how humans iteratively refine queries. This approach employs a neural re-ranker to update the expansion model as additional documents are scored dynamically, eliminating the need for re-ranking excess documents (reducing re-ranking cost by around 35%). This method significantly enhances effectiveness, particularly for complex queries.

- **Fine-Grained Query Expansion with Words and Entities**: I investigate novel query expansion models utilising feedback from re-ranking from PLMs. My investigation reveals that employing a re-ranking PLM to construct relevance models based on passage spans proves more effective than utilising entire documents. Therefore, I introduce an enhanced expansion model called Latent Entity Expansion (LEE), which applies fine-grained word and entity-based relevance modelling incorporating localised features [128], and further increasing NDCG by 2-8%. Additionally, LEE improves recall of the hardest 5% of queries by 0.6 and shows significant gains on entity-centric queries due to a strong entity representation.

- **PLM-Generated Content for Query Expansion**: In Chapter 6, I propose rethinking query expansion by leveraging pre-trained language models (PLMs) to generate synthetic documents. My contribution is the first published work [127] demonstrating that PLMs can effectively generate text for probabilistic query expansion, a method I term Generative Relevance Feedback (GRF). I show that long-form text generation, when tailored to match the style of the target corpus, is highly effective for query expansion. However, combining outputs from multiple generation strategies produces a more robust and effective expansion model than any single prompting technique in isolation. This work demonstrates that GRF improves MAP by 5–19% over RM3 expansion. Additionally, I show that GRF generalises well to dense and learned sparse retrieval, outperforming comparable PRF techniques by approximately 10% on both precision and recall-oriented metrics. These findings highlight how enhanced contextualisation through PLMs significantly improves effectiveness for complex queries.

- **Reducing Hallucinations in Generative Relevance Modeling**: In Chapter 6, analysis of Generative Relevance Feedback (GRF) reveals that hallucinations—where PLMs generate non-relevant text—can cause queries to drift off-topic. To mitigate this issue, I propose multiple solutions, including a simple PRF fusion method and a more sophisticated pipeline approach. Specifically, I introduce Relevance-Aware Sample Estimation (RASE), which improves term weighting in Generative Relevance Models (GRMs) by assessing the relevance of each generated document. This is achieved by identifying real documents from the target corpus that are similar to the generated ones and employing a neural reranker to estimate their relevance. This approach improves MAP by 6–9% and R@1k by 2–4%, outperforming previous GRF methods.

7.1.2 Conclusions

This section discusses the achievements and conclusions of this work.

Understanding and measuring retrieval effectiveness in complex queries. I begin my investigation by analyzing existing datasets and starting to understand the characteristics of complex queries. My work on DL-HARD involves augmenting the TREC Deep Learning collections with detailed annotations, including question intent categories, answer types, wikified entities, and topic categories. This enables me to systematically categorise the impact of various query types on the effectiveness of current retrieval models. I conclude that existing systems struggle with multifaceted topics that demand substantial reasoning and concern multiple entities. Nonetheless, most current information retrieval datasets focus on non-complex topics, which are factoid questions seeking short answers, where current systems are already highly effective.

This leads me to establish a complex query criteria to categorise the queries this thesis focuses on effectively. Specifically, queries that are multifaceted, concern multiple entities, and requires complex reasoning and comprehension through significant research. Furthermore, I outline an extensive annotation process that allows me to effectively assess whether a dataset is complex. After ruling out many established information retrieval datasets, such as the TREC Deep Learning collection, I ultimately identify TREC Robust 2004 as a primary complex dataset for this thesis. This process highlights the current research and resource gap around complex queries.

Therefore, I build a second complex dataset from the ground up: the Complex Document and Entity Collection (CODEC). This dataset specifically addresses complex information needs within social science domains such as history, finance, and politics. My analysis of the CODEC system reveals significant headroom for both state-of-the-art traditional models and neural rankers when handling complex queries. Furthermore, I show a positive correlation between document relevance and the presence of relevant entities, which led me to introduce a ground-truth entity query expansion method that outperforms strong baseline systems. Additionally, I illustrate the role of query reformulation in uncovering latent dimensions within complex topics which can be valuable query expansion context.

Overall, these findings show that current information retrieval datasets largely focus on easy queries for current systems that center around factoid questions. Through deep analysis, I define and show that complex queries are multifaceted, concern multiple entities, and require complex reasoning and comprehension through significant research. Through my complex criteria and new datasets, such as CODEC, I build up an understanding of these queries and lay the foundation for future analysis, including the role of entities and query reformulation.

Improving of query expansion effectiveness using PLMs to rank content. My first family of new query expansion models and pipelines centre around harnessing the ranking capabilities of pre-trained language models. I observe this in my baseline experiments on CODEC and TREC Robust datasets, where NDCG@20 increases by 25-50% with PLM re-ranking. This leads me to hypothesise that incorporating PLM-focused feedback can enhance query expansion models. I show that modifying traditional expansion pipelines to include a PLM re-ranker before query expansion increases the recall of both sparse and dense expansion methods by 5-8%.

Nonetheless, after using PLM-focused feedback, I still find that a second round of neural re-ranking is necessary to maximise precision. To avoid the computational overhead of two separate re-ranking passes, I introduce Adaptive Expansion, which dynamically refines the re-ranking pool during PLM scoring using the expansion model. This innovative approach allows me to leverage PLM feedback for both expansion and re-ranking in a single pass, thereby reducing the number of documents that require scoring by approximately 35%. As a result, this new query expansion paradigm effectively harnesses PLM capabilities to deliver accurate expansions while avoiding additional computational overhead.

I go on to propose, Latent Entity Expansion, a query expansion model designed from the ground up to leverage PLM feedback for complex queries. LEE combines word-entity hybrid expansion with fine-grained passage-level feedback from PLM re-ranking, significantly boosts R@1000 across target datasets, achieving an 8-14% improvement over RM3 expansion. I demonstrate that jointly modelling words and entities at the passage level enhances relevance by capturing entity dependencies. These findings highlight the potential of integrating PLM ranking with entity-centric query expansion models to significantly enhance effectiveness, especially for complex queries. Furthermore, by combining LEE with adaptive re-ranking, I eliminate the need for a second PLM pass, allowing us to achieve state-of-the-art results without additional document re-ranking.

Overall, these findings demonstrate that using a pre-trained language model to rank content for expansion models leads to substantial effective gains on complex topics. Specifically, I show large improvements using traditional PRF models with PMF feedback, how I can optimize efficiency using Adaptive Expansion, and produce state-of-the-art retrieval effectiveness using my LEE expansion model on complex topics.

Improving query expansion effectiveness using PLMs to generate content. My second family of new query expansion models and pipelines centres around harnessing the generative capabilities of pre-trained language models. Specifically, I propose a new query expansion paradigm, Generative Relevance Feedback (GRF), that shifts away from traditional feedback documents being used to build an expansion model and instead generates a diverse range of synthetic documents. My research reveals that long-form text generation—such as news articles, documents, and essays—tailored to match the style of the target corpus (for instance, generating news articles for a newswire corpus or creating documents for a web document corpus) provides the most effective context for GRF. Additionally, I show that combining text across various generation tasks yields better results than any generation task in isolation. I develop GRF models for sparse, dense, and learned sparse retrieval paradigms, demonstrating that GRF significantly improves both precision and recall by approximately 10% across all retrieval methods when compared to comparable PRF models.

I conduct a comprehensive query analysis comparing generative and pseudo-relevance feedback systems, revealing their distinct strengths and weaknesses. GRF offers useful external context absent in first-pass retrieval but risks hallucination. In contrast, PRF is effective in grounding topics within the corpus, but only when the initial retrieval is strong. Thus, I develop Generative Relevance Modelling that allows me to weight the relevance of generated documents differently. For this method, I use subtopic document generation and observe significant variability in the effectiveness of generated documents as contextual expansion for queries. Yet, I find that directly scoring the relevance of generated documents is ineffective without grounding them in the target corpus.

To tackle this challenge, I present my Relevance-Aware Sample Estimation method, which identifies documents in the collection that resemble each generated document and applies an external scoring function to assess their relevance. I demonstrate that combining GRM with RASE enhances the effectiveness of my query expansion models, resulting in significant improvements across my complex datasets. Specifically, utilising an PLM re-ranker for relevance estimation yields increases in MAP of 6-9% and Recall@1000 of 2-4% compared to state-of-the-art GRF, achieving the best expansion results on my target datasets. This innovative expansion pipeline effectively merges the generative and ranking capabilities of PLMs, significantly enhancing the retrieval effectiveness for complex queries and reinforcing the thesis hypothesis.

Overall, these findings demonstrate that using a pre-trained language model to generate content for expansion models leads to significant effective gains on complex topics. Specifically, I show that GRF significantly improves over traditional PRF models, explore the benefits of query expansion with real versus synthetic content, and demonstrate how using PLMs for ranking and generation context in multi-stage pipelines can set new state-of-the-art effectiveness.

7.1.3 Thesis Conclusion

In conclusion, this thesis supports the core hypothesis that integrating pre-trained language models into multi-stage query expansion pipelines significantly improves retrieval effectiveness compared to traditional expansion models. My research demonstrates that PLMs possess the ability to rank and generate contextually relevant content to help query expansion. Specifically, my ranking pipelines improve recall by 10% and generative pipelines improve recall by 5% respectively, achieving new state-of-the-art effectiveness. These capabilities prove particularly beneficial for complex queries requiring deeper reasoning and contextualisation.

7.2 Future Work

This thesis tackles the challenging task of document retrieval, specifically leveraging the new capabilities of pre-trained language models to develop novel query expansion pipelines. I show that PLMs can be incorporated into query expansion pipelines to generate and rank relevant content to contextualise complex queries. I demonstrate that these techniques result in new state-of-the-art retrieval effectiveness and lay the foundations for numerous avenues for future research. These research directions are discussed in this section.

- **Benchmarking PLMs** In this work, I leverage tuned encoder-decoder PLMs for ranking [150] and large zero-shot decoder-only PLMs for generation [22]. These models demonstrate significant improvements in retrieval effectiveness, highlighting the potential of PLMs in query expansion and ranking pipelines. Despite these gains, further research is needed to explore the impact of different model architectures, parameter scales, and training strategies on retrieval performance. For example, future work could investigate fine-tuning these models on domain-specific datasets and tasks, which should further enhance effectiveness. Understanding how to maximize the effectiveness of these model components as part of these pipelines for complex queries would further enhance understanding and performance.

- **New pipelines and agentic systems** Throughout this thesis, I demonstrate the benefits of multi-stage pipelines using PLMs for query expansion. For example, a re-ranker being leveraged before an expansion model in LEE or generation of facet-guided documents and PLM weighting in RASE. Nonetheless, current systems only scratch the surface of what is possible with flexible and rich pipelines that use PLMs to search, extract, and classify content. For example, leveraging advanced retrieval-augmented generation (RAG) techniques to address hallucination when generating synthetic context for query expansion (we found the basic RAG harmed effectiveness). There are also new threads such as "deep research" where the models more independently search and reason over the knowledge to answer complex questions - such techniques could be leveraged for query expansion. Future work could explore agentic systems that dynamically compose multi-stage PLM pipelines, integrating tools and search systems for generating, ranking, and refining content in a more autonomous manner for complex topics.

  Furthermore, our work leveraging PLMs for multi-stage is

- **Entity and knowledge representations** Throughout the creation and analysis of CODEC and the development of Latent Entity Expansion (LEE), I demonstrate that entities play a pivotal role in understanding and representing complex topics. Specifically, the success of LEE highlights how incorporating words and entities into query expansion leveraging PLMs significantly improves retrieval effectiveness. Future research could explore new query-focused entity-centric representations to enhance query expansion further. For example, encoding these entity relationships in a more contextualized manner, such as using dense embeddings that capture semantic and relational nuances more effectively. Furthermore, by incorporating structured knowledge graphs, retrieval models could better understand entity interactions, hierarchies, and dependencies. This could enable richer query understanding, disambiguation, and more precise ranking of relevant documents. Lastly, future work could investigate hybrid approaches, combining term-based approaches, symbolic knowledge, and neural representations to create more robust entity-aware retrieval systems for complex queries.

- **Improving efficiency** While this thesis primarily focuses on enhancing retrieval effectiveness, an important future research direction is optimising efficiency in query expansion pipelines leveraging PLMs. Improving efficiency is crucial for scaling retrieval systems to large datasets and real-world applications with latency constraints. This thesis demonstrates that adaptive expansion enhances efficiency by reducing re-ranking by storing query-specific relevance scores in a multi-pass retrieval system. Similar techniques, such as pre-computing partial results and relevance estimation, could be applied to other pipelines. Additionally, exploring trade-offs between large, high-performing models and smaller, more efficient alternatives could inform optimal model selection for various retrieval datasets and tasks. Model distillation—compressing large PLMs into smaller, parameter-efficient models—could reduce computational overhead and latency. Finally, investigating efficient indexing strategies, could further accelerate retrieval while maintaining accuracy.

Appendix A

Complex criterion on current datasets

Table A.1: Complex criterion annotation on TREC Deep Learning 2019

id	Query	Multifaceted	Entities	Comprehension	Knowledge	Complex
19335	anthropological definition of environment	0	1	1	1	1
47923	axon terminals or synaptic knob definition	1	1	2	2	0
87181	causes of left ventricular hypertrophy	1	1	1	1	1
87452	causes of military suicide	2	2	2	2	2
104861	cost of interior concrete flooring	0	0	0	1	0
130510	definition declaratory judgment	0	1	1	1	0
131843	definition of a sigmet	0	0	0	0	0
146187	difference between a mcdouble and a double cheeseburger	1	1	0	1	0
148538	difference between rn and bsn	1	0	1	1	1
156493	do goldfish grow	0	0	0	0	0
182539	example of monotonic function	0	0	1	0	0
183378	exons definition biology	0	1	2	1	1
207786	how are some sharks warm blooded	0	1	1	1	1
264014	how long is life cycle of flea	0	0	0	0	0
287683	how many liberty ships were built in brunswick	0	0	1	1	0
359349	how to find the midsegment of a trapezoid	0	0	1	0	0
405717	is cdg airport in main paris	0	0	0	0	0
443396	lps laws definition	0	0	1	0	0
451602	medicare's definition of mechanical ventilation	0	1	1	1	0
489204	right pelvic pain causes	1	1	1	1	1
490595	rsa definition key	0	1	1	0	0
527433	types of dysarthria from cerebral palsy	1	1	1	1	1
573724	what are the social determinants of health	2	2	2	2	2
833860	what is the most popular food in switzerland	0	0	0	0	0
855410	what is theraderm used for	0	0	1	0	0
915593	what types of food can you cook sous vide	1	1	1	1	1
962179	when was the salvation army founded	0	0	0	0	0
1037798	who is robert gray	0	0	0	0	0
1063750	why did the us volunterilay enter ww1	2	2	2	2	2
1103812	who formed the commonwealth of independent states	0	1	0	0	0
1106007	define visceral?	0	0	0	0	0
1110199	what is wifi vs bluetooth	1	1	1	0	1
1112341	what is the daily life of thai people	2	1	1	2	1
1113437	what is physical description of spruce	1	1	0	1	0
1114646	what is famvir prescribed for	0	1	0	0	0
1114819	what is durable medical equipment consist of	1	1	1	1	1
1115776	what is an aml surveillance analyst	1	1	1	1	0
1117099	what is a active margin	0	1	1	0	0
1121402	what can contour plowing reduce	1	1	0	1	0
1124210	tracheids are part of _____.	0	1	1	0	0
1129237	hydrogen is a liquid below what temperature	0	0	0	0	0
1132213	how long to hold bow in yoga	0	0	0	0	0
1133167	how is the weather in jamaica	0	0	0	1	0

Table A.2: Complex criterion annotation on TREC Deep Learning 2020

id	Query	Multifaceted	Entities	Comprehension	Knowledge	Complex
42255	average salary for dental hygienist in nebraska	0	0	0	0	0
47210	average wedding dress alteration cost	0	0	0	0	0
67316	can fever cause miscarriage early pregnancy	1	1	2	1	1
135802	definition of laudable	0	0	0	0	0
156498	do google docs auto save	0	0	0	0	0
169208	does mississippi have an income tax	0	1	0	0	0
174463	dog day afternoon meaning	0	1	0	0	0
258062	how long does it take to remove wisdom tooth	0	1	0	0	0
324585	how much money do motivational speakers make	0	1	0	1	0
330975	how much would it cost to install my own wind turbine	1	1	1	1	1
332593	how often to button quail lay eggs	0	0	0	0	0
336901	how old is vanessa redgrave	0	0	0	0	0
673670	what is a alm	0	1	0	0	0
701453	what is a statutory deed	0	1	1	0	0
730539	what is chronometer who invented it	0	1	0	1	0
768208	what is mamey	0	1	0	0	0
877809	what metal are hip replacements made of	1	1	1	1	1
911232	what type of conflict does della face in o, henry the gift of the magi	0	0	1	0	0
938400	when did family feud come out?	0	0	0	0	0
940547	when did rock n roll begin?	0	1	0	1	0
997622	where is the show shameless filmed	0	0	0	0	0
1030303	who is aziz hashim	0	1	0	0	0
1037496	who is rep scalise?	0	1	0	0	0
1043135	who killed nicholas ii of russia	0	1	0	1	0
1049519	who said no one can make you feel inferior	0	0	0	0	0
1051399	who sings monk theme song	0	0	0	0	0
1056416	who was the highest career passer rating in the nfl	0	1	0	0	0
1064670	why do hunters pattern their shotguns?	0	1	1	0	0
1071750	why is pete rose banned from hall of fame	0	0	1	0	0
1103153	who is thomas m cooley	0	1	0	0	0
1105792	define: geon	0	1	0	0	0
1108729	what temperature and humidity to dry sausage	0	1	1	1	1
1109707	what medium do radio waves travel through	0	0	1	0	0
1113256	what is reba mcentire's net worth	0	0	0	0	0
1115210	what is chaff and flare	0	1	1	0	0
1116380	what is a nonconformity? earth science	0	1	1	0	0
1119543	what does a psychological screening consist of for egg donors	2	2	2	1	2
1122767	what amino produces carnitine	0	1	1	0	0
1127540	meaning of shebang	0	0	0	0	0
1131069	how many sons robert kraft has	0	1	0	0	0
1132532	average annual income data analyst	0	0	0	0	0
1136043	difference between a hotel and motel	0	0	0	0	0
1136047	difference between a company's strategy and business model is	1	2	1	2	1
1136769	why does lacquered brass tarnish	0	1	1	0	1
1136962	why did the ancient egyptians call their land kemet, or black land?	0	1	0	1	0

Table A.3: Complex criterion annotation on DL-HARD

id	Query	Multifaceted	Entities	Comprehension	Knowledge	Complex
19335	anthropological definition of environment	0	1	1	1	1
47923	axon terminals or synaptic knob definition	1	1	2	2	0
87452	causes of military suicide	2	2	2	2	2
182539	example of monotonic function	0	0	1	0	0
443396	lps laws definition	0	0	1	0	0
451602	medicare's definition of mechanical ventilation	0	1	1	1	0
489204	right pelvic pain causes	1	1	1	1	1
527433	types of dysarthria from cerebral palsy	1	1	1	1	1
915593	what types of food can you cook sous vide	1	1	1	1	1
1103812	who formed the commonwealth of independent states	0	1	0	0	0
1106007	define visceral?	0	0	0	0	0
1112341	what is the daily life of thai people	2	1	1	2	1
1114646	what is famvir prescribed for	0	1	0	0	0
190044	foods to detox liver naturally	0	2	1	1	1
264403	how long is recovery from a face lift and neck lift	1	1	1	0	0
315637	how much does it cost to go to alabama university	0	1	0	0	0
507445	symptoms of different types of brain bleeds	2	1	1	2	1
554515	what are achieved and ascribed statuses?	1	1	1	2	1
588587	what causes heavy metal toxins in your body	1	2	1	1	1
705609	what is a virus made of	0	1	1	1	1
966413	where are the benefits of cinnamon as a supplement?	1	0	1	2	1
1056204	who was the first steam boat operator	0	1	0	0	0
1108100	what type of movement do bacteria exhibit?	0	1	1	0	0
1108939	what slows down the flow of blood	1	1	1	2	1
67316	can fever cause miscarriage early pregnancy	1	1	2	1	1
174463	dog day afternoon meaning	0	1	0	0	0
332593	how often to button quail lay eggs	0	0	0	0	0
673670	what is a alm	0	1	0	0	0
730539	what is chronometer who invented it	0	1	0	1	0
877809	what metal are hip replacements made of	1	1	1	1	1
1049519	who said no one can make you feel inferior	0	0	0	0	0
1056416	who was the highest career passer rating in the nfl	0	1	0	0	0
1103153	who is thomas m cooley	0	1	0	0	0
1105792	define: geon	0	1	0	0	0
1109707	what medium do radio waves travel through	0	0	1	0	0
1136769	why does lacquered brass tarnish	0	1	1	0	1
86606	causes of gas in large intestine	2	2	2	1	1
88495	causes of stroke?	2	2	1	2	2
144862	did prohibition increased crime	1	1	2	1	1
166046	does ethambutol treat bone infection	0	1	1	0	0
177604	eating foods that are considered warm	1	1	1	1	1
537817	vitamin e anti scar	0	1	0	1	0
655914	what drives poaching	2	2	1	2	2
794429	what is sculpture shape space	0	0	0	0	0
801118	what is supplemental security income used for	1	1	1	1	1
883915	what other brain proteins can cause dementia	0	2	2	1	1
1065636	why do some places on my scalp feel sore	0	0	1	0	0
1117817	what does unauthorized act in writing mean	0	1	1	0	0
1133485	how does vitamin c helps	1	1	1	2	1
273695	how long will methadone stay in your system	0	1	1	0	0

Table A.4: Complex criterion annotation on TREC Robust 2004 (Topics 300-350)

Topic id	Description	Multifaceted	Entities	Comprehension	Knowledge	Complex
301	Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.	1	2	1	2	1
302	Is the disease of Poliomyelitis (polio) under control in the world?	2	2	2	2	2
303	Identify positive accomplishments of the Hubble telescope since it was launched in 1991.	2	2	1	2	2
304	Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.	2	2	2	2	2
305	Which are the most crashworthy, and least crashworthy, passenger vehicles?	1	2	1	1	1
306	How many civilian non-combatants have been killed in the various civil wars in Africa?	1	1	2	2	1
307	Identify hydroelectric projects proposed or under construction by country and location. Detailed description of nature, extent, purpose, problems, and consequences is desirable.	1	2	1	2	1
308	What are the advantages and/or disadvantages of tooth implants?	2	2	2	2	2
309	Evidence that rap music has a negative effect on young people.	2	2	2	2	2
310	Evidence that radio waves from radio towers or car phones affect brain cancer occurrence.	2	2	2	2	2
311	Document will discuss the theft of trade secrets along with the sources of information: trade journals, business meetings, data from Patent Offices, trade shows, or analysis of a competitor's products.	1	2	1	2	1
312	Document will discuss the science of growing plants in water or some substance other than soil.	1	1	1	1	1
313	Commercial uses of Magnetic Levitation.	2	2	1	2	2
314	Commercial harvesting of marine vegetation such as algae, seaweed and kelp for food and drug purposes.	1	2	1	2	1
315	How many fatal highway accidents are there each year that are not resolved as to cause.	0	1	0	1	0
316	A look at the roots and prevalence of polygamy in the world today.	2	2	1	2	2
317	Have regulations been passed by the FCC banning junk facsimile (fax)? If so, are they effective?	2	1	2	2	2
318	Aside from the United States, which country offers the best living conditions and quality of life for a U.S. retiree?	2	2	2	2	2
319	What research is ongoing for new fuel sources.	2	2	2	2	2
320	Fiber optic link around the globe (Flag) will be the world's longest undersea fiber optic cable. Who's involved and how extensive is the technology on this system. What problems exist?	2	2	2	2	2
321	Pertinent documents will reflect the fact that women continue to be poorly represented in parliaments across the world, and the gap in political power between the sexes is very wide, particularly in the Third World.	2	2	2	2	2
322	Isolate instances of fraud or embezzlement in the international art trade.	1	2	1	1	1
323	Find instances of plagiarism in the literary and journalistic worlds.	2	2	2	2	2
324	Define Argentine and British international relations	2	2	2	2	2
325	Describe a cult by name and identify the cult members' activities in their everyday life.	1	2	1	2	1
326	Any report of a ferry sinking where 100 or more people lost their lives.	1	2	1	2	1
327	Identify a country or a city where there is evidence of human slavery being practiced in the eighties or nineties.	1	2	1	2	1
328	Identify an individual that has been beatified by the Pope.	1	1	1	1	1
329	Mexico City has the worst air pollution in the world. Pertinent Documents would contain the specific steps Mexican authorities have taken to combat this deplorable situation.	2	2	2	2	2
330	This query is looking for examples of cooperation or friendly ties between Iran and Iraq, or ways in which the two countries could be considered allies.	2	2	2	2	2
331	What criticisms have been made of World Bank policies, activities or personnel?	2	2	2	2	2
332	This query is looking for investigations that have targeted evaders of U.S. income tax.	1	1	1	2	1
333	Determine the reasons why bacteria seems to be winning the war against antibiotics and rendering antibiotics now less effective in treating diseases than they were in the past.	2	2	2	2	2
334	Determine the usefulness and effectiveness of continuing to maintain export controls on encryption software.	2	2	2	2	2
335	Identify the problems, and solutions to those problems, which arise in the relationships among biological parents, adoptive parents, and the child or children involved.	2	2	2	2	2
336	A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.	2	2	1	2	2
337	What research has been done on viral hepatitis and what progress has been made in its treatment?	2	2	2	2	2
338	What adverse effects have people experienced while taking aspirin repeatedly?	2	1	1	2	1
339	What drugs are being used in the treatment of Alzheimer's Disease and how successful are they?	2	2	2	2	2
340	Identify any actions being taken to propel the nations of the world toward a treaty banning the production, transfer and use of land mines.	2	2	2	2	2
341	A relevant document would discuss how effective government orders to better scrutinize passengers and luggage on international flights and to step up screening of all carry-on baggage has been.	2	1	2	2	2
342	The end of the Cold War seems to have intensified economic competition and has started to generate serious friction between nations as attempts are made by diplomatic personnel to acquire sensitive trade and technology information or to obtain information on highly classified industrial projects. Identify instances where attempts have been made by personnel with diplomatic status to obtain information of this nature.	1	2	1	2	1
343	Identify instances where a civilian policeman has been killed either during performance of his duty or because of other association with this occupation, e.g., killed for the gun, to keep from testifying, etc.	1	1	1	1	1
344	The availability of E-mail to many people through their job or school affiliation has allowed for many efficiencies in communications but also has provided the opportunity for abuses. What steps have been taken world-wide by those bearing the cost of E-mail to prevent excesses?	2	1	2	2	2
345	Health studies primarily in the U.S. have caused reductions in tobacco sales here, but the economic impact has caused U.S. tobacco companies to look overseas for customers. What impact have the health and economic factors had overseas?	2	2	2	2	2
346	There has long been a call for standards in U.S. education, these calls frequently citing the superiority of foreign school systems. Are there many countries outside the U.S. which have standards for pre-teen students? If so, which are those countries and what standards have been set?	1	2	1	2	1
347	The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?	2	2	2	2	2
348	Is the fear of open or public places (Agoraphobia) a widespread disorder or relatively unknown?	1	1	1	1	1
349	Document will discuss the chemical reactions necessary to keep living cells healthy and/or producing energy.	0	1	0	1	0
350	Is it hazardous to the health of individuals to work with computer terminals on a daily basis?	1	2	2	2	2

Table A.5: Complex criterion annotation on TREC Robust 2004 (Topics 351-400)

Topic id	Description	Multifaceted	Entities	Comprehension	Knowledge	Complex
351	What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?	1	2	1	2	1
352	What impact has the Chunnel had on the British economy and/or the life style of the British?	2	2	1	2	2
353	Identify systematic explorations and scientific investigations of Antarctica, current or planned.	1	1	1	1	1
354	Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work.	1	2	1	1	1
355	Identify documents discussing the development and application of spaceborne ocean remote sensing.	2	1	1	2	1
356	Identify documents discussing the use of estrogen by postmenopausal women in Britain.	1	1	2	2	1
357	Identify documents discussing international boundary disputes relevant to the 200-mile special economic zones or 12-mile territorial waters subsequent to the passing of the "International Convention on the Law of the Sea".	2	2	1	1	1
358	What role does blood-alcohol level play in automobile accident fatalities?	2	1	2	2	2
359	Are there reliable and consistent predictors of mutual fund performance?	2	1	2	2	2
360	What are the benefits, if any, of drug legalization?	2	2	2	2	2
361	Identify documents that discuss clothing sweatshops.	1	1	1	1	1
362	Identify incidents of human smuggling.	0	1	0	1	0
363	What disasters have occurred in tunnels used for transportation?	1	1	1	1	1
364	Identify documents discussing cases where rabies have been confirmed and what, if anything, is being done about it.	2	1	2	2	2
365	What effects have been attributed to El Nino?	2	2	2	2	2
366	What are the industrial or commercial uses of cyanide or its derivatives?	2	2	2	2	2
367	What modern instances have there been of old fashioned piracy, the boarding or taking control of boats?	1	1	1	2	1
368	Identify documents that discuss in vitro fertilization.	1	1	1	1	1
369	What are the causes and treatments of anorexia nervosa and bulimia?	2	2	2	2	2
370	What are the laws dealing with the quality and processing of food, beverages, or drugs?	2	2	1	2	1
371	What is the extent of health insurance coverage of holistic or other non-traditional medicine/medical treatments (for example, acupuncture)?	2	2	2	2	2
372	Identify documents that discuss the growth of Native American casino gambling.	2	2	2	2	2
373	Identify documents that discuss the concerns of the United States regarding the export of encryption equipment.	2	2	2	2	2
374	Identify and provide background information on Nobel prize winners.	1	2	1	1	1
375	What is the status of research on hydrogen as a feasible energy source?	2	2	2	2	2
376	What types of cases were heard by the World Court (International Court of Justice)?	1	1	1	1	1
377	Identify documents that discuss the renewed popularity of cigar smoking.	1	1	1	2	1
378	Identify documents that discuss opposition to the introduction of the euro, the European currency.	2	2	2	2	2
379	Identify documents that discuss mainstreaming children with physical or mental impairments.	2	2	2	2	2
380	Identify documents that discuss medical treatment of obesity.	2	2	2	2	2
381	What forms of alternative medicine are being used in the treatment of illnesses or diseases and how successful are they?	2	2	2	2	2
382	Identify documents that discuss the use of hydrogen as a fuel for piston driven automobiles (safe storage a concern) or the use of hydrogen in fuel cells to generate electricity to drive the car.	2	2	2	2	2
383	Identify drugs used in the treatment of mental illness.	1	2	1	1	1
384	Identify documents that discuss the building of a space station with the intent of colonizing the moon.	2	2	2	2	2
385	Identify documents that discuss the current status of hybrid automobile engines, (i.e., cars fueled by something other than gasoline only).	1	2	1	1	1
386	What methods are currently utilized or anticipated in the teaching of disabled children?	2	2	2	2	2
387	Identify documents that discuss effective and safe ways to permanently handle long-lived radioactive wastes.	2	2	2	2	2
388	Identify documents that discuss the use of organic fertilizers (composted sludge, ash, vegetable waste, microorganisms, etc.) as soil enhancers.	1	1	1	1	1
389	What specific entities have been accused of illegal technology transfer such as: selling their products, formulas, etc. directly or indirectly to foreign entities for other than peaceful purposes?	1	1	1	1	1
390	Find documents that discuss issues associated with so-called "orphan drugs", that is, drugs that treat diseases affecting relatively few people.	2	2	2	1	2
391	Identify documents that discuss the impact of the cost of research and development (R&D) on the price of drugs.	2	2	2	2	2
392	What are the applications of robotics in the world today?	2	2	2	2	2
393	Identify documents that discuss mercy killings.	2	1	2	2	2
394	Identify documents that discuss the education of children at home (home schooling).	2	1	1	2	1
395	Provide examples of successful attempts to attract tourism as a means to improve a local economy.	2	1	1	1	1
396	Identify documents that discuss sick building syndrome or building-related illnesses.	1	1	1	1	1
397	Identify documents that discuss the reasons for automobile recalls.	1	2	1	1	1
398	Identify documents that discuss the European Conventional Arms Cut as it relates to the dismantling of Europe's arsenal.	1	2	1	2	1
399	Identify documents that discuss the activities or equipment of oceanographic vessels.	1	1	1	2	1

Table A.6: Complex criterion annotation on TREC Robust 2004 (Topics 401-450)

Topic id	Description	Multifaceted	Entities	Comprehension	Knowledge	Complex
401	What language and cultural differences impede the integration of foreign minorities in Germany?	2	2	2	2	2
402	What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?	2	2	2	2	2
403	Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).	2	2	2	2	2
404	How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?	1	2	1	1	1
405	What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?	2	2	1	1	1
406	What is being done to treat the symptoms of Parkinson's disease and keep the patient functional as long as possible?	2	2	2	2	2
407	What is the impact of poaching on the world's various wildlife preserves?	2	2	1	2	2
408	What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?	2	2	1	1	1
409	What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?	2	2	2	2	2
410	Who is involved in the Schengen agreement to eliminate border controls in Western Europe and what do they hope to accomplish?	2	2	2	2	2
411	Find information on shipwreck salvaging: the recovery or attempted recovery of treasure from sunken ships.	1	1	1	1	1
412	What security measures are in effect or are proposed to go into effect in airports?	2	1	2	2	2
413	What are new methods of producing steel?	2	1	2	1	1
414	How much sugar does Cuba export and which countries import it?	1	1	1	1	1
415	What is known about drug trafficking in the "Golden Triangle", the area where Burma, Thailand and Laos meet?	2	2	2	2	2
416	What is the status of The Three Gorges Project?	1	1	1	1	1
417	Find ways of measuring creativity.	2	1	1	2	2
418	In what ways have quilts been used to generate income?	1	1	1	1	1
419	What new uses have been developed for old automobile tires as a means of tire recycling?	1	1	1	1	1
420	How widespread is carbon monoxide poisoning on a global scale?	1	1	1	2	1
421	How is the disposal of industrial waste being accomplished by industrial management throughout the world?	2	2	2	2	2
422	What incidents have there been of stolen or forged art?	1	1	0	1	1
423	Find references to Milosevic's wife, Mirjana Markovic.	0	1	0	0	0
424	Give examples of alleged suicides that aroused suspicion of the death actually being murder.	1	1	1	1	1
425	What counterfeiting of money is being done in modern times?	1	1	1	2	1
426	Provide information on the use of dogs worldwide for law enforcement purposes.	1	1	1	2	1
427	Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.	2	2	1	2	2
428	Do any countries other than the U.S. and China have a declining birth rate?	2	2	1	1	1
429	Identify outbreaks of Legionnaires' disease.	1	1	1	1	1
430	Identify instances of attacks on humans by Africanized (killer) bees.	1	1	0	1	1
431	What are the latest developments in robotic technology?	2	2	2	2	2
432	Do police departments use "profiling" to stop motorists?	1	1	1	1	1
433	Is there contemporary interest in the Greek philosophy of stoicism?	1	1	1	2	1
434	What is the state of the economy of Estonia?	1	2	1	2	1
435	What measures have been taken worldwide and what countries have been effective in curbing population growth?	2	2	2	2	2
436	What are the causes of railway accidents throughout the world?	2	2	2	2	2
437	What has been the experience of residential utility customers following deregulation of gas and electric?	2	2	2	2	2
438	What countries are experiencing an increase in tourism?	1	2	1	1	1
439	What new inventions or scientific discoveries have been made?	2	2	0	1	1
440	What steps are being taken by governments or corporations to eliminate abuse of child labor?	2	2	2	2	2
441	How do you prevent and treat Lyme disease?	2	2	2	2	2
442	Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.	1	1	0	0	0
443	What is the extent of U.S. (government and private) investment in sub-Saharan Africa?	1	2	1	2	1
444	What are the potential uses for supercritical fluids as an environmental protection measure?	2	2	2	2	2
445	What other countries besides the United States are considering or have approved women as clergy persons?	1	2	1	1	1
446	Where are tourists likely to be subjected to acts of violence causing bodily harm or death?	2	2	2	2	2
447	What new developments and applications are there for the Stirling engine?	1	1	1	1	1
448	Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.	1	1	2	1	1
449	What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?	2	2	2	2	2
450	How significant a figure over the years was the late Jordanian King Hussein in furthering peace in the Middle East?	2	2	2	2	2

Table A.7: Complex criterion annotation on TREC Robust 2004 (Topics 601-650)

Topic id	Description	Multifaceted	Entities	Comprehension	Knowledge	Complex
601	What is the effect of Turkish river control projects on Iraqi water resources?	2	2	2	2	2
602	Retrieve information regarding the process by which the Czech and Slovak republics of Czechoslovakia established separate sovereign countries.	1	2	1	2	1
603	Retrieve documents regarding U.S. lawsuits against the tobacco industry for causing health problems and/or death from cigarettes.	1	2	1	1	1
604	What evidence is there to link tick-borne Lyme disease with arthritis?	2	2	2	2	2
605	What are the pros and cons of Great Britain’s universal health care system?	2	2	2	2	2
606	Find documents that discuss banning leg traps used to capture animals.	1	1	1	1	1
607	What progress is being made in the effort to map and sequence the human genetic code?	1	1	2	1	1
608	Find articles that discuss the pros and cons of taxing U.S. social security benefits.	2	2	2	2	2
609	Find documents that discuss per capita consumption of alcohol by political entity—country, state, province, etc.	1	1	1	1	1
610	Find claims made by U.S. small businesses regarding the adverse impact on their businesses of raising the minimum wage.	2	1	2	2	2
611	What violent activities have Kurds, or members of the Workers Party of Kurdistan (PKK), carried out in Germany?	1	2	1	2	1
612	What has been the outcome for the pro-independence protesters in Tibet who were arrested by Chinese authorities?	1	2	1	1	1
613	How were pieces of the Berlin wall disposed of after their removal?	1	1	1	1	1
614	Find information about the first genetically modified food product to go on the market, Flavr Savr (also Flavor Saver) Tomato developed by Calgene.	0	1	1	1	1
615	What is the extent of U.S. raw timber exports to Asia, and what effect do these exports have on the U.S. lumber industry?	2	2	2	2	2
616	What is the history and extent of Volkswagen production in Mexico?	1	1	1	2	1
617	What effect has the reduction of Russian support had on the Cuban economy?	2	1	2	2	2
618	Find documents that describe the death of Iranian President Ayatollah Khomeini and the ramifications of his death.	2	2	2	2	2
619	What part did Winnie Mandela herself play in the kidnapping, beating and murder scandal in South Africa in December 1988 through January 1989?	2	2	2	2	2
620	How has France responded to protests against its nuclear testing in the South Pacific?	2	2	2	2	2
621	What are the arguments for and against Great Britain’s approval of women being ordained as Church of England priests?	2	2	2	2	2
622	Identify companies or corporations that have been accused or indicted of price fixing including the product or type of product involved.	1	2	1	1	1
623	Gather any information that mentions ricin, sarin, soman, or anthrax as a toxic chemical used as a weapon.	1	2	1	2	1
624	What are the pros and cons of developing the Strategic Defense Initiative (SDI) also known as "Star Wars"?	2	2	2	2	2
625	Identify documents that provide information on the arrest and/or conviction of the bombers of the World Trade Center (WTC) in February 1993.	1	1	1	1	1
626	Find reports of human stampedes that have resulted in 20 or more deaths.	1	2	1	1	1
627	What steps are being taken by the U.S. to help Russia solve the food crisis in Russia?	2	2	2	2	2
628	What justification was used by the U.S. government to invade Panama, and why did some oppose the invasion?	2	2	2	2	2
629	What is the incidence of violent attacks on abortion clinics and the doctors and staff of the clinics by anti-abortionists?	1	1	1	2	1
630	Retrieve documents containing information about the symptoms of individuals suffering from ’Gulf War Syndrome’ as a result of serving in the Gulf War.	2	2	2	2	2
631	Find documents relating to the election of Nelson Mandela as president of the Republic of South Africa.	1	2	1	2	1
632	What are the major tin mining countries in southeast Asia?	1	2	1	2	1
633	What is the history of the Welsh devolution movement?	2	2	2	2	2
634	How many deaths are attributed to having taken tainted L-tryptophan dietary supplements?	1	1	2	1	1
635	What are the arguments for and against doctor assisted suicide in the U.S.?	2	2	2	2	2
636	Find documents that discuss reasons why people may be exempted from serving on a jury.	1	1	1	1	1
637	What are the pros and cons of adults using human growth hormone (HGH)?	2	1	2	2	2
638	Find documents that discuss freed prisoners who have been wrongfully convicted based on faulty forensic evidence, poor police work, or false testimony.	2	1	1	2	1
639	What factors contributed to the growth of consumer on-line shopping?	2	2	2	2	2
640	What are the maternity leave policies of various governments?	1	2	1	2	1
641	What was the impact of the Exxon Valdez oil spill on the marine life and wildlife of the area?	2	2	2	2	2
642	What happened to protesters arrested in connection with the Tiananmen Square demonstrations in Beijing in the spring of 1989?	2	2	2	2	2
643	What harm have power dams in the Pacific northwest caused to salmon fisheries?	2	2	2	2	2
644	Identify documents that discuss exotic species of animals that are imported into the U.S. or U.K.	1	1	1	1	1
645	Find documents that discuss the financial impact of software piracy upon the software-producing industry.	1	1	2	1	1
646	Find documents that discuss an increase in the number of people receiving food stamp benefits.	2	1	2	2	2
647	Has the use of windmill technology to generate electricity been economically productive?	2	2	2	2	2
648	Identify documents that discuss details of a family leave law, such as how long, compensation, if any, for what reason allowed, etc.	2	1	1	2	1
649	How do computers get infected by computer viruses?	2	1	2	2	2
650	Identify individuals or corporations that have been indicted on charges of tax evasion of more than two million dollars in the U.S. or U.K.	1	1	0	1	1

Table A.8: Complex criterion annotation on TREC Robust 2004 (Topics 651-700)

Topic id	Description	Multifaceted	Entities	Comprehension	Knowledge	Complex
651	How is the ethnic make-up of the U.S. population changing?	2	2	2	1	2
652	What was the OIC's involvement in the Balkans in 1990-94?	1	2	2	2	2
653	Find documents that describe the activities of ETA, the Basque terrorist organization, in Spain.	1	1	1	1	1
654	What are the advantages and disadvantages of same-sex schools?	2	1	2	2	2
655	How is Attention Deficit Disorder (ADD) diagnosed and treated in young children?	2	2	2	2	2
656	How are young children being protected against lead poisoning from paint and water pipes?	2	1	1	2	1
657	Has prayer in U.S. schools been banned completely?	1	1	1	1	1
658	Find documents that discuss teenage pregnancy in the United States: the birth rate for teenage mothers, causes and results of teenage pregnancies, and steps taken to reduce the number of teenage pregnancies.	2	1	2	2	2
659	What standards do cruise ships use for health and safety maintenance?	1	2	1	2	1
660	Find information about whale watching off the coast of California.	1	1	1	1	1
661	What are the causes and treatments for melanoma?	2	2	2	2	2
662	How can consumers protect against telemarketers?	1	1	1	1	1
663	What were the health effects of Vietnam veterans' exposure to Agent Orange?	2	2	2	2	2
664	What are the plans for a National Museum of the American Indian?	1	1	1	1	1
665	How extensive is poverty in sub-Saharan Africa?	1	1	1	1	1
666	Find documents that discuss the impact Prime Minister Margaret Thatchers' resignation may have on U.S. and U.K. relations.	2	2	2	2	2
667	Find documents that discuss the increasing trend toward creation of unmarried-partner households in the U.S.	2	1	2	2	2
668	What is the relationship between poverty and disease?	2	1	2	2	2
669	What were the causes for the Islamic Revolution relative to relations with the U.S.?	2	2	2	2	2
670	Why is there such apathy in U.S. elections?	2	1	2	2	2
671	Find documents that cite the specific benefits the Salvation Army provides those in need.	1	2	1	2	1
673	What factors led to the withdrawal of Soviet troops from Afghanistan?	2	2	2	2	2
674	Has Greenpeace been prosecuted or its members arrested for any of its actions?	1	2	1	1	1
675	Find information regarding training for Olympic swim meets.	1	1	0	1	1
676	Find information on poppy cultivation and export worldwide.	1	1	1	2	1
677	What efforts are being made to stabilize the Leaning Tower of Pisa, and how successful have the efforts been?	2	1	2	1	2
678	Find information on joint/shared custody's impact on children.	2	2	2	2	2
679	Find documents that discuss the U.S. debate about the opening of sealed adoption records to adoptees.	2	1	2	2	2
680	Find documents that discuss how the use of Spanish in U.S. schools has improved the lives of Mexican immigrants.	2	1	2	2	2
681	Where are wind power installations located?	1	1	1	1	1
682	What is being done to teach English to recently admitted adult immigrants?	1	1	1	1	1
683	Find information on the breakup of Czechoslovakia into the Czech Republic and Slovakia and its social and political impact on the two countries' people.	2	2	2	2	2
684	What businesses or government entities give medical or other benefits to part-time workers?	1	2	1	1	1
685	How are Oscar winners selected?	0	1	0	1	0
686	What are the negative impacts of Argentina's policy of pegging their peso to the U.S. dollar?	2	2	2	2	2
687	What businesses and industries form the basis of the economy of Northern Ireland?	1	2	1	2	2
688	What bias exists in the media of countries other than the U.S.?	2	2	2	2	2
689	To which countries does the U.S. provide aid to support family planning, and for which countries has the U.S. refused or limited support?	1	1	1	2	1
690	Find documents which describe an advantage in hiring potential or increased income for graduates of U.S. colleges.	2	1	2	2	2
691	What are the objections to the practice of "clear-cutting"?	2	2	2	2	2
692	Find information on prostate cancer detection and treatment.	1	1	1	2	1
693	What has been the effect of the electronic media on the newspaper industry?	2	2	2	2	2
694	How do you make a compost pile?	0	1	0	1	0
695	What is the usual sentence for those convicted of white collar crimes?	1	1	1	1	1
696	Find documents that discuss the safety of or the hazards of cosmetic plastic surgery.	2	1	1	2	1
697	What are working conditions and pay for U.S. air traffic controllers?	1	1	1	1	1
698	What are literacy rates in African countries?	1	2	1	1	1
699	What are the pros and cons of term limits?	2	2	2	2	2
700	What are the arguments for and against an increase in gasoline taxes in the U.S.?	2	2	2	2	2

Table A.9: Complex criterion annotation on CODEC

Topic id	Query	Facets	Entities	Comprehension	Knowledge	Complex
economics-1	How has the UK's Open Banking Regulation benefited challenger banks?	2	2	2	2	2
economics-2	What technological challenges does Bitcoin face to becoming a widely used currency?	2	2	2	2	2
economics-3	Why are many commentators arguing NFTs are the next big investment category?	2	1	2	2	2
economics-4	Why has value investing underperformed growth over the last decade?	2	2	2	2	2
economics-6	Why are some economists sceptical about the EU's monetary union without a shared fiscal system?	2	2	2	2	2
economics-8	How is the push towards electric cars impacting the demand for raw materials?	2	2	2	2	2
economics-12	What are the common problems or criticisms aimed at public sector enterprises?	2	1	2	2	2
economics-13	Why do many economists argue against fixed exchange rates?	2	2	2	2	2
economics-17	Why is scaling a hardware business more capital intensive than a software business?	2	2	2	2	2
economics-18	Was the crash that followed the dot-com bubble an overreaction considering the ultimate success of the internet?	2	2	2	2	2
economics-19	Is diversification the best strategy to get rich?	2	1	2	2	2
economics-20	Are private capital markets so plentiful that there is no need for startups to IPO?	2	2	2	2	2
economics-21	How much of a threat are ETFs to actively-managed Asset Managers?	1	1	2	2	2
economics-23	Offering non-accounting services arguably creates a conflict of interest for the Big Four. Is this the reason for their inability to uncover recent financial scandals?	2	2	2	2	2
history-1	Would the United Kingdom have been ready for WWII without the time gained through Appeasement?	2	2	2	2	2
history-6	What were the lasting social changes brought about by the Black Death?	2	2	2	2	2
history-11	Francesco Petrarch coined the term Dark Ages. Was this a fair description of this period of history?	2	2	2	2	2
history-12	Why did England have a reformation of religion under Henry VIII?	2	2	2	2	2
history-13	Why did Japan attack the United States at Pearl Harbour?	2	2	2	2	2
history-15	Why did Winston Churchill lose the 1945 General Election after winning World War II?	2	2	2	2	2
history-16	Would Adolf Hitler have won World War II if he had invaded England instead of the Soviet Union?	2	2	2	2	2
history-17	How significant was Smallpox in the Spanish defeat of the Aztecs?	2	2	2	2	2
history-18	Were the Crusades driven by religious devotion or political and economic gain?	2	2	2	2	2
history-19	How close did the world come to nuclear war during the Cuban Missile Crisis?	2	2	2	2	2
history-20	How vital was French support during the American Revolutionary War?	2	2	2	2	2
history-23	Why did the Treaty of Versailles fail?	2	2	2	2	2
history-24	How did American media coverage of the war in Vietnam shape public attitudes and opinions?	2	2	2	2	2
history-25	How responsible was Rasputin for the fall of the Romanov dynasty?	2	2	2	2	2
politics-1	Is Scottish Independence inevitable?	2	2	2	2	2
politics-3	Are nations with nuclear capabilities a threat to world peace or a deterrent?	2	2	2	2	2
politics-4	What has been the impact of the Umbrella Movement on Hong Kong politics?	2	2	2	2	2
politics-5	How did Colin Kaepernick impact the political discourse about racism in the United States?	2	2	1	2	2
politics-6	Would Obama have won the 2012 US Presidential Elections without the Latino vote?	2	2	2	2	2
politics-7	Should the Electoral College system in United States elections be abolished?	2	2	2	2	2

Appendix B

Dense and learned sparse GRF effectivenss per subtask

Table B.1: Effectiveness of **dense** generative relevance feedback based on different generation subtasks on target datasets, *bold* depicts best system.

	Robust04 - Title			**CODEC**		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.494	0.250	0.667	0.373	0.227	0.757
Entities	0.492	0.249	0.663	0.377	0.230	0.767
COT-Keywords	0.482	0.244	0.652	0.377	0.236	0.788
COT-Entities	0.479	0.243	0.661	0.366	0.234	0.778
Queries	0.502	0.259	0.681	0.359	0.231	0.786
Summary	0.504	0.264	0.675	0.381	0.244	0.801
Facts	0.513	0.266	0.689	0.372	0.246	0.792
Document	0.510	0.264	0.676	0.397	0.251	0.790
Essay	0.511	0.266	0.692	**0.408**	0.255	0.794
News	0.513	0.268	0.693	0.379	0.254	0.802
Combined	**0.517**	0.276	0.700	0.385	**0.261**	**0.821**

Table B.2: Effectiveness of **dense** generative relevance feedback based on different generation subtasks on target datasets, *bold* depicts best system.

	DL-19			**DL-20**		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.687	0.365	0.670	0.623	0.433	0.788
Entities	0.686	0.362	0.664	0.594	0.409	0.773
COT-Keywords	0.701	0.380	0.684	0.583	0.395	0.770
COT-Entities	0.692	0.399	0.710	0.553	0.384	0.765
Queries	0.707	0.380	0.697	0.620	0.439	0.789
Summary	0.684	0.399	0.708	0.644	0.454	0.795
Facts	0.634	0.376	0.695	0.628	0.451	0.811
Document	**0.690**	0.411	0.710	0.597	0.439	0.803
Essay	0.663	0.381	0.687	0.626	0.445	0.801
News	0.671	0.392	0.697	0.614	0.443	0.796
Combined	0.683	**0.418**	**0.743**	**0.634**	**0.457**	**0.812**

Table B.3: Effectiveness of **learned sparse** generative relevance feedback based on different generation subtasks on target datasets, *bold* depicts best system.

	Robust04 - Title			**CODEC**		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.427	0.234	0.703	0.317	0.191	0.744
Entities	0.417	0.229	0.693	0.321	0.193	0.739
CoT-Keywords	0.405	0.223	0.679	0.313	0.200	0.753
CoT-Entities	0.412	0.229	0.687	0.308	0.197	0.765
Queries	0.432	0.240	0.703	0.333	0.200	0.759
Summary	0.453	0.246	0.702	0.345	0.212	0.762
Facts	0.451	0.251	0.701	0.324	0.205	0.757
Document	0.438	0.243	0.697	0.332	0.216	0.766
Essay	0.455	0.252	0.709	0.343	0.222	0.784
News	0.462	0.256	0.706	0.318	0.211	0.775
Combined	**0.462**	**0.265**	**0.730**	**0.337**	**0.222**	**0.785**

Table B.4: Effectiveness of **learned sparse** generative relevance feedback based on different generation subtasks on target datasets, *bold* depicts best system.

	DL-19			**DL-20**		
	NDCG@10	MAP	R@1k	NDCG@10	MAP	R@1k
Keywords	0.594	0.342	0.677	0.533	0.383	0.803
Entities	0.600	0.323	0.662	0.533	0.366	0.789
COT-Keywords	0.583	0.356	0.680	0.489	0.341	0.800
COT-Entities	0.585	0.369	0.702	0.480	0.338	0.794
Queries	0.595	0.334	0.674	0.515	0.366	0.800
Summary	0.605	0.392	0.721	0.566	0.412	0.829
Facts	0.554	0.351	0.696	**0.594**	**0.432**	0.830
Document	0.623	0.391	0.703	0.549	0.400	0.815
Essay	0.619	0.373	0.685	0.529	0.383	0.812
News	0.585	0.364	0.699	0.539	0.385	0.818
Combined	**0.642**	**0.407**	**0.732**	0.553	0.415	**0.839**

Bibliography

[1] Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series p. 189 (2004)

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[3] Ahmad, W., Chakraborty, S., Ray, B., Chang, K.: Unified pre-training for program understanding and generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021)

[4] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics. pp. 1638–1649 (2018)

[5] Alaparthi, S., Mishra, M.: Bert: A sentiment analysis odyssey. Journal of Marketing Analytics **9**(2), 118–126 (2021)

[6] Alharbi, A., Stevenson, M.: Refining boolean queries to identify relevant studies for systematic review updates. Journal of the American Medical Informatics Association **27**(11), 1658–1666 (2020)

[7] Allen, K., Berry, M.M., Luehrs Jr, F.U., Perry, J.W.: Machine literature searching viii. operational criteria for designing information retrieval systems. American Documentation (pre-1986) **6**(2), 93 (1955)

[8] Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)

[9] Angles, R., Gutierrez, C.: The expressive power of sparql. In: International Semantic Web Conference. pp. 114–129. Springer (2008)

[10] Arabzadeh, N., Khodabakhsh, M., Bagheri, E.: Bert-qpp: contextualized pre-trained transformers for query performance prediction. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 2857–2861 (2021)

[11] Arabzadeh, N., Meng, C., Aliannejadi, M., Bagheri, E.: Query performance prediction: From fundamentals to advanced techniques. In: European Conference on Information Retrieval. pp. 381–388. Springer (2024)

[12] Arabzadeh, N., Mitra, B., Bagheri, E.: Ms marco chameleons: challenging the ms marco leaderboard with extremely obstinate queries. In: proceedings of the 30th ACM international conference on information & knowledge management. pp. 4426–4435 (2021)

[13] Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Al-Obeidat, F., Bagheri, E.: Neural embedding-based specificity metrics for pre-retrieval query performance prediction. Information Processing & Management **57**(4), 102248 (2020)

[14] Arampatzis, A., Kamps, J., Robertson, S.: Where to stop reading a ranked list? threshold optimization using truncated score distributions. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 524–531 (2009)

[15] Baeza-Yates, R.: Modern information retrieval. Addison Wesley google schola **2**, 127–136 (1999)

[16] Bahri, D., Tay, Y., Zheng, C., Metzler, D., Tomkins, A.: Choppy: Cut transformer for ranked list truncation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1513–1516 (2020)

[17] Balog, K.: Entity-oriented search. Springer Nature (2018)

[18] Belkin, N.J., Oddy, R.N., Brooks, H.M.: Ask for information retrieval: Part i. background and theory. Journal of documentation (1982)

[19] Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on freebase from question-answer pairs. In: EMNLP (2013)

[20] Bolotova, V., Blinov, V., Scholer, F., Croft, W.B., Sanderson, M.: A non-factoid question-answering taxonomy. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1196–1207 (2022)

[21] Bonifacio, L., Abonizio, H., Fadaee, M., Nogueira, R.: Inpars: Unsupervised dataset generation for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2387–2392 (2022)

[22] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)

[23] BUCKLEY, C.: Evaluating evaluation measure stability. In: ACM SIGIR 2000 Proceedings (2000)

[24] Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.Y., Maiorano, S., Miller, G., et al.: Issues, tasks and program structures to roadmap research in question & answering (q&a). In: Document Understanding Conferences Roadmapping Documents. pp. 1–35 (2001)

[25] Cambazoglu, B.B., Tavakoli, L., Scholer, F., Sanderson, M., Croft, B.: An intent taxonomy for questions asked in web search. Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (2021)

[26] Cao, N.D., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=5k8F6UU39V

[27] Chatterjee, S., Mackie, I., Dalton, J.: Dreq: Document re-ranking using entity-based query understanding. In: European Conference on Information Retrieval. pp. 210–229. Springer (2024)

[28] Chen, L., Zhang, D., Mark, L.: Understanding user intent in community question answering. In: Proceedings of the 21st international conference on world wide web. pp. 823–828 (2012)

[29] Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators (2020)

[30] Cohen, D., Croft, W.B.: End to end long short term memory networks for non-factoid question answering. In: Proceedings of the 2016 ACM international conference on the theory of information retrieval. pp. 143–146 (2016)

[31] Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759 (2009)

[32] Cornolti, M., Ferragina, P., Ciaramita, M., Rüd, S., Schütze, H.: Smaph: A piggyback approach for entity-linking in web queries. ACM Trans. Inf. Syst. **37**, 13:1–13:42 (2019)

[33] Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. In: Text REtrieval Conference (TREC). TREC (2021)

[34] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the trec 2019 deep learning track. arXiv preprint arXiv:2003.07820 (2020)

[35] Crestani, F., Rijsbergen, C.J.v.: A model for adaptive information retrieval. Journal of Intelligent Information Systems **8**(1), 29–56 (1997)

[36] Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice, vol. 520. Addison-Wesley Reading (2010)

[37] Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 299–306 (2002)

[38] Culpepper, J.S., Faggioli, G., Ferro, N., Kurland, O.: Do hard topics exist? a statistical analysis. In: IIR (2021)

[39] Dai, Z., Callan, J.: Deeper text understanding for ir with contextual neural language modeling. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 985–988 (2019)

[40] Dai, Z., Callan, J.: Context-aware term weighting for first stage passage retrieval. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 1533–1536 (2020)

[41] Dai, Z., Xiong, C., Callan, J., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: Proceedings of the eleventh ACM international conference on web search and data mining. pp. 126–134 (2018)

[42] Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 365–374 (2014)

[43] Dalton, J., Xiong, C., Callan, J.: Trec cast 2019: The conversational assistance track overview **abs/2003.13624** (2020)

[44] Dalton, J., Xiong, C., Callan, J.: Trec cast 2019: The conversational assistance track overview. arXiv preprint arXiv:2003.13624 (2020)

[45] Das, R., Godbole, A., Kavarthapu, D., Gong, Z., Singhal, A., Yu, M., Guo, X., Gao, T., Zamani, H., Zaheer, M., et al.: Multi-step entity-centric information retrieval for multi-hop question answering. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 113–118 (2019)

[46] Datta, S., Ganguly, D., Greene, D., Mitra, M.: Deep-qpp: A pairwise interaction-based deep learning model for supervised query performance prediction. In: Proceedings of the fifteenth ACM international conference on web search and data mining. pp. 201–209 (2022)

[47] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, `https://doi.org/10.18653/v1/n19-1423`

[48] Dietz, L.: Ent rank: Retrieving entities for topical information needs through entity-neighbor-text relations. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 215–224 (2019)

[49] Dietz, L., Verma, M., Radlinski, F., Craswell, N.: Trec complex answer retrieval overview. In: TREC (2017)

[50] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)

[51] Emerson, S.L., Darnovsky, M., Bowman, J.: The practical SQL handbook: using structured query language. Addison-Wesley Longman Publishing Co., Inc. (1989)

[52] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., Auli, M.: Eli5: Long form question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3558–3567 (2019)

[53] Ferragina, P., Scaiella, U.: Fast and accurate annotation of short texts with wikipedia pages. IEEE software **29**(1), 70–75 (2011)

[54] Ferraretto, F., Laitz, T., Lotufo, R., Nogueira, R.: Exaranker: Explanation-augmented neural ranker. arXiv preprint arXiv:2301.10521 (2023)

[55] Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansion model for first stage ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2288–2292 (2021)

[56] Ganguly, D., Jones, G.J.: A non-parametric topical relevance model. Information Retrieval Journal **21**(5), 449–479 (2018)

[57] Ganguly, D., Yilmaz, E.: Query-specific variable depth pooling via query performance prediction. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2303–2307 (2023)

[58] Gao, L., Callan, J.: Long document re-ranking with modular re-ranker (2022)

[59] Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels. arXiv preprint arXiv:2212.10496 (2022)

[60] Gemmell, C., Dalton, J.: Generate, transform, answer: Question specific tool synthesis for tabular data. arXiv preprint arXiv:2303.10138 (2023)

[61] Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM international on conference on information and knowledge management. pp. 55–64 (2016)

[62] Gupta, D., Pujari, R., Ekbal, A., Bhattacharyya, P., Maitra, A., Jain, T., Sengupta, S.: Can taxonomy help? improving semantic question matching using question taxonomy. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 499–513 (2018)

[63] Guy, I., Pelleg, D.: The factoid queries collection. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 717–720 (2016)

[64] Harman, D.: Overview of the second text retrieval conference (trec-2). Information Processing & Management **31**(3), 271–289 (1995)

[65] Hashemi, H., Aliannejadi, M., Zamani, H., Croft, W.B.: Antique: A non-factoid question answering benchmark. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42. pp. 166–173. Springer (2020)

[66] Hasibi, F., Nikolaev, F., Xiong, C., Balog, K., Bratsberg, S.E., Kotov, A., Callan, J.: Dbpedia-entity v2: a test collection for entity search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1265–1268 (2017)

[67] Hauff, C., Azzopardi, L., Hiemstra, D.: The combination and evaluation of query performance prediction methods. In: European Conference on Information Retrieval. pp. 301–312. Springer (2009)

[68] Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 1419–1420 (2008)

[69] He, B., Ounis, I.: Query performance prediction. Information Systems **31**(7), 585–594 (2006)

[70] Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-based sentiment analysis using bert. In: Proceedings of the 22nd nordic conference on computational linguistics. pp. 187–196 (2019)

[71] Hölbl, M., Kompara, M., Kamišalić, A., Nemec Zlatolas, L.: A systematic review of the use of blockchain in healthcare. Symmetry **10**(10), 470 (2018)

[72] Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear **7**(1), 411–420 (2017)

[73] Hovy, E., Hermjakob, U., Ravichandran, D.: A question/answer typology with surface text patterns. In: Proceedings of the Human Language Technology conference (HLT). pp. 247–251. Citeseer (2002)

[74] Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 2333–2338 (2013)

[75] Hui, K., Yates, A., Berberich, K., De Melo, G.: Pacrr: A position-aware neural ir model for relevance matching. arXiv preprint arXiv:1704.03940 (2017)

[76] Hui, K., Yates, A., Berberich, K., De Melo, G.: Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In: Proceedings of the eleventh ACM international conference on web search and data mining. pp. 279–287 (2018)

[77] van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: Rel: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’20, ACM (2020)

[78] Huston, S., Croft, W.B.: Evaluating verbose query processing techniques. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 291–298 (2010)

[79] Huston, S., Croft, W.B.: Parameters learned in the comparison of retrieval models using term dependencies. Ir, University of Massachusetts (2014)

[80] Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing. pp. 604–613 (1998)

[81] Jafarzadeh, P., Amirmahani, Z., Ensan, F.: Learning to rank knowledge subgraph nodes for entity retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2519–2523 (2022)

[82] Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Information storage and retrieval **7**(5), 217–240 (1971)

[83] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)

[84] Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems **33**(2), 494–514 (2021)

[85] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)

[86] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)

[87] Jin, N., Siebert, J., Li, D., Chen, Q.: A survey on table question answering: recent advances. In: China Conference on Knowledge Graph and Semantic Computing. pp. 174–186. Springer (2022)

[88] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)

[89] Kadry, A., Dietz, L.: Open relation extraction for support passage retrieval: Merit and open issues. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1149–1152 (2017)

[90] Karimi, S., Pohl, S., Scholer, F., Cavedon, L., Zobel, J.: Boolean versus ranked querying for biomedical systematic reviews. BMC medical informatics and decision making **10**, 1–20 (2010)

[91] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)

[92] Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proc. of SIGIR. pp. 39–48 (2020)

[93] Kim, H., Komachi, M.: Tmu nmt system with japanese bart for the patent task of wat 2021. In: Proceedings of the 8th Workshop on Asian Translation (WAT2021). pp. 133–137 (2021)

[94] Kim, Y., Seo, J., Croft, W.B.: Automatic boolean query suggestion for professional search. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 825–834 (2011)

[95] Kwok, K.L., Chan, M.: Improving two-stage ad-hoc retrieval for short queries. In: SIGIR 1998 (1998)

[96] Lassance, C., Clinchant, S.: Naver labs europe (splade)@ trec deep learning 2022. arXiv preprint arXiv:2302.12574 (2023)

[97] Lawrie, D., Mayfield, J., Oard, D.W., Yang, E.: Hc4: A new suite of test collections for ad hoc clir. In: European Conference on Information Retrieval. pp. 351–366. Springer (2022)

[98] Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model (2023)

[99] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

[100] Lehnert, W.G.: A conceptual theory of question answering. In: Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1. pp. 158–164 (1977)

[101] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)

[102] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)

[103] Li, B.Z., Min, S., Iyer, S., Mehdad, Y., Yih, W.t.: Efficient one-pass end-to-end entity linking for questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6433–6441 (2020)

[104] Li, C., Yates, A., MacAvaney, S., He, B., Sun, Y.: Parade: Passage representation aggregation for document reranking. arXiv preprint arXiv:2008.09093 (2020)

[105] Li, H., Mourad, A., Zhuang, S., Koopman, B., Zuccon, G.: Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. ArXiv **abs/2108.11044** (2021)

[106] Li, H., Zhuang, S., Mourad, A., Ma, X., Lin, J., Zuccon, G.: Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I. pp. 599–612. Springer (2022)

[107] Li, X., Roth, D.: Learning question classifiers. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)

[108] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281 (2023)

[109] Lin, J., Ma, X.: A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. arXiv preprint arXiv:2106.14807 (2021)

[110] Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2356–2362 (2021)

[111] Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Springer Nature (2022)

[112] Lin, S.C., Yang, J.H., Lin, J.: In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). pp. 163–173 (2021)

[113] Lin, X.V., Socher, R., Xiong, C.: Bridging textual and tabular data for cross-domain text-to-sql semantic parsing. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4870–4888 (2020)

[114] Liu, L., Li, M., Lin, J., Riedel, S., Stenetorp, P.: Query expansion using contextual clue sampling with language models. arXiv preprint arXiv:2210.07093 (2022)

[115] Liu, X., Chen, F., Fang, H., Wang, M.: Exploiting entity relationship for query expansion in enterprise search. Information Retrieval **17**(3), 265–294 (2014)

[116] Liu, X., Fang, H.: Latent entity space: a novel retrieval approach for entity-bearing queries. Information Retrieval Journal **18**(6), 473–503 (2015)

[117] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[118] Losada, D.E., Parapar, J., Barreiro, Á.: Feeling lucky? multi-armed bandits for ordering judgements in pooling-based evaluation. In: proceedings of the 31st annual ACM symposium on applied computing. pp. 1027–1034 (2016)

[119] Luccioni, A.S., Viguier, S., Ligozat, A.L.: Estimating the carbon footprint of bloom, a 176b parameter language model. arXiv preprint arXiv:2211.02001 (2022)

[120] Lv, Y., Zhai, C.: Adaptive relevance feedback in information retrieval. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 255–264 (2009)

[121] MacAvaney, S., Macdonald, C., Murray-Smith, R., Ounis, I.: Intent5: Search result diversification using causal language models. arXiv preprint arXiv:2108.04026 (2021)

[122] MacAvaney, S., Macdonald, C., Ounis, I.: Streamlining evaluation with ir-measures. In: European Conference on Information Retrieval. pp. 305–310. Springer (2022)

[123] MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, O.: Expansion via prediction of importance with contextualization. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 1573–1576 (2020)

[124] MacAvaney, S., Tonellotto, N., Macdonald, C.: Adaptive re-ranking with a corpus graph. In: CIKM (2022)

[125] MacAvaney, S., Tonellotto, N., Macdonald, C.: Adaptive re-ranking with a corpus graph. In: 31st ACM International Conference on Information and Knowledge Management (2022). https://doi.org/10.1145/3511808.3557231, `https://arxiv.org/abs/2208.08942`

[126] MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: Cedr: Contextualized embeddings for document ranking. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 1101–1104 (2019)

[127] Mackie, I., Chatterjee, S., Dalton, J.: Generative relevance feedback with large language models. 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (2023)

[128] Mackie, I., Chatterjee, S., MacAvaney, S., Dalton, J.: Re-rank-expand-repeat: Adaptive query expansion for document retrieval using words and entities. ECIR Workshop on Knowledge-Enhanced Information Retrieval (KEIR) (2023)

[129] Mackie, I., Dalton, J.: Query-specific knowledge graphs for complex finance topics. AKBC Workshop, Knowledge Graphs in Finance and Economics (2022)

[130] Mackie, I., Dalton, J., Yates, A.: How deep is your learning: The dl-hard annotated deep learning dataset. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2335–2341 (2021)

[131] Mackie, I., Owoicho, P., Gemmell, C., Fischer, S., MacAvaney, S., Dalton, J.: Codec: Complex document and entity collection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2022)

[132] Malaviya, C., Shaw, P., Chang, M.W., Lee, K., Toutanova, K.: Quest: A retrieval dataset of entity-seeking queries with implicit set operations. arXiv preprint arXiv:2305.11694 (2023)

[133] Mallia, A., Khattab, O., Suel, T., Tonellotto, N.: Learning passage impacts for inverted indexes. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1723–1727 (2021)

[134] Manning, C.D.: An introduction to information retrieval. Cambridge university press (2009)

[135] Meij, E., Trieschnigg, D., De Rijke, M., Kraaij, W.: Conceptual language models for domain-specific retrieval. Information Processing & Management **46**(4), 448–469 (2010)

[136] Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 472–479 (2005)

[137] Metzler, D., Croft, W.B.: Latent concept expansion using markov random fields. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 311–318 (2007)

[138] Mihalcea, R., Csomai, A.: Wikify! linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242 (2007)

[139] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[140] Mitra, B., Craswell, N.: An updated duet model for passage re-ranking. arXiv preprint arXiv:1903.07666 (2019)

[141] Mitra, B., Diaz, F., Craswell, N.: Learning to match using local and distributed representations of text for web search. In: Proceedings of the 26th international conference on world wide web. pp. 1291–1299 (2017)

[142] Mizuno, J., Akiba, T., Fujii, A., Itou, K.: Non-factoid question answering experiments at ntcir-6: Towards answer type detection for realworld questions. In: NTCIR (2007)

[143] Moosavi, N.S., Fan, A., Shwartz, V., Glavaš, G., Joty, S., Wang, A., Wolf, T.: Proceedings of sustainlp: Workshop on simple and efficient natural language processing. In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing (2020)

[144] Mulwa, C., Lawless, S., Sharp, M., Wade, V.: The evaluation of adaptive and personalised information retrieval systems: a review. International Journal of Knowledge and Web Intelligence **2**(2-3), 138–156 (2011)

[145] Naseri, S., Dalton, J., Yates, A., Allan, J.: Ceqe: Contextualized embeddings for query expansion. In: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43. pp. 467–482. Springer (2021)

[146] Nguyen, T., MacAvaney, S., Yates, A.: Adapting learned sparse retrieval for long documents. 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (2023)

[147] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268v1 (2016)

[148] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human-generated machine reading comprehension dataset (2016)

[149] Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)

[150] Nogueira, R., Jiang, Z., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. arXiv preprint arXiv:2003.06713 (2020)

[151] Nogueira, R., Lin, J., Epistemic, A.: From doc2query to docttttquery. Online preprint **6** (2019)

[152] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons & challenges. Queue **17**, 48–75 (2019)

[153] Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.: Evaluation of information retrieval for e-discovery. Artificial Intelligence and Law **18**, 347–386 (2010)

[154] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)

[155] Overwijk, A., Xiong, C., Liu, X.X., VandenBerg, C., Callan, J.: Clueweb22: 10 billion web documents with visual and semantic information (2022), `https://api.semanticscholar.org/CorpusID:254221261`

[156] Pan, J.X., Fang, K.T., Pan, J.X., Fang, K.T.: Maximum likelihood estimation. Growth curve models and statistical diagnostics pp. 77–158 (2002)

[157] Paranjpe, P.P.: Patent information and search. DESIDOC Journal of Library & Information Technology **32**(3) (2012)

[158] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[159] Pereira, J., Fidalgo, R., Lotufo, R., Nogueira, R.: Visconde: Multi-document qa with gpt-3 and neural reranking. In: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II. pp. 534–543. Springer (2023)

[160] Perevalov, A., Diefenbach, D., Usbeck, R., Both, A.: Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). pp. 229–234. IEEE (2022)

[161] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., et al.: Kilt: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2523–2544 (2021)

[162] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2463–2473 (2019)

[163] Piccinno, F., Ferragina, P.: From tagme to wat: a new entity annotator. In: Proceedings of the first international workshop on Entity recognition & disambiguation. pp. 55–62 (2014)

[164] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: ACM SIGIR Forum. vol. 51, pp. 202–208. ACM New York, NY, USA (2017)

[165] Project, A.C.R., Cleverdon, C., Mills, J., Keen, M.: Factors Determining the Performance of Indexing Systems: Design. College of Aeronautics (1966)

[166] Radev, D.R., Qi, H., Wu, H., Fan, W.: Evaluating web-based question answering systems. In: LREC. Citeseer (2002)

[167] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[168] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (20)

[169] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)

[170] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016)

[171] Raviv, H., Kurland, O., Carmel, D.: Document retrieval using entity-based language models. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 65–74 (2016)

[172] Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94. pp. 232–241. Springer (1994)

[173] Rocchio, J.: Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing pp. 313–323 (1971)

[174] Russell-Rose, T., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: A comparison of professional search practices. Information Processing & Management **54**(6), 1042–1057 (2018)

[175] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)

[176] Samarinas, C., Dharawat, A., Zamani, H.: Revisiting open domain query facet extraction and generation. In: Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 43–50 (2022)

[177] Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2339–2352 (2021)

[178] Scholak, T., Schucher, N., Bahdanau, D.: Picard: Parsing incrementally for constrained auto-regressive decoding from language models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9895–9901 (2021)

[179] Sciavolino, C., Zhong, Z., Lee, J., Chen, D.: Simple entity-centric questions challenge dense retrievers. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6138–6148 (2021)

[180] Shehata, D., Arabzadeh, N., Clarke, C.L.: Early stage sparse retrieval with entity linking. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4464–4469 (2022)

[181] Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2014)

[182] Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd international conference on world wide web. pp. 373–374 (2014)

[183] Singh, M., Jakhar, A.K., Pandey, S.: Sentiment analysis on the impact of coronavirus in social life using the bert model. Social Network Analysis and Mining **11**(1), 33 (2021)

[184] Soare, E., Mackie, I., Dalton, J.: Docut5: Seq2seq sql generation with table documentation. arXiv preprint arXiv:2211.06193 (2022)

[185] Soleimani, A., Monz, C., Worring, M.: Nlquad: A non-factoid long question answering data set. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1245–1255 (2021)

[186] Sousa, M.G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P.H., Fernandes, E.R., Matsubara, E.T.: Bert for stock market sentiment analysis. In: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI). pp. 1597–1601. IEEE (2019)

[187] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)

[188] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13693–13696 (2020)

[189] Suryakusuma, M.R., Shiddiq, M.F.A., Lucky, H., Iswanto, I.A.: Investigating t5 generation neural machine translation performance on english to german. In: 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS). pp. 12–15. IEEE (2023)

[190] Suzuki, J., Taira, H., Sasaki, Y., Maeda, E.: Question classification using hdag kernel. In: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering. pp. 61–68 (2003)

[191] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663 (2021)

[192] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[193] Tran, H.D., Yates, A.: Dense retrieval with entity views. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 1955–1964 (2022)

[194] Van Rijsbergen, C.: Information retrieval: theory and practice. In: Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems. vol. 79, pp. 1–14 (1979)

[195] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

[196] Verberne, S., Boves, L., Oostdijk, N., Coppen, P.A.: Data for question answering: The case of why. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (2006)

[197] Voorhees, E.M.: The trec question answering track. Natural Language Engineering **7**(4), 361–378 (2001)

[198] Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004). pp. 52–69. Gaithersburg, Maryland (2004)

[199] Voorhees, E.M.: The trec robust retrieval track. In: ACM SIGIR Forum. vol. 39, pp. 11–20. ACM New York, NY, USA (2005)

[200] Wang, B., Shin, R., Liu, X., Polozov, O., Richardson, M.: Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7567–7578 (2020)

[201] Wang, X., Macdonald, C., Tonellotto, N., Ounis, I.: Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. ACM Transactions on the Web (2022)

[202] Wang, Y., Le, H., Gotmare, A., Bui, N., Li, J., Hoi, S.: Codet5+: Open code large language models for code understanding and generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 1069–1088 (2023)

[203] Wang, Y., Wang, W., Joty, S., Hoi, S.C.: Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8696–8708 (2021)

[204] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems

[205] Willett, P.: The porter stemming algorithm: then and now. Program **40**(3), 219–223 (2006)

[206] Wu, L.Y., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Zero-shot entity linking with dense entity retrieval. In: EMNLP (2020)

[207] Xiong, C., Callan, J.: Esdrank: Connecting query and documents through external semi-structured data. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management. pp. 951–960. CIKM '15, ACM,

New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806456, `http://doi.acm.org/10.1145/2806416.2806456`

[208] Xiong, C., Callan, J.: Query expansion with freebase. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. p. 111–120. ICTIR '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2808194.2809446, `https://doi.org/10.1145/2808194.2809446`

[209] Xiong, C., Callan, J., Liu, T.Y.: Bag-of-entities representation for ranking. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. p. 181–184. ICTIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2970398.2970423, `https://doi.org/10.1145/2970398.2970423`

[210] Xiong, C., Callan, J., Liu, T.Y.: Word-entity duet representations for document ranking. In: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval. pp. 763–772 (2017)

[211] Xiong, C., Dai, Z., Callan, J., Liu, Z., Power, R.: End-to-end neural ad-hoc ranking with kernel pooling. In: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval. pp. 55–64 (2017)

[212] Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International Conference on Learning Representations

[213] Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International Conference on Learning Representations (2020)

[214] Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 59–66. SIGIR '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1571941.1571954, `https://doi.org/10.1145/1571941.1571954`

[215] Yang, E., Grossman, D., Frieder, O., Yurchak, R.: Effectiveness results for popular e-discovery algorithms. In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. pp. 261–264 (2017)

[216] Yang, Z.: Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)

[217] Yates, A., Nogueira, R., Lin, J.: Pretrained transformers for text ranking: Bert and beyond. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. pp. 1154–1156 (2021)

[218] Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying bert to document retrieval with birch. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 19–24 (2019)

[219] Yu, H., Xiong, C., Callan, J.: Improving query representations for dense retrieval with pseudo relevance feedback. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3592–3596 (2021)

[220] Zeng, H., Killingback, J., Zamani, H.: Scaling sparse and dense retrieval in decoder-only llms. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2679–2684 (2025)

[221] Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the tenth international conference on Information and knowledge management. pp. 403–410 (2001)

[222] Zhao, C., Xiong, C., Qian, X., Boyd-Graber, J.: Complex factoid question answering with a free-text knowledge graph. In: Proceedings of The Web Conference 2020. pp. 1205–1216 (2020)

[223] Zheng, Z., Hui, K., He, B., Han, X., Sun, L., Yates, A.: Bert-qe: Contextualized query expansion for document re-ranking. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4718–4728 (2020)

[224] Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 543–550 (2007)

[225] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., Wen, J.R.: Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107 (2023)

[226] Zhuang, H., Qin, Z., Jagerman, R., Hui, K., Ma, J., Lu, J., Ni, J., Wang, X., Bendersky, M.: Rankt5: Fine-tuning t5 for text ranking with ranking losses. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2308–2313 (2023)