Regulating Hallucination Risks in Large Language Models Across Their Lifecycle: Lessons for China from the European Union

Linyi Xu

Submitted in fulfilment of the requirements of the Degree of Master of Laws (LLM) by Research

School of Law

College of Social Sciences

University of Glasgow

August 2025

Abstract

The rapid development of large language models (LLMs) has raised scholarly and regulatory concerns about the phenomenon of hallucinations, which means LLMs produce factually inaccurate, misleading, or fabricated content in a confident way. This thesis argues that hallucinations constitute a distinct legal and regulatory challenge that cannot be adequately addressed through general artificial intelligence (AI) governance alone. undertakes a comparative legal analysis of how the European Union (EU) and China understand, regulate, and seek to mitigate the harms caused by hallucinations in LLMs, combining doctrinal method with tech and law method to assess both the theoretical coherence and the practical enforceability of hallucination-related regulations. To better address the uncertainty that technological change creates for legal regulation, this thesis develops a lifecycle-based analytical framework that divides the lifecycle of LLMs into five interrelated stages: model design, pretraining, fine-tuning and alignment, deployment and interaction, and monitoring and iteration. On this basis, it examines how hallucination risks arise across the lifecycle, how they generate harm for stakeholders at both the micro and macro levels, and how regulation responds to those harms in practice. It further compares the legal approaches adopted in the EU and China, identifying both convergences and divergences, as well as the normative gaps and enforcement dilemmas that remain in each jurisdiction. Through this comparison, the thesis draws lessons that may inform the future regulation of LLMs in China.

This thesis advances two main arguments encompassing both theoretical and practical dimensions. Theoretically, although the EU and China recognise hallucinations as part of broader Artificial Intelligence (AI) risks and incorporate them into their AI or generative AI governance frameworks, the specific harms they pose raise distinct technical and legal challenges that are insufficiently addressed by existing general regulations. Neither jurisdiction has yet developed

a targeted legal response. However, key differences remain. The EU employs a horizontally integrated, legally codified framework anchored in the AI Act, with provisions spanning the entire lifecycle. In contrast, China adopts a vertically layered model and focuses more on the design, monitoring, and iteration stages. Practically, both jurisdictions implement measures to mitigate hallucination-related harms at both macro and micro levels. The EU's approach tends to be broader in scope, while China's is more operationally focused and implementable.

Overall, the thesis contributes to the broader discourse on AI regulation by proposing that hallucination risks demand not only lifecycle-based legal frameworks but also cross-jurisdictional learning to ensure both innovation and accountability in the governance of LLMs.

Key Words

Artificial Intelligence, Hallucination, Large Language Model, Artificial Intelligence Act

Table of Contents

Abstract II
Key Words III
Author's declaration X
Chapter 1 Introduction 1
1.1 Research Background 1
1.2 Research Questions 10
1.2.1 What are the causes and potential harms of hallucination risks in LLMs? 10
1.2.2 How can legislation be used to regulate LLMs hallucinations in the EU and China? 11
1.2.3 How can China draw the lessons on the AIA to regulate LLMs hallucinations? 11
1.2.4 How should such lessons be generalised into EU and other jurisdictions' AI legislation? 12
1.3 Research Methodology 12
1.3.1 Comparative Law Method 12
1.3.2 Legal Doctrinal Method 15
1.3.3 Technical and Law Method 17
1.4 Structure 18
Chapter 2 - Technological Foundation for Legal Research: Hallucination Harms in LLMs' Lifecycle 20
2.1 Introduction 20
2.2 The Definition of Hallucination 21
2.3 Hallucinations Generated in LLMs Lifecycle 24
2.3.1 LLMs Lifecycle 24
2.3.2 Hallucinations in LLMs Lifecycle 26
2.4 The Harms Caused by Hallucinations Risks 33
2.4.1 Harms on Micro Level 38
2.4.2 Harms on Macro Level 41
2.5 Conclusion 45
Chapter 3 - Legal Perspective Based on Technology: LLMs Regulation in the EU and China 47
3.1 Introduction 47
3.2 LLMs Regulation in the EU 48
3.2.1 Theoretical Framework: Regulatory Approach and Legal Provisions 48
3.2.2 Practical Assessment: Effectiveness of the EU's Regulation 56
3.3 AI Regulation in China 65

3.3.1 Theoretical Framework: Regulatory Approach and Legal Provisions 65
3.3.2 Practical Assessment: Effectiveness of China's Regulation..... 73
3.4 Conclusion..... 79
Chapter 4 - Comparison and Analysis of Reasons: The Similarities and Differences between the Regulation of Hallucination in EU and China 82
4.1 Introduction..... 82
4.2 Similarities Between the EU and China..... 84
4.2.1 Issues Already Addressed in Both Jurisdictions..... 84
4.2.2 Issues Yet to Be Resolved in Both Jurisdictions 88
4.3 Key Differences Between the EU and China 97
4.3.1 Differences at the Theoretical Level 97
4.3.2 Differences at the Practical Level..... 107
4.4 Conclusion..... 115
Chapter 5 - Lessons: Experiences and Insights from the EU and China 118
5.1 Introduction..... 118
5.2 Lessons for China in Regulating LLMs Hallucinations 119
5.2.1 Enhancing Global Regulatory Influence..... 119
5.2.2 Balancing Innovation and Risk..... 120
5.2.3 Covering the Full Lifecycle of LLMs 123
5.2.4 Expanding the Toolbox of Regulatory Mechanisms..... 123
5.3 Broader Lessons for Hallucinations Regulation in the EU and Other Jurisdictions 124
5.3.1 Enhancing Flexibility in Governance 125
5.3.2 Supporting Implementation through Secondary Instruments..... 126
5.3.3 Strengthening Technical and Innovation Infrastructure..... 127
5.4 Conclusion..... 129
Chapter 6 - Conclusion 131
6.1 Summary of Key Findings 131
6.2 Policy and Legislative Recommendations 133
6.3 Future Research Directions..... 135
Bibliography..... 139

List of Legislation and Policy Documents

1. EU Legislation

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L119/1

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act) [2022] OJ L265/1

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) [2022] OJ L277/1

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data (Data Act) [2023] OJ L 2023/2854

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689

2. China National Laws and Regulations

Cybersecurity Law of the People's Republic of China (中华人民共和国网络安全法) (adopted 7 November 2016, effective 1 June 2017; amended 28 October 2025, effective 1 January 2026)

Data Security Law of the People's Republic of China (中华人民共和国数据安全法) (adopted 10 June 2021, effective 1 September 2021)

Personal Information Protection Law of the People's Republic of China (中华人民共和国个人信息保护法) (adopted 20 August 2021, effective 1 November 2021)

Provisions on the Administration of Algorithmic Recommendations for Internet Information Services (互联网信息服务算法推荐管理规定) (Cyberspace Administration of China and others, 31 December 2021, effective 1 March 2022)

Provisions on the Administration of Deep Synthesis in Internet Information Services (互联网信息服务深度合成管理规定) (Cyberspace Administration of China and others, 25 November 2022, effective 10 January 2023)

3. China Local Legislation and Policy Documents

Ethical Norms for the New Generation Artificial Intelligence (新一代人工智能伦理规范) (Ministry of Science and Technology of the People's Republic of China, 26 September 2021)

Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) (Cyberspace Administration of China and others, 10 July 2023, effective 15 August 2023)

New Generation Artificial Intelligence Development Plan (新一代人工智能发展规划) (State Council of the People's Republic of China, 20 July 2017)

New Generation Artificial Intelligence Governance Principles: Developing Responsible Artificial Intelligence (新一代人工智能治理原则——发展负责任的人工智能) (National New Generation Artificial Intelligence Governance Professional Committee, 17 June 2019)

Regulations of Shanghai Municipality on Promoting the Development of the Artificial Intelligence Sector (上海市促进人工智能产业发展条例) (Shanghai Municipal People's Congress Standing Committee, 22 September 2022, effective 1 October 2022)

Shenzhen Special Economic Zone Regulations on Promoting the Artificial Intelligence Industry (深圳经济特区人工智能产业促进条例) (Shenzhen Municipal People's Congress Standing Committee, 30 August 2022, effective 1 November 2022)

4. 技术标准与规范 (Technical Standards)

State Administration for Market Regulation (SAMR), 'Information technology—Artificial intelligence—Platform computing resource specification' (信息技术 人工智能 平台计算资源规范, GB/T 42018-2022, 14 October 2022)

SAMR, 'Information technology—Artificial intelligence—Terminology' (信息技术 人工智能 术语, GB/T 41867-2022, 14 October 2022)

5. International and Soft Law Documents

European Commission, Ethics Guidelines for Trustworthy AI (*High-Level Expert Group on AI*, 8 April 2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai accessed 17 June 2025

OECD, Recommendation of the Council on Artificial Intelligence (*OECD*, 22 May 2019) https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 accessed 17 July 2025

UNESCO, Recommendation on the Ethics of Artificial Intelligence (*UNESCO*, 23 November 2021) https://unesdoc.unesco.org/ark:/48223/pf0000381137 accessed 17 July 2025

Author's declaration

"I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution."

Printed Name: ____Linyi Xu_____

Signature: _____

Chapter 1 Introduction

1.1 Research Background

Generative AI (Gen AI) such as ChatGPT, DeepSeek, Google Gemini, Microsoft Copliot, and Meta's Llama, has brought significant convenience to society and ushered in a transformative era in the field of artificial intelligence (AI). Gen AI can run on different models that use different mechanisms to do training and output new content. Among these, Large Language Models (LLMs) represent a specialised subset focused on producing human-like text,[1] which are trained on massive text datasets to understand and produce coherent, contextually relevant language.[2] While LLMs offer unprecedented capabilities in natural language processing, generation, and understanding, they also face critical challenges, particularly the growing concern over hallucinations.[3] Since LLMs are large statistical models that predict the next word, phrase, sentence, or paragraph based on input data, they do not truly understand the text. Instead, they rely on

---

[1] In this thesis, 'LLMs' refer to text-based language models unless otherwise specified; Abdenour Hadid, Tanujit Chakraborty and Daniel Busby, 'When Geoscience Meets Generative AI and Large Language Models: Foundations, Trends, and Future Challenges' (2024) 41(10) Expert Systems e13654.

[2] Malik Sallam, 'ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns' (2023) 11(6) Healthcare 887; Raju Vaishya, Anoop Misra and Abhishek Vaish, 'ChatGPT: Is This Version Good for Healthcare and Research?' (2023) 17 Diabetes & Metabolic Syndrome: Clinical Research & Reviews 102744.

[3] Catalina Goanta and others, 'Regulation and NLP (RegNLP): Taming Large Language Models' in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) 8715 <https://aclanthology.org/2023.emnlp-main.539/> accessed 24 March 2026.

algorithms to make probabilistic predictions.[4] As a result, these models may produce hallucinations, namely outputs that appear credible or factually accurate but are not supported by real data or verifiable facts.[5] These outputs are often presented confidently, with logical explanations, even when the answers are incorrect.[6]

Hallucinations generated by LLMs are not only a technical issue, but also a social problem with important legal implications. They create significant difficulties in practical applications and may give rise to serious consequences.[7] For example, in 2025, as DeepSeek advanced in AI capabilities, some Chinese writers expressed concern that its hallucinations could disrupt the reliability of online information in China.[8] Besides, in 2023, attorney Stephen Schwartz was fined $5,000 for relying on ChatGPT, which fabricated six legal precedents, underscoring the risks

---

[4] Dorottya Demszky and others, 'Using Large Language Models in Psychology' (2023) 2(11) Nature Reviews Psychology 688.

[5] Elijah Berberette, Jack Hutchins and Amir Sadovnik, 'Redefining "Hallucination" in LLMs: Towards a Psychology-Informed Framework for Mitigating Misinformation' (arXiv, 1 February 2024) <http://arxiv.org/abs/2402.01769> accessed 28 October 2024.

[6] Joshua Maynez and others, 'On Faithfulness and Factuality in Abstractive Summarization' in Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2020) 1906 <https://aclanthology.org/2020.acl-main.173/> accessed 26 November 2024.

[7] Susmit Jha and others, 'Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting' *2023 IEEE International Conference on Assured Autonomy (ICAA)* (IEEE 2023) 149 <https://ieeexplore.ieee.org/abstract/document/10207581/> accessed 12 December 2024.

[8] Xi Lan, 'DeepSeek's fabrication is flooding the Chinese internet.' (DeepSeek 的胡编乱造，正在淹没中文互联网) (*Sina Technology*, 6 March 2025) <https://finance.sina.com.cn/tech/roll/2025-03-06/doc-inensrzp6931316.shtml> accessed 10 March 2025.

of AI generating fake court citations.[9] Similarly, in 2022, Meta AI introduced Galactica, a scientific LLM, which was quickly withdrawn after users discovered it produced authoritative-sounding content with fictitious citations and fabricated papers. [10] Furthermore, AI hallucinations may threaten democracy by misrepresenting election procedures. A February 2024 study found that five major LLMs provided incorrect information, including erroneous voter ID requirements, which could lead to ballot refusals.[11] In the scientific use of ChatGPT, such as drafting scientific texts, hallucinations can pose significant limitations by generating plausible-sounding citations with complete bibliographic details that do not actually exist. [12] These examples highlight the harms caused by hallucinations, emphasising the urgent need to address these issues in LLMs.

Besides, in examining hallucination as a distinct type of "risk" and the resulting "harm" it may produce, the concepts of risk and harm should be interpreted with reference to their established legal definitions. The concept of "risk" in this thesis

[9] Ramishah Maruf, CNN, 'Lawyer apologizes for fake court citations from ChatGPT' (*CNN Business*, 28 May 2023) <https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers> accessed 20 November 2024.

[10] Benj Edwards, 'New Meta AI demo writes racist and inaccurate scientific literature, gets pulled' (*arstechnica*, 18 November 2022) <https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/> accessed 20 November 2024.

[11] Julia Angwin, Alondra Nelson, Rina Palta, 'Seeking Reliable Election Information? Don't Trust AI' (*Proof News*, 27 February 2024) <https://www.proofnews.org/seeking-election-information-dont-trust-ai/ >accessed 15 January 2025.

[12] Teresa Kubacka, 'Today I asked ChatGPT about the topic I wrote my PhD about' (*Lookalikes and Meanders*, 6 December 2022) <https://lookalikes.substack.com/p/today-i-asked-chatgpt-about-the-topic> accessed 13 March 2025.

refers to the definition established in the EU AI Act (AIA)[13], which is based on a risk-based regulatory system. The narrow conception of harm in Article 5(1)(a) of the AIA, limited to physical or psychological injury, is unduly restrictive and fails to account for the more diffuse ways in which AI systems can cause damage. As Veale and Zuiderveen Borgesius argue[14], such a definition overlooks cumulative, cognitive, and societal harms that are no less significant. To avoid these shortcomings, this thesis adopts a broader conception of harm in Chapter 2, which provides a more comprehensive and reliable basis for analysing the harms of LLM hallucinations.

Against this background, the following section outlines the existing body of research. Firstly, existing research primarily focuses on AI rules[15] or other risks associated with LLMs, such as bias,[16] output memorised passages from their

[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L, 2024/1689.

[14] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22(4) Computer Law Review International 97.

[15] Jonas Schuett, 'Risk Management in the Artificial Intelligence Act' (2024) 15(2) European Journal of Risk Regulation 367; Philipp Hacker, 'AI Regulation in Europe: From the AI Act to Future Regulatory Challenges' (arXiv, 6 October 2023) <https://doi.org/10.48550/arXiv.2310.04072> accessed 20 September 2025; Teresa Rodríguez de Las Heras Ballell, 'Mapping Generative AI Rules and Liability Scenarios in the AI Act, and in the Proposed EU Liability Rules for AI Liability' (2025) 1 Cambridge Forum on AI: Law and Governance e5.

[16] Emilio Ferrara, 'Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models' (2023) 28(11) First Monday; Jwala Dhamala and others, 'BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation' *in Proceedings of the 2021 ACM Conference on Fairness*, Accountability, and Transparency (ACM 2021) 862 <https://doi.org/10.1145/3442188.3445924> accessed 11 March 2025.

training data,[17] and unpredictable phenomena,[18] or general risks with LLMs.[19] However, legal scholarship specifically addressing hallucinations remains limited. Secondly, there is less interdisciplinary research that combines legal perspectives Technological studies seldom assess regulatory compliance,[20] while legal studies often lack a concrete technical foundation. For legal research, a classification grounded in legal theory is more suitable than one based solely on technical features. Existing studies variously distinguish between intrinsic, extrinsic, and model-specific hallucinations;[21] fact inconsistency, query inconsistency, and

[17] Nicholas Carlini and others, 'Extracting Training Data from Large Language Models' in 30th USENIX Security Symposium (USENIX Security 21) (USENIX Association 2021) 2633 <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting> accessed 11 March 2025.

[18] Jason Wei and others, 'Emergent Abilities of Large Language Models' (arXiv, 15 June 2022) <https://doi.org/10.48550/arXiv.2206.07682> accessed 8 March 2025.

[19] Ziwei Ji and others, 'Survey of Hallucination in Natural Language Generation' (2023) 55(12) ACM Computing Surveys art 248.

[20] Eloise Ainsworth, Justin Wycliffe and Florence Winslow, 'Reducing Contextual Hallucinations in Large Language Models through Attention Map Optimization' (*TechRxiv*, 23 July 2024) <https://www.techrxiv.org/users/807662/articles/1206169-reducing-contextual-hallucinations-in-large-language-models-through-attention-map-optimization> accessed 14 December 2024.

[21] Timothy R McIntosh and others, 'A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination' (2024) 5(6) IEEE Transactions on Artificial Intelligence 2739; Nouha Dziri and others, 'Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding' in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2021) 2197 <https://aclanthology.org/2021.emnlp-main.168/> accessed 22 March 2025; Ben Goodrich and others, 'Assessing the Factual Accuracy of Generated Text' in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM 2019) <https://doi.org/10.1145/3292500.3330955> accessed 22 March 2025; Chunting Zhou and others, 'Detecting Hallucinated Content in Conditional Neural Sequence Generation' in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Association for Computational Linguistics 2021) 1393 <https://aclanthology.org/2021.findings-acl.120/> accessed 22 March 2025.

tangentiality;[22] or factual and faithfulness hallucinations.[23] Existing theoretical research often leans towards technological aspects, especially analysis hallucination mitigation techniques such as improving explainability [24] and transparency, identify hallucination [25], hallucination detection, [26] categorise hallucination mitigation techniques for LLMs,[27] introduce Retrieval-Augmented

[22] Ziwei Ji and others, 'Towards Mitigating Hallucination in Large Language Models via Self-Reflection' (arXiv, 10 October 2023) <https://doi.org/10.48550/arXiv.2310.06271> accessed 14 December 2024.

[23] Lei Huang and others, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions' (2025) 43(2) ACM Transactions on Information Systems.

[24] Haiyan Zhao and others, 'Explainability for Large Language Models: A Survey' (2024) 15(2) ACM Transactions on Intelligent Systems and Technology.

[25] Yifu Qiu and others, 'Think While You Write: Hypothesis Verification Promotes Faithful Knowledge-to-Text Generation' in *Findings of the Association for Computational Linguistics: NAACL 2024* (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.findings-naacl.106/> accessed 19 March 2025; Yifu Qiu and others, 'Detecting and Mitigating Hallucinations in Multilingual Summarisation' in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) 8940 <https://aclanthology.org/2023.emnlp-main.551/> accessed 19 March 2025; Shuo Zhang and others, 'The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models' in *Findings of the Association for Computational Linguistics: ACL 2024* (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.findings-acl.121/> accessed 19 March 2025.

[26] Potsawee Manakul, Adian Liusie and Mark J F Gales, 'SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models' in Houda Bouamor, Juan Pino and Kalika Bali (eds), in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) <https://aclanthology.org/2023.emnlp-main.557.pdf> accessed 23 March 2025.

[27] SM Towhidul Islam Tonmoy and others, 'A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models' (arXiv, 2 January 2024) <http://arxiv.org/abs/2401.01313> accessed 17 October 2024; Timothy R McIntosh and others, 'A Culturally Sensitive Test to

Generation (RAG) and with limited analysis of regulatory frameworks. Thirdly, most research focuses on reducing hallucinations without recognising their potential benefits or the need for a balanced regulatory approach, overlooking the fact that hallucinations cannot be eliminated.[28] Some research predominantly focuses on risk avoidance,[29] often overlooking the fact that hallucinations can be mitigated but not eliminated.[30] Finally, comparative legal analysis of Gen AI regulatory frameworks remains underdeveloped, especially in relation to the regulation of hallucinations. Existing comparative studies, even those comparing China and the EU, primarily focus on broad legislation,[31] such as whether China

Evaluate Nuanced GPT Hallucination' (2024) 5(6) *IEEE Transactions on Artificial Intelligence* 2739.

[28] Roman Capellini, Frank Atienza and Melanie Sconfield, 'Knowledge Accuracy and Reducing Hallucinations in LLMs via Dynamic Domain Knowledge Injection' <https://www.researchsquare.com/article/rs-4540506/latest> accessed 28 October 2024; Jichang Chen, Xinnan Huang and Yongping Li, 'Dynamic Supplementation of Federated Search Results for Reducing Hallucinations in LLMs' (OSF Preprints, 2024) <https://doi.org/10.31219/osf.io/x5vge> accessed 28 October 2024; Minhyeok Lee, 'A Mathematical Investigation of Hallucination and Creativity in GPT Models' (2023) 11(10) Mathematics 2320.

[29] Claudio Novelli and others, 'Taking AI Risks Seriously: A New Assessment Model for the AI Act' (2024) 39(5) AI & Society 2493; Yoshua Bengio and others, 'Managing Extreme AI Risks amid Rapid Progress' (2024) 384(6698) Science 842.

[30] Ziwei Xu, Sanjay Jain and Mohan Kankanhalli, 'Hallucination is Inevitable: An Innate Limitation of Large Language Models' (arXiv, 22 January 2024) <https://arxiv.org/abs/2401.11817> accessed 19 March 2025.

[31] Daniel Albrecht, 'Chinese First Personal Information Protection Law in Contrast to the European GDPR' (2022) 23(1) Computer Law Review International 1; Wenlong Li and Jiahong Chen, 'From Brussels Effect to Gravity Assists: Understanding the Evolution of the GDPR-Inspired Personal Information Protection Law in China' (2024) 54 Computer Law & Security Review 105994.

should emulate the EU in establishing a dedicated AI law.[32] Alfiani et al[33] and Roberts et al[34] compare China and the EU, focusing on regulatory objectives, sectoral approaches, and policy beneficiaries, while offering recommendations for ethical AI governance. Hasan[35] analyses AI legislation in South Asia against broader global frameworks, identifying regulatory disparities and the structural constraints affecting implementation. Keith[36] examines the limited participation of Southeast Asian states in the Global Partnership on Artificial Intelligence (GPAI) and proposes more inclusive governance strategies centred on human capital development. Hine[37] et al use a philosophy-of-technology framework and natural language processing to compare US and Chinese AI policies and their historical foundations. Luna[38] et al provide a cross-regional comparison of six jurisdictions to identify commonalities and divergences in regulatory coverage.

---

[32] Huijuan Dong and Junkai Chen, 'Meta-Regulation: An Ideal Alternative to the Primary Responsibility as the Regulatory Model of Generative AI in China' (2024) 54 Computer Law & Security Review 106016; Angela Huyue Zhang, 'The Promise and Perils of China's Regulation of Artificial Intelligence' (2025) 63 Columbia Journal of Transnational Law 1.

[33] Francisca Romana Nanik Alfiani and Faisal Santiago, 'A Comparative Analysis of Artificial Intelligence Regulatory Law in Asia, Europe, and America' in *SHS Web of Conferences vol 204* (EDP Sciences 2024) 07006.

[34] Huw Roberts and others, 'Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes' (2023) 39(2) The Information Society 79.

[35] Mahmud Hasan, 'Regulating Artificial Intelligence: A Study in the Comparison between South Asia and Other Countries' (2024) 5(1) Legal Issues in the Digital Age 122.

[36] Andrew J Keith, 'Governance of Artificial Intelligence in Southeast Asia' (2024) 15(5) Global Policy 937.

[37] Emmie Hine and Luciano Floridi, 'Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies' (2024) 39(1) AI & Society 257.

[38] Jose Luna and others, 'Navigating Governance Paradigms: A Cross-Regional Comparative Study of Generative AI Governance Processes &

Based on the above context, this thesis focuses on the AI-related laws, specifically examining how it addresses hallucinations throughout the lifecycle of LLMs and evaluates its effectiveness in mitigating the associated harms. The term "AI-related laws" in this thesis refers to regulations specifically designed to address AI-related issues, such as those targeting Gen AI, deep synthesis, and deepfake technologies, etc. Although frameworks such as the Personal Information Protection Law of the People's Republic of China (PIPL)[39], General Data Protection Regulation (GDPR)[40], the Digital Markets Act (DMA)[41], the Digital Services Act (DSA)[42], and the Data Act (DA)[43] address certain AI-related issues,[44] these

---

Principles' (2024) 7(1) in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 917.

[39] Personal Information Protection Law of the People's Republic of China (中华人民共和国个人信息保护法) (adopted 20 August 2021, effective 1 November 2021).

[40] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

[41] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L 265/1.

[42] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1.

[43] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) [2023] OJ L 2023/2854.

[44] Philipp Hacker, Johann Cordes and Janina Rochon, 'Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access, and Fairness under the Digital Markets Act, the General Data Protection Regulation, and beyond' (2024) 15(1) European Journal of Risk Regulation 49.

legislations horizontally strengthen data flow and competition rules but are not AI-specific laws,[45] and thus fall outside the scope of this thesis.[46]

1.2 Research Questions

The research question is: ***what lessons can China learn from the EU's regulation of LLM hallucination risks across the model lifecycle?*** This question can be further broken down into several smaller research questions.

1.2.1 What are the causes and potential harms of hallucination risks in LLMs?

Firstly, the concept of hallucination must be clarified, as the term originates from other disciplines. What does hallucination mean in the context of AI and law, and why do LLMs generate it? Secondly, understanding the lifecycle of LLMs is crucial. How can it be reasonably divided into distinct stages, and at which points do hallucinations occur? A well-defined lifecycle framework provides a benchmark

---

[45] Lilian Edwards, 'The EU AI Act: A Summary of Its Significance and Scope' (Ada Lovelace Institute, 11 April 2022) <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf> accessed 26 March 2026.

[46] The choice of AI-specific legal frameworks, particularly those adopted by the EU and China to address hallucination risks in LLMs, which is grounded in several considerations. Firstly, it enhances the feasibility of the research. The regulation of hallucinations involves not only data-related laws and AI-specific regulations but also product liability, content regulation, and other legal domains. This creates a fragmented and complex research scope. Moreover, since different EU member states adopt varying approaches to AI regulation, studying a broad range of laws would hinder the ability to effectively summarise the EU's overall approach and conduct meaningful comparative research. Secondly, AI-related law is a relatively new and evolving field, with ongoing academic exploration. In contrast, existing research on frameworks such as the GDPR, DSA, and China's PIPL and data protection laws is already extensive, which would limit the innovative contribution of this thesis.

for assessing whether the legal system adequately addresses all phases of development and deployment. Thirdly, the taxonomy of harms associated with hallucinations in LLMs must be examined. Such classification underscores the need for regulation and offers a practical basis for evaluating the effectiveness of relevant legislation.

1.2.2 How can legislation be used to regulate LLMs hallucinations in the EU and China?

The regulation of LLMs hallucinations can be examined from both theoretical and practical perspectives, raising two central questions: What legislative frameworks exist in the EU and China for regulating LLMs hallucinations, and how effective are these frameworks in practice? The theoretical inquiry concerns the macro-level regulatory approaches and specific legal provisions adopted in each jurisdiction. The practical inquiry focuses on how these regulations are implemented and how effectively they mitigate the harms caused by hallucinations.

1.2.3 How can China draw the lessons on the AIA to regulate LLMs hallucinations?

Drawing on developments in both the EU and China, several key questions arise. First, how might China refine its regulatory approach so as to avoid some of the limitations associated with the EU AIA, particularly the risk of imposing disproportionate burdens on technological innovation? Moreover, can China build on its experience with the PIPL, which draws on the GDPR, to develop AI regulations that strike a balance between innovation and the management of hallucination risks?

1.2.4 How should such lessons be generalised into EU and other jurisdictions' AI legislation?

While China can draw lessons from the EU's regulatory approach, the EU may also benefit from China's theoretical and practical experiences in AI governance as it refines its framework through instruments such as the Code of Practice and future legal amendments. A comparative study of the two jurisdictions can yield broadly applicable insights for developing AI regulations elsewhere and reveal common limitations and challenges that remain unresolved across regulatory contexts.

1.3 Research Methodology

This thesis adopts a comparative approach, combining both theoretical and practical perspectives on AI regulation, with particular attention to hallucinations. Given the multidimensional nature of hallucinations, which encompass technical, legal, ethical, and societal dimensions, the thesis employs a dual methodological framework consisting of legal doctrinal analysis and tech and law method.

1.3.1 Comparative Law Method

The purpose of comparative law method is looking to other, foreign legal systems for illumination and insight in the hope that wisdom and understanding are to be gained, either from a foreign legal system or our own.[47] This thesis adopts the functional method of comparative law, and the key comparative question is not "What does the law say?" but rather "How does the law resolve a given issue?"

[47] Edward J Eberle, 'The Methodology of Comparative Law' (2011) 16(1) Roger Williams University Law Review 51.

The functional method is therefore necessary because it compares the EU and China in terms of how they address the same challenge through different legal mechanisms, rather than merely comparing formal rules. This is particularly important for hallucination risks, where regulatory responses are shaped not only by legal design but also by broader institutional, social, economic, and cultural conditions. On this basis, the thesis identifies both commonalities and divergences in the two jurisdictions' approaches to hallucination governance.

Furthermore, the comparison is carried out across two levels: theoretical design and practical enforcement. This dual lens enables a more nuanced evaluation of each jurisdiction's strengths and weaknesses in governing hallucination-related harms. Specifically, the analysis will assess how both the EU and China approach the reduction of hallucination risks through three dimensions: overall regulatory strategy, specific legal provisions, and real-world effectiveness. The functional purpose of regulating hallucination-related harms is grounded in the concrete problems identified in Chapter 2, which maps how hallucinations arise at different stages of the LLM lifecycle and categorises the micro- and macro-level harms they produce. This lifecycle-based harm typology provides the foundation for evaluating whether existing legal measures effectively fulfil their intended regulatory function.

The choice of comparing the EU and China stems from several factors: firstly, the EU is a global frontrunner in AI regulation, with its comprehensive AIA and ongoing development of a Code of Practice, providing a valuable model for analysis. Meanwhile, China's rapidly evolving AI legislation makes it crucial to learn from the EU's experiences. Secondly, previous policy borrowing examples, such as China's PIPL drawing from the EU's GDPR,[48] highlight the feasibility of comparing

[48] Igor Calzada, 'Citizens' Data Privacy in China: The State of the Art of the Personal Information Protection Law (PIPL)' (2022) 5(3) Smart Cities 1129; Graham Greenleaf, 'China's Completed Personal Information Protection

these jurisdictions. Lastly, both jurisdictions provide rich legislative and academic materials, making a thorough comparative study feasible. In both China and the EU, AI regulation has developed within existing legal frameworks, especially those relating to data protection. Yet, despite some similarities, their AI governance strategies differ significantly. Understanding these differences is important not only for legal comparison but also for firms preparing for compliance in both markets.

However, this comparative law method also has certain limitations:

First, comparison between the EU and China is constrained by major structural differences in their legal and regulatory systems. The EU adopts a supranational, rights-based framework, whereas China relies on a more centralised and policy-driven approach. The EU AIA also has broader extraterritorial effects, while China's regulatory regime remains primarily domestic. In addition, the EU provides a relatively unified binding framework, whereas China governs LLMs through a more fragmented mix of regulations, policy documents, and technical standards. These differences limit the comparability of the two systems and make direct evaluation of regulatory outcomes more difficult. Second, the regulation of hallucination risks remains at an early stage in both jurisdictions. Neither the EU nor China explicitly treats hallucinations as a distinct legal category. As a result, this thesis necessarily relies on broader legal provisions and interpretive analysis, which increases the level of abstraction and limits the precision of normative conclusions.

Law: Rights plus Cyber-Security' (2021) 172 Privacy Laws & Business International Report 20.

Notwithstanding these limitations, the comparative law method remains valuable. It helps identify differences in regulatory logic, reveal areas for mutual learning, and generate context-sensitive insights into the governance of LLM hallucinations.

1.3.2 Legal Doctrinal Method

Legal doctrinal method involves the systematic analysis of rules, principles, and concepts within a specific legal system. It guides the application and development of law through the interpretation of sources such as cases, statutes, and regulations,[49] and serves three core objectives: description, prescription, and justification.[50] This thesis applies the doctrinal method to describe and analyse the legal frameworks governing hallucinations in the EU and China, focusing on legislative texts relevant to the LLM lifecycle and assessing their coherence and legal soundness.[51]

First, this thesis adopts doctrinal analysis to examine legislation relevant to LLMs hallucinations in the EU and China at both the macro and micro levels. At the macro level, it considers the broader regulatory approaches to AI, focusing specifically on dedicated AI legislation for reasons of precision and feasibility. At the micro level, it compares specific legal provisions, with particular attention to

[49] Bhagyamma G, 'A Comparative Analysis of Doctrinal and Non-Doctrinal Legal Research' (2023) 1(1) ILE Journal of Governance and Policy Review 88.

[50] Jan M Smits, 'What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research' in Rob van Gestel, Hans-W Micklitz and Edward L Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 207.

[51] Sofia Terzi and Ioannis Stamelos, 'Architectural Solutions for Improving Transparency, Data Quality, and Security in eHealth Systems by Designing and Adding Blockchain Modules, While Maintaining Interoperability: The eHDSI Network Case' (2024) 14 Health and Technology 451.

the EU AIA and China's Interim Measures for the Management of Generative Artificial Intelligence Services (Interim Measures).[52]

Second, this thesis examines whether existing legal frameworks can adequately address hallucination-related risks within the broader body of AI law. If current rules can reasonably be interpreted as covering liability and other risks arising from AI hallucinations, this may suggest that existing law provides a sufficient regulatory basis and may offer lessons for other jurisdictions. If, however, the current legal framework lacks clear and specific measures to address LLM hallucination risks, this would indicate the need for further regulatory development.

However, doctrinal analysis also has important limitations in this context. First, its focus on legal texts may overlook how rules operate in practice, including enforcement gaps, implementation difficulties, and behavioural responses. Second, it often assumes a degree of coherence and completeness within the legal system, whereas AI regulation remains fragmented, evolving, and frequently open to interpretation, especially in relation to hallucinations. Third, doctrinal research tends to prioritise formal legal sources while giving less weight to soft law, policy documents, and technical standards, even though these play a significant role in AI governance,[53] particularly in China. Finally, because AI develops rapidly,

[52] Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) (Cyberspace Administration of China and others, 10 July 2023, effective 15 August 2023).

[53] Terry Hutchinson, 'The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law' (2015) 8(3) Erasmus Law Review 130; Pradeep M D, 'Legal Research-Descriptive Analysis on Doctrinal Methodology' (2019) 4(2) International Journal of Management, Technology, and Social Sciences 95.

doctrinal analysis is limited in its ability to anticipate emerging challenges or assess future regulatory needs.[54]

1.3.3 Technical and Law Method

This thesis adopts a combined technical and law method. The advent of any new technology raises a host of techlaw questions particular to that technology and the legal context.[55] The role of the legal scholar is to accept extant or anticipated technology and puzzle out how law and legal institutions might adapt to it.[56] In this thesis, a technical perspective is necessary to explain the lifecycle of LLMs, identify where hallucinations arise, and show how they propagate in practice. Without such grounding, legal analysis risks becoming overly abstract and disconnected from the operational realities of AI systems. However, technical measures alone are insufficient. Exclusive reliance on them risks techno-solutionism and cannot resolve questions of accountability, rights protection, institutional responsibility, or regulatory justification. A legal perspective is therefore needed to assess hallucination risks within a normative and institutional framework, particularly in relation to the EU AIA and China's evolving AI governance regime. Bringing these perspectives together allows this thesis to link the causes of hallucinations to the regulatory responses designed to address them.

---

[54] Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' (2012) 17(1) Deakin Law Review 83; Bhagyamma G, 'A Comparative Analysis of Doctrinal and Non-Doctrinal Legal Research' (2023) 1(1) ILE Journal of Governance and Policy Review 88.

[55] BJ Ard and Rebecca Crootof, 'Legal Responses to Techlaw Uncertainties' in Bartosz Brożek, Olia Kanevskaia and Przemysław Pałka (eds), *Research Handbook on Law and Technology* (Edward Elgar 2023) 28.

[56] Ryan Calo, *Law and Technology: A Methodical Approach* (OUP 2025) 3.

This approach nevertheless has limits: the technical analysis relies mainly on secondary sources, while the legal analysis must engage with a rapidly changing technological field. Even so, their combination provides a necessary methodological foundation for studying hallucination regulation.

1.4 Structure

This research consists of six chapters.

Chapter 1 introduces the research by outlining the background, formulating the research questions and detailing the methodology employed.

Chapter 2 examines the technological causes of hallucinations in LLMs. It begins by defining hallucination, discussing the term itself, and explaining why such outputs arise. Using a lifecycle approach, it identifies the stages at which hallucinations are most likely to occur and analyses the risks and harms they generate. This chapter provides the technological foundation for assessing the legal frameworks discussed later.

Chapter 3 turns to the regulation of LLMs hallucinations in the EU and China. It evaluates the strengths and weaknesses of existing approaches from both theoretical and practical perspectives. The chapter first examines the regulatory approaches and legal provisions adopted in each jurisdiction and then considers their implementation and effectiveness in mitigating hallucination-related harms.

Chapter 4 presents a comparative law method of the regulatory frameworks governing hallucination risks in the EU and China. It identifies and critically

assesses the similarities and differences between the two jurisdictions, and examines the legal, institutional, economic, cultural, and political factors that help explain them. With regard to similarities, the chapter considers both the issues that have been addressed relatively effectively in each jurisdiction and those that remain unresolved. This helps to reveal shared regulatory gaps and potential areas for mutual learning. As for differences, the comparison highlights the distinctive strengths of each jurisdiction and analyses them in light of their respective social contexts. The chapter also identifies continuing challenges faced by both jurisdictions, thereby providing a basis for further research and future policy development.

Chapter 5 outlines potential lessons for improving China's regulatory framework from EU based on the above findings, while also reflecting on how China's and EU's experience may inform AI governance in other jurisdictions.

Chapter 6 summarises the research findings and outlines directions for future research.

Chapter 2 - Technological Foundation for Legal Research: Hallucination Harms in LLMs' Lifecycle

2.1 Introduction

Due to the technical and law method of this thesis, it is essential to first consider key technological aspects that could inform the legal discussion. This chapter does not aim to provide an exhaustive overview of technological perspectives. Instead, it outlines the core technical principles most relevant to the legal analysis that follows, serving as a focused foundation for the legal research rather than a comprehensive technological survey.

Defining hallucinations in LLMs is a necessary first step, as it forms the basis for subsequent discussions on risks and harms. Without a clear definition, it is impossible to systematically analyse how hallucinations manifest, what risks and harms they pose, and how regulatory frameworks can address them. This chapter will firstly analysis the lifecycle and harms of LLMs in technology theory to provide technological standards for the subsequent legal analysis.

Secondly, due to the complexity process of content generated by LLMs, direct regulation across the entire lifecycle may be impractical. Such an approach would likely be overly broad and fail to precisely target the issue of hallucinations. Therefore, this thesis applies the LLMs lifecycle framework to identify the specific stages at which hallucinations arise and the mechanisms through which they occur. Understanding these stages is essential for evaluating whether legal provisions in the EU and China effectively address risks across the full LLMs lifecycle. To support this analysis, the thesis adopts a refined five-stages approach, including Model Design, Pretraining, Fine-tuning and Alignment, Deployment and Interaction, and

Monitoring and Iteration, providing a more structured and targeted foundation for legal assessment.

Thirdly, this chapter categorises the harms caused by hallucinations by continuing to build on the lifecycle framework, aiming to conduct an in-depth legal analysis of the harms caused by hallucinations throughout the lifecycle of LLMs. Since hallucinations occurring at different stages pose varying risks to different entities, this research will further analyse and summarise the types of harms at two levels—macro and micro. The micro level primarily refers to the users directly engaging with AI, practitioners, and producers of LLMs, as well as individuals and legal entities affected by hallucinations. The macro level concerns the direct or indirect harm hallucinations may cause to society and the regulatory landscape. By mapping hallucination risks to specific harms, the analysis establishes a standard for evaluating the effectiveness of regulations in mitigating these harms.

In conclusion, this chapter establishes the technological foundation for the subsequent legal analysis. It introduces a lifecycle framework through which the causes of hallucinations can be examined at each stage and, by analysing the resulting harms at both macro and micro levels, provides practical criteria for evaluating the effectiveness of legal regulation.

2.2 The Definition of Hallucination

To begin with, the appropriateness of the term "hallucination" has been questioned by some scholars, who propose alternative terminology. For instance, the term "confabulation" has been suggested as a more accurate characterisation, aligning with the medical definition where "patients produce false memories

without intent to deceive".[57] Other scholars argue that LLMs like ChatGPT are better described as producing "bullshit", as defined by Harry Frankfurt, which lacks concern for truth or falsehood rather than being deliberate deception.[58] Some scholars also propose the term "careless speech" to describe subtle inaccuracies in LLM outputs, which do not fit neatly into categories like libel, misinformation, or disinformation.[59] Careless speech captures the unique behaviour of LLMs and focuses on their potential impact beyond deliberate inaccuracy.

However, this thesis says that other terms do not fully describe the problem. The word “confabulation” is too broad and not specific enough to explain what happens with LLMs. It also does not bring new ideas. Confabulation usually means making up information, but the problem studied here is not just about making things up. It also includes wrong, misunderstood, or incomplete facts, which the word does not cover well. The word “bullshit” suggests that false content from LLMs is not made on purpose, but it ignores that this content can still have some value. It also does not show how the model is affected by outside factors or its tendency to cause misunderstandings. The term “careless speech” is more creative but suggests a deliberate human action because “speech” usually means a planned way of speaking. In contrast, LLMs operate passively in response to prompts. Additionally, the notion of carelessness inaccurately suggests negligence, whereas the errors LLMs produce are better understood because of training data limitations and model architecture.

---

[57] Andrew L Smith, Felix Greaves and Trishan Panch, ‘Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models’ (2023) 2(11) PLOS Digital Health e0000388.

[58] Michael Townsen Hicks, James Humphries and Joe Slater, ‘ChatGPT is bullshit’ (2024) 26 Ethics and Information Technology.

[59] Sandra Wachter, Brent Mittelstadt and Chris Russell, ‘Do large language models have a legal duty to tell the truth?’ (2024) 11(8) Royal Society Open Science 240197

To ensure scientific rigor and terminological consistency, this thesis retains the term “hallucination”. This term not only offers greater conceptual precision and novelty, but also vividly captures the generation of incorrect or incomplete content by LLMs because of both internal mechanisms and external factors.

Furthermore, researchers hold differing perspectives on the meaning of “hallucination”. The first category of scholars describes it as “the generation of seemingly realistic sensory experiences unconnected to any real-world input”[60] or “distinct undesirable outputs of LLMs” [61] . The second category defines hallucinations more narrowly as nonfactual statements.[62] Both categories address false content but differ in focus. The first emphasises input-output coherence, while the second ensures factual accuracy. This distinction shapes regulation, as the definition to some extent determines the goals and scope of regulatory efforts. An overly broad or vague definition may lead to unfocused regulation that lacks precision in addressing hallucinations effectively. Conversely, an overly narrow definition risks limiting the scope of regulation and may fail to account for emerging or unforeseen risks associated with future developments in LLMs.

However, as the definition of hallucination originates from pathology and psychology,[63] its definition should not be detached from its original context and

---

[60] Hussam Alkaissi and Samy I McFarlane, ‘Artificial Hallucinations in ChatGPT: Implications in Scientific Writing’ (2023) 15(2) Cureus e35179.

[61] Lei Huang and others, ‘A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions’ (2025) 43(2) ACM Transactions on Information Systems.

[62] Peter Henderson, Tatsunori Hashimoto and Mark A Lemley, ‘Where’s the Liability in Harmful AI Speech?’ (2023) 3 Journal of Free Speech Law 589.

[63] Fiona Macpherson and Dimitris Platchias (eds), *Hallucination: Philosophy and Psychology* (MIT Press 2013).

discussed solely in relation to LLMs. A better way to understand hallucinations in LLMs is to look at them from multiple disciplinary angles. By combining the original meaning of hallucination with how LLMs work, this thesis builds a definition that makes sense in this specific context. This means considering how LLMs function and why they produce certain kinds of errors. Regarding the mechanisms behind hallucinations in LLMs, some scholars describe these models as "stochastic parrots",[64] emphasising that LLMs generate text based on statistical patterns rather than true semantic understanding. While their outputs may appear fluent, they lack genuine comprehension and context, often producing plausible yet factually incorrect content.[65] Therefore, this thesis suggests that hallucinations in LLMs should be defined as content generated by models that appears logical, plausible, helpful, and confident, but contradicts the original source and no regard for the truth.[66]

2.3 Hallucinations Generated in LLMs Lifecycle

2.3.1 LLMs Lifecycle

---

[64] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) 610.

[65] Yejin Bang and others, 'A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity' in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers) (Association for Computational Linguistics 2023) 675; Nuno M Guerreiro and others, 'Hallucinations in Large Multilingual Translation Models' (2023) 11 Transactions of the Association for Computational Linguistics 1500.

[66] Katja Filippova, 'Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data' in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Association for Computational Linguistics 2020) 864; Luke Munn, Liam Magee and Vanicka Arora, 'Truth Machines: Synthesizing Veracity in AI Language Models' (2024) 39 AI & Society 2759.

Breaking down the life cycle of LLMs is essential for studying the generation of hallucinations because it is helpful to analyse and address the potential sources of these issues at each stage of development and deployment.

Currently, there is no consensus on how to classify the lifecycle of LLMs. Some discussion tends to focus on dividing the lifecycle of AI in general or Gen AI, rather than specifically addressing LLMs.[67] Furthermore, the methods for classification often swing between being overly simplistic and excessively complicated. For example, one approach divides the LLMs lifecycle into three stages: design, development, and deployment. Each phase requires specific human expertise and can be further divided into several stages, totalling 19 stages. [68] Another method uses a knowledge-flow approach, segmenting the LLMs lifecycle into five critical periods: knowledge acquisition, representation, probing, editing, and application.[69] Similarly, the Digital Health Centre of Excellence (DHCoE) outlines the LLMs lifecycle in seven phases, [70] a level of granularity that may be overburdensome. Such fine-grained segmentation is not adopted here because it risks creating unnecessary complexity and administrative burdens. Requiring

[67] Charlotte Stix, 'A Survey of the European Union's Artificial Intelligence Ecosystem' (arXiv, 28 December 2020) <https://doi.org/10.48550/arXiv.2101.02039> accessed 5 August 2025.

[68] Daswin De Silva and Damminda Alahakoon, 'An Artificial Intelligence Life Cycle: From Conception to Production' (2022) 3(6) Patterns 100489.

[69] Boxi Cao and others, 'The Life Cycle of Knowledge in Big Language Models: A Survey' (2024) 21(2) Machine Intelligence Research 217.

[70] Troy Tazbaz, and John Nicol, 'Blog: A Lifecycle Management Approach toward Delivering Safe, Effective AI-enabled Health Care' (*U.S. Food and Drug Administration*, 25 July 2024) <https://www.fda.gov/medical-devices/digital-health-center-excellence/blog-lifecycle-management-approach-toward-delivering-safe-effective-ai-enabled-health-care> accessed 31 March 2025.

separate documentation, assessment, and compliance processes for each stage may fragment governance, duplicate effort, and reduce regulatory efficiency.

This thesis proposes a classification method that considers both the relationship between knowledge flow and hallucinations while retaining the characteristics of the model itself. This approach introduces a lifecycle that facilitates the differentiation of hallucinations at various stages. Specifically, the lifecycle includes the following stages (Fig 2-1): Model Design, Pretraining, Fine-tuning and Alignment, Deployment and Interaction, and Monitoring and Iteration. With regard to the relevant actors, this thesis draws on the definitions of providers, deployers, and other actors set out in Article 3 of the AIA, while also introducing the category of users in light of the practical operation of LLMs.

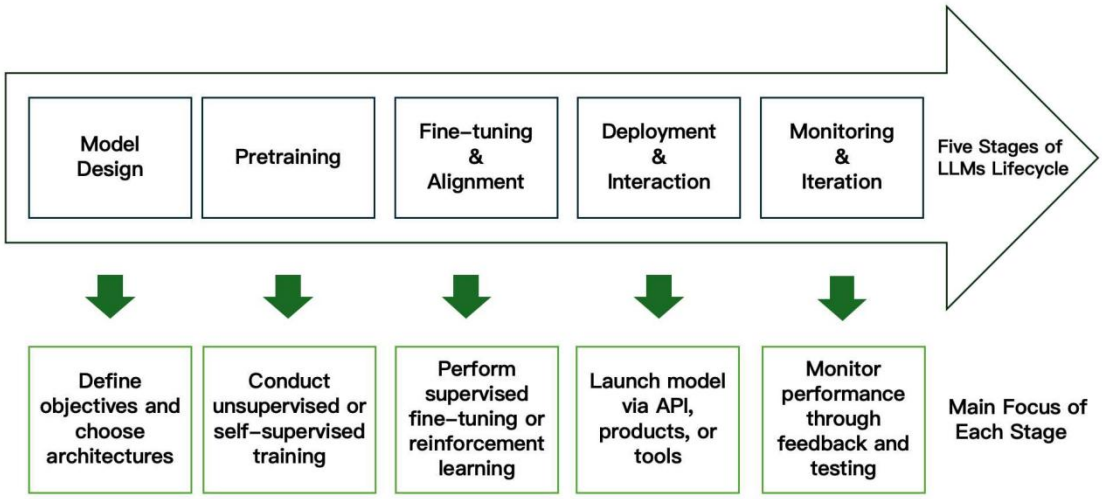

Fig. 2-1 Five stages and focus on lifecycle of knowledge in LLMs

2.3.2 Hallucinations in LLMs Lifecycle

2.3.2.1 Model Design

In the model design stage, providers define the model's purpose, architecture, and training objectives. A fundamental cause of hallucinations lies in the objectives and constraints set at the design stage. Most LLMs are optimised for next-token prediction rather than factual verification, so plausibility may be treated as sufficient even where accuracy is required. The standard language modelling objective of predicting the next token from large-scale internet datasets is inherently misaligned with the goal of answering user queries truthfully. Without design features that promote factual accuracy or enable verification, the model has no inherent capacity to distinguish truth from falsehood and may produce confident but fabricated responses.

Another design-level factor is the scope and capability boundaries set for the model. If the model's intended domain or use-case is not clearly scoped, it may be deployed in zero-shot or few-shot settings to perform tasks beyond its reliable capacity, thereby increasing the risk of hallucinations and unreliable outputs.[71] For example, a general-purpose model may be used to answer specialised legal or medical questions. Without a refusal or fallback mechanism for queries beyond its competence, it may generate a plausible but hallucinated answer.

Moreover, where the model lacks grounding mechanisms, such as retrieval tools or structured knowledge bases, it must rely solely on internal statistical patterns rather than verified facts, increasing the likelihood of plausible but incorrect outputs. Despite the long scholarly discussions, algorithms trained on biased data continue to perpetuate existing inequalities and lead to adverse outcomes on

[71] Zihao Zhao and others, 'Calibrate before Use: Improving Few-Shot Performance of Language Models'in *Proceedings of the 38th International Conference on Machine Learning* (PMLR 2021) 12697 <http://proceedings.mlr.press/v139/zhao21c.html> accessed 2 August 2025.

social, cultural, and political plans.[72] This is the reason why the regulatory frameworks are expected to incorporate mechanisms for auditing and mitigating bias risks, which is a task difficult to be fulfilled by proprietary algorithms used by private companies.[73] Empirical studies have shown that grounding LLMs with external retrieval systems significantly reduces the incidence of hallucinations and improves factual accuracy.[74] For example, many domain-specific systems use retrieval-augmented generation (RAG)[75] to connect LLMs to external knowledge sources and reduce hallucinations in areas such as law and finance. By contrast, models without such grounding mechanisms are more likely to produce unsupported outputs because they rely solely on internal parameters rather than verifiable external information. However, even retrieval-augmented approaches, such as RAG, are not immune to hallucination, as they may retrieve irrelevant or outdated content, misinterpret retrieved documents, or generate inaccurate outputs based on loosely related sources. In summary, design choices regarding objectives and knowledge integration set the stage for hallucination propensity. If factual accuracy, domain limitations, and uncertainty handling are not baked into the model's design, hallucinations are likely to emerge downstream.

---

[72] Russell Belk, 'Ethical Issues in Service Robotics and Artificial Intelligence' (2021) 41(13—14) The Service Industries Journal 860.

[73] Sara Hajian, Francesco Bonchi and Carlos Castillo, 'Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining' in P*roceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM 2016) 2125.

[74] Patrick Lewis and others, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks' in *Advances in Neural Information Processing Systems* 33 (2020) 9459; Sebastian Borgeaud and others, 'Improving Language Models by Retrieving from Trillions of Tokens' in *Proceedings of the 39th International Conference on Machine Learning* (PMLR 2022) 2206.

[75] Samar AboulEla and others, 'Exploring RAG Solutions to Reduce Hallucinations in LLMs' in *2025 IEEE International Systems Conference* (SysCon) (IEEE 2025) 1; Grégoire Mialon and others, 'Augmented Language Models: A Survey' (arXiv, 15 February 2023) <https://doi.org/10.48550/arXiv.2302.07842> accessed 2 August 2025.

2.3.2.2 Pretraining

After design, LLMs enter the pretraining stage, where they learn linguistic patterns from vast text corpora through self-supervised learning. Because this process relies on statistical pattern recognition rather than truth verification, it creates an important source of hallucination risk.

Firstly, the training data itself is sourced largely from the internet, which contains a significant amount of false, inaccurate, out-of-date, or contradictory information.[76] LLMs rely on massive amounts of textual data sourced from the internet, constituting the foundational element that underpins their operations. These models collect extensive datasets from the internet to train their algorithms, wherein the textual data serves as training samples during acquisition, commonly known as training data.[77] Based on this principle, hallucination also arises due to the training phase's pattern generation techniques and the absence of real-time internet updates, [78] language barriers and legal barriers, contributing to discrepancies in the information output.

---

[76] Ekin Akyürek and others, 'Towards Tracing Knowledge in Language Models Back to the Training Data' in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Association for Computational Linguistics 2022) 2429.

[77] Guilherme Penedo and others, 'The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only' (arXiv, 1 June 2023) <https://arxiv.org/abs/2306.01116> accessed 26 November 2024.

[78] Partha Pratim Ray, 'Web3: A Comprehensive Review on Background, Technologies, Applications, Zero-Trust Architectures, Challenges and Future Directions' (2023) 3 Internet of Things and Cyber-Physical Systems 213; Claudia E Haupt and Mason Marks, 'AI-Generated Medical Advice—GPT and Beyond' (2023) 329(16) JAMA 1349.

Secondly, hallucinations may result from inappropriate training corpora. Where the data includes large amounts of social media or fictional content, the model may learn fluent but unreliable patterns, which becomes particularly problematic in contexts requiring high factual accuracy. In some advanced scenarios such as medical diagnosis, there might be not enough training data.[79] In this situation, LLMs trained on general corpus might not be able to generalise well to specific domains or new knowledge due to the lack of domain-specific knowledge or new training data and may incorrectly diagnose a disease.[80] Moreover, legal and ethical limits on data use, such as restrictions relating to copyright or personal data, may also create knowledge gaps in training. Although often necessary, such constraints can leave the model without access to important information, increasing the risk of fabrication. Similar problems arise from language barriers, where the model lacks sufficient exposure to information available primarily in other languages.

2.3.2.3 Fine-tuning and Alignment

In the fine-tuning and alignment stage, models are trained to align with human intentions and values, aiming to develop trustworthy, responsible, and beneficial AI.[81] This is achieved through supervised learning and reinforcement techniques

[79] Shirui Pan and others, 'Unifying Large Language Models and Knowledge Graphs: A Roadmap' (2024) 36 (7) *IEEE Transactions on Knowledge and Data Engineering* 3580.

[80] Jindong Wang and others, 'On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective' (arXiv, 22 February 2023) <https://doi.org/10.48550/arXiv.2302.12095> accessed 24 March 2025.

[81] Tianhao Shen and others, 'Large Language Model Alignment: A Survey' (arXiv, 26 September 2023) <https://doi.org/10.48550/arXiv.2309.15025> accessed 9 July 2025.

such as Supervised Fine-Tuning (SFT)[82] and Reinforcement Learning from Human Feedback (RLHF)[83], which further refine the model for specific tasks. While these approaches improve alignment with human preferences, they also introduce new sources of hallucination. SFT improves model capabilities through annotated (instruction, response) pairs and enhances generalisation, but it may also reinforce subtle inaccuracies or misleading patterns.[84] RLHF aligns outputs with human preferences using a preference model, yet it can reward persuasive but false responses, increasing high-confidence hallucinations.[85]

Fine-tuning may also prioritise goals such as helpfulness, harmlessness, or user satisfaction over factual accuracy. As a result, a model may appear polite and well aligned while still hallucinating, especially where it is trained to avoid uncertainty and respond rather than admit that it does not know.

Moreover, the fine-tuning process is also inherently limited in scope. It typically targets a subset of common tasks or domains, leaving many unaligned areas where the model may revert to pretraining behaviours and generate hallucinations.

---

[82] Yue Zhang and others, 'Alleviating Hallucinations of Large Language Models through Induced Hallucinations' in *Findings of the Association for Computational Linguistics: NAACL 2025* (Association for Computational Linguistics 2025) 8218.

[83] Long Ouyang and others, 'Training Language Models to Follow Instructions with Human Feedback' in Advances in Neural Information Processing Systems 35 (2022) 27730.

[84] Zorik Gekhman and others, 'Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?' in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2024) 7765.

[85] Patrick Fernandes and others, 'Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation' (2023) 11 Transactions of the Association for Computational Linguistics 1643.

2.3.2.4 Deployment and Interaction

At this stage, the model is formally deployed and integrated into real-world applications, where it interacts with users in dynamic and often unpredictable environments. Hallucinations during this phase stem from several interrelated factors.

A primary factor is the nature of user prompts. In practice, prompts may be ambiguous, incomplete, or based on false assumptions. Because LLMs typically respond to the prompt as given rather than verify or clarify it, they may reinforce false premises or fill in missing details, thereby generating hallucinations. This risk is especially high where user queries are vague, or complex and the model lacks a mechanism to seek clarification.

Secondly, the complexity and diversity of deployment scenarios, which range from legal advice chatbots to educational platforms, introduce further risks. Models that are trained on general-purpose data may struggle to meet specific domain requirements, resulting in inappropriate or inaccurate outputs in high-risk settings. However, the models lack mechanisms to reject answering due to business pressures or design constraints, instead forcibly responding to user demands.

2.3.2.5 Monitoring and Iteration

The final stage of the lifecycle involves monitoring the model in operation and iteratively improving it based on feedback, error reports, and new data. Even after deployment, developers and organisations must continuously observe how

the LLMs are performing, especially tracking incidents of hallucinations. Hallucinations at this stage might not constitute a completely new category, as they often reflect unresolved issues originating from earlier phases such as design, training, and interaction. However, the crucial factor is whether these issues are detected and addressed.

A major difficulty is the delayed detection of hallucinations. Unlike overtly harmful content, subtle factual errors are often hard to identify and may go unreported. As a result, developers may remain unaware of repeated inaccuracies until after they have already influenced users or spread more widely.

Furthermore, for hallucination issues that have already been reported, due to the need to quantify and accurately identify the root causes, tracking needs to be done appropriately based on retrospective records, which means that the speed of modification and correction cannot match the demands for iteration and optimisation.

2.4 The Harms Caused by Hallucinations Risks

Hallucinations in LLMs can cause harms at both the micro and macro levels, affecting individuals and organisations as well as wider social and systemic interests. Existing study[86] has begun to map such harms by linking failures, harms, and stakeholders, or by classifying risks according to the responsible entity, its intent, and the stage at which the risk arises. Their work contributes to an understanding of the sociotechnical factors that precipitate AI risks and harms,

[86] Peter Slattery and others, 'The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence' (arXiv, 14 August 2025) <https://doi.org/10.48550/arXiv.2408.12622> accessed 9 July 2025.

which may help develop novel harm mitigation strategies and identify the stakeholders best positioned to address specific risks. However, in addition to identifying only coarse-grained causal factors, their work does not identify who tends to experience each type of risk or harm, limiting its utility to prioritise risks and map accountability pathways. To address this gap, Vel´azquez[87] applied a "what-where-who" framework to analyse 639 AI incident reports. They used large language models to operationalise their framework and automatically classify each incident based on what type of harm occurred, where it originated, and who was harmed. They found that most incidents in their dataset harmed the individual directly interacting with the AI system. Hutiri, Papakyriakopoulos and Xiang[88] arrived at a similar conceptual framework to characterise the harms of speech generators, mapping harms from their responsible to affected entities. However, they found that most harms did not affect the individual who interacted with the speech generator. Taken together, these papers highlight the importance of mapping harms to the stakeholders they affect. They also suggest that there are critical differences between traditional AI harms and Generative AI harms in terms of who they affect and how they materialise. Some scholars construct a taxonomy specifically for Gen AI failures and map them to the harms they're associated with in the real world. Although this study provides a valuable analysis of 499 publicly reported incidents of harm involving generative AI, its taxonomy of harms is overly complex for practical classification purposes.[89] Taken together, however, the

[87] Julia De Miguel Velázquez and others, 'Decoding Real-World Artificial Intelligence Incidents' (2024) 57(11) Computer 71.

[88] Wiebke Hutiri, Orestis Papakyriakopoulos and Alice Xiang, 'Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators' in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024) 359.

[89] Megan Li and others, 'A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents' (2025) 8(2) *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 1561.

existing studies reveal common patterns that support a more simplified and analytically useful categorisation of harms.

Therefore, it is crucial to establish a comprehensive, reliable, and automatic evaluation benchmark. Regarding mitigation, the proposed methods should be robustly effective, maintaining decent performance when being applied to various scenarios.[90]

The impact of hallucinations depends on the stage of the LLM lifecycle at which they arise and how they reach end users. This thesis therefore classifies hallucination related harms into two levels, micro and macro, according to the scale and subjects affected (Table 2.1).[91] Micro level harms affect individual users or specific organisations, whereas macro level harms arise when such risks accumulate and affect wider social systems or functions. This distinction captures both the immediate and structural consequences of hallucinations and helps clarify the different regulatory responses they require.

Harms in Lifecycle Stages	Micro-Level Harms	Macro-Level Harms
Model Design	Misunderstanding	Structural flaws

[90] Yue Zhang and others, 'Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models' (2025) 51(4) Computational Linguistics 1373.

[91] This thesis does not adopt a meso-level framework to capture operational, legal, and group-level harms, such as strategic missteps, compliance risks, or representational biases. Instead, it adopts a micro-macro distinction for the sake of analytical clarity. These intermediate phenomena are addressed either at the micro level when they concern identifiable entities, such as a company's legal exposure, or at the macro level when they relate to broader institutional patterns or collective outcomes.

	Credibility confusion	Public distrust
Pretraining	Cognitive biases	Bias amplification
	Personal accountability	Accountability loss
Fine-Tuning & Alignment	Over trust	Stereotypes
	Usability vs. integrity	Integrity erosion
Deployment & Interaction	Misjudgement	Information pollution
	Confidence confusion	Institutional failure
Monitoring & Iteration	Trust loss	Regulatory gap
	Normalisation	Governance decline

Table 2-1 Harms in Lifecycle Stages

	Micro level	Macro level
Misinformation	Cognitive Misguidance and the Acceptance of Misinformation	Public Information Pollution and Social Cognitive Disruption
Trust Erosion	Overreliance and Decline of Trust	Erosion of Public Trust and Acceptance Toward the AI Industry
Regulatory and Governance Gaps	Loss of Effective Feedback Mechanisms	Innovation Chilling

Table 2-2 Harms at micro level and macro levels

Based on the table above, hallucinations can arise at various stages throughout the lifecycle of LLMs. However, the hallucinations at different stages often overlap in nature and effect. Therefore, similar types of hallucinations can be grouped and systematised into broader categories of harms, providing a more structured and coherent analytical framework. Drawing on the mapping of hallucinations in Table 2-1 and summarising the overlaps across lifecycle stages, this thesis classifies the primary types of harms into three categories (Table 2-2): (1) misinformation, (2) trust erosion, and (3) regulatory and governance gaps.

In the context of this thesis, misinformation refers to inaccurate, unverified, or distorted content generated by hallucinations in LLMs. Such information is typically unintentional, arising from technical limitations such as insufficient data or contextual errors or from systemic flaws, rather than from malicious manipulation. This distinguishes misinformation from disinformation, which refers to deliberately fabricated or misleading content spread with the intent to manipulate, deceive, or harm. Conceptually, misinformation is closer to a cognitive level "misunderstanding," whereas disinformation is anchored in intentionality.

Trust erosion refers to the gradual loss of trust by users, businesses, and institutions in the LLMs itself, its developers, and even the broader industry, following repeated encounters with hallucinated content. The decline in trust isn't just about how individual users feel; it can have far-reaching consequences across systems. On one hand, distrust poses a threat to innovation. If the public or regulators lose confidence in Gen AI, they might impose tougher regulations, restrict data usage, or make the approval process for new products more challenging. Such barriers can slow progress, limit opportunities for research and experimentation, and lead companies to opt for safer, less imaginative designs. In critical fields like healthcare, law, or public services, if leaders remain sceptical

of Gen AI and decide against its implementation, the anticipated advancements in efficiency and quality may never be realised.

On the other hand, a loss of trust can also lead to users leaving. When people stop trusting the system, they may use it less or stop using it completely. This can hurt the growth and health of the AI market. For LLMs, which need user feedback, interaction, and data to keep getting better, this is a serious problem that can slow down improvement of technology.

Finally, missing rules and weak governance mean that today's laws, industry guidelines, and ethical standards are not enough to fully address the harms caused by hallucinations in LLMs. These harms affect different stakeholders in different ways and unfold across both micro and macro levels, creating distinct impacts that require careful analysis.

2.4.1 Harms on Micro Level

2.4.1.1 Cognitive Misguidance and the Acceptance of Misinformation

Misinformation caused by hallucinations makes it hard for users to tell if a model's output is correct, especially when the model gives information in a clear and confident way, with easily, widely, and cheaply by attributing a high probability to false or misleading claims.[92] This clear style makes the wrong information seem more true and harder to spot. Misinformation may facilitate fraud, scams,

[92] Claudio Novelli and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (2024) 55 Computer Law & Security Review 106066.

targeted and non-targeted manipulation and cyber attacks.[93] If users trust wrong answers without checking, they might form false knowledge in areas like learning, decision-making, and communication.

For example, in medical consultations, a user who believes false health advice from the model might wait too long to get real help or take the wrong steps to care for their health. In academic writing, citing fabricated sources or references generated by the model can undermine academic integrity and damage scholarly standards. In sharing news, people may get the wrong idea about social events if they read or pass on false or incorrect details created by the model.

Moreover, hallucinations can arise early in the model's lifecycle. For example, false or biased data used during pretraining can introduce errors. During fine-tuning and alignment, the model may be overfit to user preferences or reinforce existing biases in the data. As a result, users might receive stereotyped, discriminatory, or misleading content, which limits diverse perspectives and undermines fairness. Over time, repeated use of models prone to frequent hallucinations may erode users' knowledge structures, creating a self-reinforcing chain of misinformation where even if the model later ceases to hallucinate, the user's internalised cognitive biases persist and propagate errors independently.

2.4.1.2 Overreliance or Decline in Trust

The occurrence of hallucinations can significantly affect users' cognition, ultimately weakening individual cognitive capacities. Gen AI, due to its powerful content-generation capabilities, has lowered the cost of knowledge acquisition

[93] Laura Weidinger and others, 'Ethical and Social Risks of Harm from Language Models' (arXiv, 8 December 2021) <https://doi.org/10.48550/arXiv.2112.04359> accessed 9 July 2025.

and become a kind of modern “encyclopaedia,” offering substantial convenience to users. However, Gen AI does not possess true understanding or knowledge; it merely recombines and reassembles data to produce new content. As a result, the information it generates is often interspersed with false or misleading “knowledge.”

When individuals engage deeply with AI systems, they become highly susceptible to absorbing this “knowledge,” which results in what can be called a knowledge hallucination, a distortion of their personal knowledge framework and cognitive accuracy. Over time, seeing these mistakes repeatedly can slowly weaken thinking skills. It becomes harder to tell what is true or false in a large and confusing information space. This harm to thinking can make daily choices and judgment worse, leading to serious problems for the person.

Also, LLMs talk in a smooth and confident way and can talk about many topics. This makes users think the model can be trusted and that “everything it says is correct.” This is especially common in new users or people who do not have expert knowledge, who may develop overreliance on the system and come to treat it as an authoritative source.

Conversely, when users encounter repeated hallucinations, such as discovering that the model’s outputs are wrong, made up, or don’t match, their initial trust may collapse. This breakdown of trust can even extend to scepticism toward the broader AI system or technology. As a result, user engagement with LLMs products may decline, indirectly impeding the development and advancement of LLMs technologies.

2.4.1.3 Loss of Effective Feedback Mechanisms

In the deployment and user interaction stages, current LLMs require robust feedback and improvement mechanisms to help reduce hallucinations. User feedback is not only a critical channel for detecting hallucinations but also a key driver for model enhancement. However, hallucinations themselves can, to some extent, hinder the effectiveness of feedback mechanisms: users often cannot promptly recognise hallucinations, making it even more difficult to provide meaningful feedback.

Firstly, many users fail to detect hallucinations when they occur. For example, when dealing with information that requires verification, users may only discover errors after completing external checks by which point the window for submitting feedback has often passed. More typically, when users rely on LLMs to acquire new knowledge, they may remain entirely unaware that the information provided is erroneous.

Secondly, even when users do identify hallucinations, the available feedback channels are often unfriendly or inaccessible. For instance, there may be no simple button to report “this answer is incorrect,” or the feedback entry point may be deeply buried in the interface. Even when there are ways to give feedback, users often do not trust them. They worry that their comments will be ignored or will not lead to real changes in the system.

Without clear responses or noticeable improvements, users slowly stop giving feedback and become disappointed. This weakens the system’s ability to improve itself and may cause users to leave.

2.4.2 Harms on Macro Level

2.4.2.1 Public Information Pollution and Social Cognitive Disruption

When hallucination content spreads widely on social media, news sites, search engines, and other platforms, it causes big problems for the information system such as fake news, consists of false claims that may seem verifiable but are based on fabricated facts, often distorted or manipulated from real events, and it is designed to trigger an emotional response, aiming to deceive readers and influence their opinions through implicit conclusions,[94] meaning the goal is persuasion, not informing.[95] Selakovic[96] highlighted that fake news can significantly influence various aspects of government functioning, while other scholars, such as Fakhry et al.[97], emphasised that it is increasingly used as a tool in the political arena. Furthermore, some researchers[98] have pointed out the difficulties in governing public discourse shaped by AI systems like ChatGPT.

---

[94] Nicolas Belloir, Wassila Ouerdane and Oscar Pastor, 'Characterizing Fake News: A Conceptual Modeling-based Approach' in Jolita Ralyté and others (eds), *Conceptual Modeling: 41st International Conference, ER 2022, Hyderabad, India, October 17–20, 2022, Proceedings* (Lecture Notes in Computer Science vol 13607, Springer 2022) 115.

[95] Edson C Tandoc Jr, Zheng Wei Lim and Richard Ling, 'Defining "Fake News": A Typology of Scholarly Definitions' (2018) 6(2) Digital Journalism 137.

[96] Marko Selaković, 'Fake News and Foreign Direct Investment Inflows: Is there a Relationship?' (2022) 14(2) European Journal of Interdisciplinary Studies 24.

[97] Baher Fakhry, Anna Tarabasz and Marko Selakovic, 'Social Media & Uprisings: The Case of the Egyptian Revolution in 2011' (2023) 377 *MATEC Web of Conferences* 02002.

[98] Rihan Huang and Haolong Yao, '"Reshaping" and "High Risk": The Impact of Generative Artificial Intelligence on Public Opinion Security' ('"再塑造"与"高风险": 生成式人工智能对舆论安全的影响') (2024) 43(4) Journal of Intelligence 121.

Firstly, users have trouble telling the difference between trusted information and content made by Gen AI. This makes information seem less reliable and harder to trust.

Secondly, hallucinated content can help spread false information. It can travel fast and reach many people, especially during important events or on sensitive topics. Made-up facts or unclear answers from LLMs often get shared repeatedly on different platforms and grow quickly across networks.

Finally, information pollution causes confusion in groups. Different users may believe opposite things because of hallucinated content. This breaks the shared facts that people need to agree on issues. Without common facts, public conversations become unclear, and it gets harder to have fair discussions or make group decisions. In a media environment increasingly shaped by algorithmically generated content, hallucinations may exacerbate epistemic fragmentation, where different segments of the population operate under diverging informational realities.

2.4.2.2 Erosion of Public Trust and Acceptance Toward the AI Industry

The frequent exposure of hallucination problems gradually erodes public trust in AI technologies, thereby weakening the push for innovation.

Firstly, regular users feel misled and start to doubt the technology when they see wrong information. This makes them use it less or stop using it completely. Business clients are also careful because they worry that AI tools with hallucinations might hurt their reputation, break rules, or cause financial loss.

Because of this, using AI in important areas like healthcare, finance, and law is often delayed.

Secondly, producers and operators may avoid testing and new ideas because they are afraid of hallucination problems. At the same time, governments may start to doubt whether tech companies are being honest or can do their job properly. This can lead to stricter rules, which may slow down the growth of the industry.

Overall, this loss of trust hurts the growth and progress of the whole AI system. It gets in the way of using new ideas and slows down innovation.

2.4.2.3 Harms to the Public from LLM Hallucinations as a Barrier to Innovation

The current legal and regulatory frameworks significantly hinder innovation by failing to adequately address the hallucination problem in LLMs.

Firstly, unclear allocation of responsibility creates a regulatory bottleneck that stifles innovation. When hallucinations cause harm, the ambiguous liability among developers, deployers, and third-party providers creates a “grey zone” where no actor is clearly accountable. This uncertainty discourages bold advancements, as innovators face potential but undefined legal risks. Furthermore, the lack of universally accepted standards for detecting, reporting, and mitigating hallucinations prevents the establishment of consistent governance norms, limiting the industry’s ability to self-regulate effectively.

Secondly, the slow pace of regulatory adaptation exacerbates the innovation barrier. Existing legal instruments and regulatory bodies are often reactive rather

than proactive, stepping in only after incidents occur or public pressure mounts. This delayed response hampers the rapid deployment and iterative improvement of LLM technologies, curtailing their full potential.

Together, these governance gaps erode public trust and create vulnerabilities that may be exploited commercially or maliciously, further deterring investment and slowing the sustainable growth of AI innovation ecosystems. Without timely and clear regulatory guidance, the innovation momentum in LLM development risks being significantly constrained.

2.5 Conclusion

This chapter examined the causes and risks of hallucinations in large language models across five key stages of their lifecycle: model design, pretraining, fine-tuning, deployment, and monitoring. It provided an interdisciplinary foundation for assessing regulatory responses. The chapter defined hallucination broadly to include both factual inaccuracies and structural distortions and adopts this term due to its widespread use and regulatory relevance.

Hallucinations stem from various sources, including design goals that prioritise fluency over factuality, the use of unreliable training data, reinforcement of errors during fine-tuning, misleading user prompts, and insufficient post-deployment oversight. Table 1 outlines the risks associated with each stage.

Despite differences across stages, the harms caused by hallucinations fall into three main categories: misinformation, trust erosion, and regulatory and governance gaps. As shown in Table 2, these harms occur at both the micro level, affecting users, developers, and specific groups, and the macro level, where they

contribute to the spread of false information, declining public trust in AI, and the hindrance of innovation.

By mapping these harms, the chapter showed that hallucinations are not only technical flaws but also socio-technical challenges requiring layered legal and policy interventions. Without effective regulation, hallucinations can undermine trust, impair decision-making, and cause widespread harm. The following chapter explores how the EU and China address these issues in their AI regulatory frameworks, assessing their strengths, weaknesses, and areas in need of reform.

Chapter 3 - Legal Perspective Based on Technology: LLMs Regulation in the EU and China

3.1 Introduction

Considering the technical underpinnings of hallucinations in LLMs established in Chapter 2, it is imperative to examine how the law respond to these challenges. The global race for AI dominance has positioned AI governance as a critical national priority, with countries striving to enhance their competitiveness and influence in the field.[99] As a result, national AI strategies across different countries often exhibit remarkably similar narrative constructions, portraying AI as a disruptive technological force capable of fundamentally transforming society and politics.[100] This chapter shifts focus from the technological roots of hallucinations to the regulatory landscape in the EU and China. It analyses how each jurisdiction's approach solves (or fails to address) the problem of hallucination harms at different stages of the LLMs lifecycle theoretically and practically, bridging the gap between technology and law.

This chapter is a "legal perspective based on technology", meaning the regulatory analysis is structured around the LLMs lifecycle. By aligning legal obligations with the stages where hallucinations emerge (from model design and pretraining to deployment and monitoring), we can critically evaluate whether current laws effectively target the points of vulnerability identified in Chapter 2. The EU and China provide an instructive contrast: the EU is on the verge of implementing a

[99] Fernando Filgueiras, 'The Politics of AI: Democracy and Authoritarianism in Developing Countries' (2022) 19(4) Journal of Information Technology & Politics 449.

[100] Jana Bareis and Christian Katzenbach, 'Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics' (2022) 47(5) Science, Technology, & Human Values 855.

comprehensive, risk-based AIA, while China has rapidly issued targeted rules (such as the Interim Measures) without a single omnibus AI law. Both approaches aspire to balance innovation with risk management, yet they reflect different legal cultures and governance philosophies. In addition, this chapter adopts a tech and law method, recognising that LLMs operate within complex socio-technical systems, which can be analysed by real cases. Thus, regulations are not assessed in a vacuum; their practical enforcement, industry uptake, and broader governance context are considered to gauge real-world effectiveness.

Structurally, Section 3.2 examines the EU's approach to LLMs regulation. Firstly, it outlines the theoretical framework of EU AI governance, including the multi-layered regulatory approach and key legal provisions in AIA relevant to hallucinations. This theoretical analysis maps how EU law attempts to ensure data quality, transparency, and accountability throughout the LLM lifecycle, as these factors affect hallucinations. It then looks at how these rules work in practice, including the challenges in applying them and whether they can reduce harm. Section 3.3 moves to China's approach. It first gives a look at national plans and rules for different sectors, then checks how well these rules are carried out. China has been active in making rules for algorithms and Gen AI, but problems like overlapping rules and weak enforcement still bring doubt about how well hallucinations are being handled.

3.2 LLMs Regulation in the EU

3.2.1 Theoretical Framework: Regulatory Approach and Legal Provisions

3.2.1.1 Regulatory Approach in the EU

The EU, as a supranational entity characterised by multilevel governance, aims to harmonize fragmented decision-making among member states to foster integration and enhance global competitiveness.[101] Guided by "human-centred" cultural values,[102] the EU enforces strict market regulations that influence global standards through the "Brussels Effect" [103], whereby non-EU entities adopt EU standards to access its extensive market. The EU has taken a proactive stance with regard to social effects of AI, implementing stringent regulations aimed at fostering competitiveness while prioritising ethical considerations, enhancing privacy protections, and mitigating potential harms.[104]

In the context of AI governance, the EU prioritises its regulatory power to establish ethical standards for AI, taking an incremental approach to ensure that AI systems align with citizens' rights and European values. These efforts aim to position the EU as a global leader in AI ethics, while simultaneously promoting market integration, coordinating member states, and maintaining competitiveness through a "state-market regime" (state actors recognise the necessity of establishing new regulations to reshape market framework conditions).[105]

---

[101] Troels Krarup and Maja Horst, 'European Artificial Intelligence Policy as Digital Single Market Making' (2023) 10(1) Big Data & Society; Ronit Justo-Hanani, 'The Politics of Artificial Intelligence Regulation and Governance Reform in the European Union' (2022) 55(1) Policy Sciences 137.

[102] Stephen Cory Robinson, 'Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI)' (2020) 63 Technology in Society 101421.

[103] Anu Bradford, 'The Brussels Effect' (2015) 107 Northwestern University Law Review 1, 5.

[104] Jon Chun, Christian Schroeder de Witt and Katherine Elkins, 'Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US' (arXiv, 5 October 2024) <https://doi.org/10.48550/arXiv.2410.21279> accessed 21 October 2025.

[105] Troels Krarup and Maja Horst, 'European Artificial Intelligence Policy as Digital Single Market Making' (2023) 10(1) Big Data & Society 205395172311538; Huw Roberts and others, 'Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical

The EU controls LLMs using a system that mixes special AI laws with current legal rules. It uses both strict and flexible tools. This way helps deal with hallucinations by pushing for trusted AI through rules about good data, being clear, and being responsible. These rules can be used to manage hallucinations.

In the area of hard law, the AIA represents the main legislative instrument for regulating LLMs. As pioneering legislation, the AIA provides a comprehensive approach to mitigating risks posed by AI. It adopts a risk-based classification framework, distinguishing AI systems into categories of unacceptable risk, high-risk, limited risk, and minimal risk. Although general-use LLMs, such as GPT-4 or Gemini, are not automatically classified as "high-risk" under the AIA, the legislation explicitly acknowledges the need for oversight of general-purpose AI (GPAI) systems, frequently referred to as foundation models, upon which LLMs are typically built. Notably, during the legislative drafting process, EU policymakers extensively debated the scope and mechanisms for regulating foundation models. Consequently, the finalised Act contains explicit provisions aimed specifically at providers of GPAI systems, colloquially termed the "AI Pact" for foundation models.[106] These provisions impose duties related to transparency, performance, and risk management on the providers of LLMs, even when the models are subsequently adapted for diverse end-use scenarios. Such measures serve to

Outcomes' (2023) 39(2) The Information Society 79; Louisa von Essen and Marinus Ossewaarde, 'Artificial Intelligence and European Identity: The European Commission's Struggle for Reconciliation' (2024) 25(2) European Politics and Society 375.

[106] European Commission, 'AI Pact' (*European Commission*, updated 19 June 2025) <https://digital-strategy.ec.europa.eu/en/policies/ai-pact> accessed 20 June 2025; European Commission, 'Over a hundred companies sign EU AI Pact pledges' (*European Commission*, 25 September 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4864> accessed 20 June 2025.

promote reliable AI operation and significantly reduce associated risks, including hallucinations.

Other EU legal instruments also contribute to the governance of AI outputs. However, this thesis focuses exclusively on legislation explicitly designed for the regulation of AI. Accordingly, general legislative frameworks that may indirectly address hallucination risks are referenced where relevant but excluded from the core legal analysis such as the DSA or the GDPR.

In summary, while hard law, particularly the AIA, constitutes a comprehensive regulatory framework addressing AI broadly, it indirectly tackles the phenomenon of hallucinations. Although hallucinations are not explicitly identified as standalone regulatory targets, legislative measures emphasising trustworthy AI significantly mitigate associated risks. The EU adopts this multi-layered legislative approach because AI technologies, including LLMs, encompass complex, rapidly evolving, and broad-ranging applications, rendering a singular legislative measure insufficient. This overarching legislative strategy grants flexibility for subsequent Codes of Practice and permits member states to legislate in alignment with national specificities. By integrating dedicated AI-specific regulations with existing legal frameworks and employing both hard and soft law mechanisms, the EU aims for a flexible yet robust governance framework capable of adapting to technological advancements, while maintaining clear standards for safety, transparency, and accountability. This strategy is particularly advantageous for nuanced issues such as hallucinations, as it fosters trustworthy AI practices both directly and indirectly through measures addressing data quality, transparency obligations, and accountability. Nevertheless, it remains undeniable that, owing to the AIA's prominent risk-based approach, the lack of specific focus on Gen AI, and the absence of hallucinations as an explicitly addressed regulatory concern, further regulatory efforts are necessary to comprehensively tackle the risks posed by hallucinations.

3.2.1.2 Legal Provisions Governing LLMs in the EU

Following the earlier examination of regulatory approaches, the EU AIA provides the foundational framework for promoting AI trustworthiness[107] and addressing LLM hallucinations through a structured, risk-based regulatory framework designed to tackle challenges at every stage of the LLMs lifecycle. Recital 70 highlights that obtaining comprehensible information about the development and operational processes of high-risk AI systems throughout their lifecycle is essential for achieving system traceability. Similarly, mitigating hallucinations requires tracing back across various stages of the lifecycle and utilising traceable information for retrospective analysis whether through technological advancements or regulatory interventions. Thus, evaluating whether the AIA, as a primary legislative instrument, effectively meets theoretical requirements for regulating hallucinations necessitates examining the legislative provisions in relation to each lifecycle stage. Specifically, Article 15 of the AIA mandates that high-risk AI systems "shall be designed and developed to achieve accuracy, robustness, and cybersecurity throughout their lifecycle." Furthermore, Article 53 outlines general obligations for providers of GPAI models, emphasising the necessity of ensuring the trustworthiness of outputs as comprehensively as possible. In addition to legally binding provisions, the recitals of the Act also contain relevant guidance applicable to various stages of the LLMs lifecycle.

The AIA fundamentally distinguishes between the roles of provider, deployer, importer and distributor. Providers are those “placing on the market or putting into service AI systems or placing on the market general-purpose AI models in the

[107] Johann Laux, Sandra Wachter and Brent Mittelstadt, ‘Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk’ (2024) 18(1) Regulation & Governance 3.

Union irrespectively of their location. They have pre-market obligations including an initial risk assessment, risk-specific compliance, as well as risk-specific post-market obligations. Deployers are users of AI systems that are based on the EU and that don't fall within a small number of non-professionals use cases. The Act furthermore distinguishes between AI models or systems provided or deployed from within the EU, and those from outside. This distinguish is useful for select the obligation for different parts in lifecycle.

Specifically, in different stages of the lifecycle:

Firstly, during the model design stage, transparency is pivotal for preventing hallucinations. From the start of model design, the way the model works should be clear. If the model is intended to support user interaction, mechanisms for users to interact with the AI should be designed in advance. This helps make sure rules can be followed and sets a strong base for later tuning and use. Article 13 of the AIA requires deployers of high-risk AI systems to ensure transparency, enabling users to understand and interpret system outputs. Article 14 talks about human oversight. In paragraph 1, it says high-risk systems must have good ways for people to watch over them during use. This supports later steps in the LLMs lifecycle. Article 53 gives general rules for makers of GPAI models and says their outputs should be as trustworthy as possible. Besides these strict rules, the AIA's recitals also give helpful advice for different steps in the LLMs lifecycle.

Secondly, the quality of data in the pretraining stage directly affects the chance of hallucinations. Article 10 of the AIA sets strong rules for data management. It asks providers to make sure training data is relevant, correct, and complete, especially for high-risk AI systems. Specifically, rule (f) says to check for biases that could harm health, safety, or basic rights. This helps lower hallucination risks from bad data. Article 11 requires developers of high-risk AI systems to keep detailed technical records that clearly show data sources and their uses. These

rules help make sure data is collected legally and carefully checked, which cuts down hallucination risks early on. Also, Recital 68 says that people using high-risk AI should collect and use good quality data for their fields. Recital 107 adds that there should be more openness about the datasets used in pretraining and training GPAI models.

Thirdly, during fine-tuning and alignment, LLMs show how they keep knowledge through interaction and specific tests. Article 14 gives a legal basis for watching over these adjustments with human oversight. Also, Article 60 introduces regulatory sandboxes, which allow safe and controlled testing to find and reduce hallucination risks.

Fourthly, in the deployment and interaction stage, Article 15(4) primarily addresses cybersecurity by requiring high-risk AI systems to be designed with maximum resilience against errors, faults, or inconsistencies arising within the system or its operational environment. While this provision focuses on protecting against cyber threats, the required technical and organisational measures also indirectly enhance system reliability during deployment and interaction, underscoring the importance of robust safeguards in ensuring stable and trustworthy AI operations. Article 17(d) sets clear duties for providers of high-risk AI systems to fix problems. Providers must do regular checks, careful testing, and ongoing validation during the system's whole development. This helps improve output accuracy and lowers hallucinations by keeping the system updated.

Fifthly, at the monitoring and iteration stage, Article 20 sets rules for fixing problems and sharing information. Article 23 requires importers to maintain thorough records, conduct risk assessments, track AI systems, and cooperate with regulators. Although primarily aimed at regulatory compliance, these measures can indirectly contribute to the mitigation of hallucinations. These steps help make sure AI outputs are reliable before they reach users. Similarly, Article 24

gives distributors the job to check compliance, manage risks, and fix problems when needed. Article 26 tells deployers to watch how high-risk AI systems work and follow the instructions for use. This means if hallucinations from GPAI models cause harm, deployers must watch and act properly. Chapter IX of the AIA, called "Post-Market Monitoring, Information Sharing and Market Surveillance," sets the rules for ongoing oversight and helps find and lower hallucination risks after deployment. Additionally, Recital 31 highlights the prohibition of unacceptable scoring methods by AI systems that result in discriminatory or exclusionary outcomes for specific groups. Although regulation at this stage is typically post hoc in nature, it plays a critical role in limiting the proliferation of hallucination-induced harms and thus remains an essential component of the broader governance framework.

Although the AIA broadly addresses the entire lifecycle of LLMs, several limitations remain. Firstly, hallucination is sometimes not recognised as a significant risk; even when it is acknowledged, it is often treated merely as one of many risks within the broader general risk framework. The AIA aims to prevent and mitigate aggregated risks but does not explicitly identify hallucination as a distinct or exceptional category. Consequently, it lacks targeted measures that address the unique characteristics and societal implications of hallucinations generated by LLMs. Secondly, under Articles 2, 25, and 53 of the EU AIA, certain obligations are waived for open-source LLMs, effectively granting them partial exemption from liability. However, with the increasing proliferation of open-source models such as China's DeepSeek and others, this limitation has become increasingly prominent. The exclusion of such models from the regulatory scope raises concerns about risk governance and accountability in open development environments. Thirdly, while several recitals contain detailed and insightful provisions that could help mitigate hallucinations, sometimes even more comprehensively than the operative articles themselves, recitals are not legally binding. Their non-binding nature restricts their enforceability, limiting the practical impact of those potentially valuable

provisions. Lastly, many of the AIA's articles that are most relevant to mitigating hallucination risks, such as Article 12 on record-keeping and Article 13 on transparency and the provision of information to deployers, apply only to high-risk AI systems. This limited scope may reduce the effectiveness of the Act in addressing hallucination risks associated with general-purpose or lower-risk LLMs, suggesting the existence of regulatory gaps that deserve further consideration.

3.2.2 Practical Assessment: Effectiveness of the EU's Regulation

3.2.2.1 Implementation of EU AI Regulations

As of 2025, the main legal basis for the EU to regulate hallucination is the AIA. However, because hallucinations occur frequently and their nature keeps changing, the AIA may not be able to keep up with the speed of these developments. Since 2022, the rapid progress of Gen AI has greatly changed both the direction of AI technology and the structure of the AI industry. As Gen AI evolves, hallucinations have become more common due to the way LLMs work, and the harms they cause have also increased. Although the AIA has gone through several rounds of revision, its core framework remains relatively outdated and may not match the fast pace of Gen AI updates.

In addition, the breadth of the AIA's scope and the novelty of its provisions mean that implementation requires substantial administrative preparation. Each Member State must designate or establish national supervisory authorities for AI, and a new EU AI Office will coordinate enforcement across the Union. Setting up these bodies, training staff in the technical intricacies of AI, and developing procedures for compliance checks and sandbox programs is a work in progress. The AIA contains wide-ranging exemptions for providers of certain AI systems provided under free and open source software licenses Article 53-54. To be

exempt, the systems may not contain GPAI models that fall within the systemic risk category or otherwise exhibit unacceptable behaviour. AI regulatory sandboxes are controlled frameworks offering innovative companies a safe space to develop, train, validate, and test an innovative AI system.[108] The complexity of the task raises concerns about regulatory capacity: will national authorities have enough expertise and resources to audit LLMs providers for things like data quality or accuracy metrics? The EU is attempting to address this by facilitating knowledge-sharing and possibly concentrating expertise in the AI Office, but uneven readiness among Member States could mean inconsistent enforcement in the early years.

Another aspect of implementation is the development of technical standards to flesh out the AIA's requirements. The AIA deliberately uses broad terms (e.g., "appropriate level of accuracy" or "state of the art" risk mitigation) and delegates to European standardisation organisations (like CEN/CENELEC) the task of developing harmonised standards that, if followed, confer a presumption of compliance. Work has begun on standards for metrics of accuracy, robustness, bias detection, etc. For hallucinations, this might entail industry-agreed testing protocols to measure an LLM's factual accuracy rate on benchmark queries, or standards for dataset documentation quality. However, as of now, no finalised standards exist specifically for hallucination reduction, and the AIA's effectiveness will partly hinge on these technical guidelines. The absence of finalised standards and codes of practice now of the Act's coming into force could leave providers in a murky situation and they know the broad obligations but lack detailed "how-to" guidance. This gap may result in a compliance lag: some companies might take a cautious approach, delaying AI deployments until standards clarify the expectations, while others might proceed and risk later adjustments if their

[108] Thomas Buocz, Sebastian Pfotenhauer and Iris Eisenberger, 'Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?' (2023) 15(2) Law, Innovation and Technology 357.

interpretations don't match regulators'. To mitigate this, the European Commission has convened expert groups and is encouraging a voluntary Code of Conduct for generative AI in the interim. In mid-2023, for instance, major AI developers (OpenAI, Google, Meta, etc.) were invited to commit to voluntary safeguards ahead of the law. Several firms did pledge to watermark AI-generated content and share information under this AI Pact.[109] These voluntary moves, while non-binding, indicate that industry is bracing for the AIA and, in some cases, even starting to implement its transparency measures early (such as watermarking images from generative models).[110]

3.2.2.2 Mitigation of Hallucination Harms

This section employs tech and law method to analyse real-world case reported in the media to evaluate the effectiveness of the AIA and other regulatory frameworks in mitigating associated harms. The evaluation criteria are based on the standards outlined in Section 2.4, providing a benchmark for assessing mitigation efforts.

3.2.2.2.1 Mitigation of Harms on Micro Level

Firstly, in relation to mitigating cognitive misguidance and the spread of misinformation, hallucination-induced false outputs must be addressed

[109] Diane Bartz and Krystal Hu, 'OpenAI, Google, others pledge to watermark AI content for safety, White House says' (*Reuters*, 22 July 2023) <https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/> accessed 10 June 2025.

[110] OpenAI, 'Understanding the Source of What We See and Hear Online' (*OpenAI*, 7 May 2024) <https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online> accessed 10 June 2025.

throughout the entire lifecycle of LLMs, with particular emphasis on ensuring the quality and legality of data at the pretraining stage. High-quality and lawfully obtained training data are essential to minimising the generation of factually incorrect or unlawful outputs, thereby reducing the risk of misinformation.

A notable example occurred in 2024, when the data protection organisation NOYB filed a complaint against OpenAI, citing ChatGPT's repeated misrepresentation of a public figure's date of birth. This inaccuracy was deemed a violation of the GDPR's principle of data accuracy and led to further scrutiny by the Austrian and Italian data protection authorities. In response, OpenAI implemented citation functions and source attribution mechanisms in its services for EU users.[111] Such measures demonstrate a lifecycle-conscious approach to mitigating hallucinations: by addressing the legality and accuracy of data during the pretraining phase, they help prevent the infringement of individual rights through hallucinated outputs. Moreover, they reflect the expectations under the AIA concerning data quality and lawful data sourcing and represent a regulatory-compliant attempt to reduce downstream risks of hallucination-induced harm.

Elsewhere, practical measures have been adopted at the monitoring and iteration stage to detect hallucinated outputs in real time, prevent their presentation to users, and thereby reduce cognitive misguidance. For instance, clinical AI developers such as Mistral and Gemma 2 have implemented retrieval-augmented generation (RAG) pipelines combined with output verification scoring methods such as the Normalised Misinformation Score (NMISS) to evaluate the factual accuracy of generated responses. These mechanisms reportedly achieved over a

[111] Noyb, 'ChatGPT provides false information about people, and OpenAI can't correct it' (*noyb*, 29 April 2024) <https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it> accessed 27 June 2025.

30% reduction in hallucinated outputs in diagnostic simulation contexts.[112] Such practices exemplify a post-deployment strategy for hallucination mitigation, highlighting the importance of continuous monitoring and iterative improvement. They also reflect an emerging industry norm of integrating technical safeguards into the lifecycle of LLMs deployment, aligned with the AIA's emphasis on transparency, accountability, and performance monitoring throughout the AI system's operational phase. These tools, though technical in nature, are directly responsive to growing regulatory pressures around output integrity.

Secondly, when addressing overreliance or declining user trust in LLMs, current responses often avoid the root problem. Some impose restrictions or bans on AI-generated content to reduce dependence, which may ease short-term issues like regulatory uncertainty or liability but fail to address deeper cognitive risks of overdependence. [113] Tagging AI-generated content is a common mitigation strategy, but it has limitations. Users may ignore or misunderstand tags, and inconsistent application across platforms reduces their effectiveness in encouraging critical engagement.[114] A more effective regulatory approach would allow probabilistic reasoning and generative functions within clear, context-specific limits. This supports users in creative or assistive roles while preventing unchecked reliance. Defining application boundaries and safeguards helps balance innovation and risk, reflecting the proportionality principle in adaptive regulation.

---

[112] Maria Paola Priola, 'Addressing Hallucinations with RAG and NMISS in Italian Healthcare LLM Chatbots' (arXiv, 5 December 2024) <https://doi.org/10.48550/arXiv.2412.04235> accessed 30 June 2025.

[113] Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press 2018); Ryan Calo, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51 UC Davis Law Review 399; Corinne Cath and others, 'Artificial Intelligence and the "Good Society": the US, EU, and UK Approach' (2018) 24 (2) Science and Engineering Ethics 505.

[114] Matthew Crain, 'The Limits of Transparency: Data Brokers and Commodification' (2018) 20(1) New Media & Society 88..

Thirdly, in response to the loss of effective feedback mechanisms, various AI companies have implemented measures to enhance user feedback channels. These efforts aim to detect and suppress hallucinations promptly after they occur, thereby mitigating associated risks. In practice, developers like Anthropic have implemented structured feedback mechanisms to support this goal. For instance, users of the Claude model are encouraged to flag incorrect or misleading outputs, with such feedback integrated into ongoing model improvement processes.[115] In its Claude 3 release, Anthropic also emphasised improved factual accuracy and announced plans to integrate source attribution features in future updates.[116] While not yet fully automated, these iterative feedback loops represent a meaningful step toward post-deployment hallucination control, particularly within regulatory contexts like the EU. This early-stage development directly aligns with the AIA's emphasis on post-market monitoring. Legal AI company Doctrine, based in France, operationalised feedback capture via embedded "report error" buttons, enabling both model improvement and evidence gathering for regulatory compliance purposes. Furthermore, Hugging Face's COMPL-AI initiative provides an open-source audit framework that supports continuous hallucination tracking in hosted models, helping companies document output risks in anticipation of AIA enforcement.[117]

---

[115] Anthropic, 'Claude is providing incorrect or misleading responses — what's going on?' (*Anthropic*, 2024) <https://support.anthropic.com/en/articles/8525154-claude-is-providing-incorrect-or-misleading-responses-what-s-going-on> accessed 10 June 2025.

[116] Anthropic, 'The Claude 3 model family' (*Anthropic,* 4 March 2024) <https://www.anthropic.com/news/claude-3-family> accessed 10 June 2025.

[117] Carlo Giovine and others, 'Building AI trust: The key role of explainability' (*McKinsey*, 26 November 2024) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability> accessed 17 July 2025.

3.2.2.2.2 Mitigation of Harms on Macro Level

Firstly, in mitigating the harm of public information pollution and social cognitive disruption, the AIA mandates that users must be informed when they are interacting with AI-generated content. This transparency requirement is intended to alert users to the potential inaccuracy and epistemic limitations of such outputs, thereby reducing the likelihood that hallucinations will distort public understanding. In practice, several platforms and jurisdictions have already implemented precautionary labelling systems to comply with this principle.

For example, TikTok began automatically labelling AI-generated images and videos in May 2024, applying descriptors such as "AI-generated" or "Content Credentials" to content created using its internal tools. These labels are also extended to media produced outside TikTok where metadata includes C2PA (Coalition for Content Provenance and Authenticity) standards.[118] Similarly, Meta, the parent company of Facebook, Instagram and Threads, expanded its "Made with AI" labelling policy in April 2024 to cover visual and audio content, relying on both user disclosures and AI detection systems.[119] Although an internal oversight review found inconsistencies in the application of these labels, it reaffirmed the importance of visible AI attribution in protecting users from misinformation.[120] At the national level, Spain has proposed legislation that would

[118] The FactCheckHub, 'TikTok Begins Auto-Labelling of AI-Generated Content' (*The FactCheckHub*, 15 May 2024) <https://factcheckhub.com/tiktok-begins-auto-labelling-of-ai-generated-content> accessed 10 June 2025.

[119] Mia Sato, 'Meta Will Require AI-Generated Content Labels on Facebook, Instagram and Threads' The Verge (*The Verge*,5 April 2024) <https://www.theverge.com/2024/4/5/24121978/meta-ai-generated-content-label-requirements-deepfakes> accessed 10 June 2025.

[120] Cynthia Kroet, 'Meta's AI Labelling "Inconsistent", Internal Oversight Board Finds' Euronews Next (*euro news*, 25 June 2025)

impose fines of up to €35 million or 7% of global turnover on companies that fail to label AI-generated content, especially deepfakes.[121] This initiative echoes the labelling provisions in the AIA and illustrates how national regulators are operationalising EU-level rules in concrete enforcement strategies. In parallel, Germany's public broadcaster ZDF, with support from the EU-funded MediaFaktCheck project, introduced hallucinations detection systems for AI-generated news summaries, with editorial review layers added prior to publication.[122] These emerging practices demonstrate growing recognition that the visibility of AI provenance through accurate, standardised labelling, can serve as a first line of defence against hallucination-driven public confusion.

Secondly, in response to the harm of erosion of public trust and acceptance toward the AI industry, many companies have adopted a strategy of transparent disclosure to acknowledge the existence of hallucination problems and demonstrate a proactive commitment to rectification. By openly reporting system limitations and planned mitigation measures, these companies aim to reduce public concern and rebuild confidence in the responsible development of AI technologies.

Independent audits and model documentation practices have begun to play a crucial role in enhancing transparency and trust around hallucination risks in

---

<https://www.euronews.com/next/2025/06/25/metas-ai-labelling-inconsistent-internal-oversight-board-finds> accessed 30 June 2025.

[121] Reuters, 'Spain to Impose Massive Fines for Not Labelling AI-Generated Content' Reuters (Reuters, 11 March 2025) <https://www.reuters.com/technology/artificial-intelligence/spain-impose-massive-fines-not-labelling-ai-generated-content-2025-03-11> accessed 10 June 2025.

[122] Laura Mascarell, Ribin Chalumattu and Annette Rios, 'German Also Hallucinates! Inconsistency Detection in News Summaries with the Absinth Dataset' in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (ELRA and ICCL 2024) 7696.

deployed LLM systems. For instance, Hugging Face's model cards for BLOOM, Falcon, and other hosted models include detailed disclosures concerning hallucination-related risks, dataset provenance, evaluation methods, and proposed mitigation strategies.[123] These documentation practices have been praised for promoting responsible AI development and enabling downstream users, including researchers, regulators, and developers, to better understand the behavioural characteristics of foundation models.[124] Meanwhile, proposals for formal hallucination audit protocols have emerged in recent research. These include frameworks for measuring a model's "Hallucination Vulnerability Index" (HVI), enabling independent third parties to report and benchmark LLM performance across tasks and domains.[125] Such initiatives represent promising steps toward institutionalising transparency, safety, and accountability within the LLMs lifecycle.

Thirdly, in addressing the harm of regulatory vacuums and governance deficits, both public authorities and AI developers have taken steps to improve alignment between regulators and the regulated. On the one hand, national authorities are increasingly engaging in in-depth investigations and research to better understand the operational needs and risks of LLMs. For instance, in April 2025, Ireland's Data Protection Commission launched an investigation into xAI's Grok model over alleged scraping of EU user data for training purposes, signalling that enforcement

---

[123] Shiyu Jin and others, 'Reasoning Grasping via Multimodal Large Language Model' in *Proceedings of the 8th Conference on Robot Learning* (PMLR 2025) 3809.

[124] Margaret Mitchell and others, 'Model Cards for Model Reporting' *in Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM 2019) 220.

[125] Vipula Rawte and others, 'The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations' in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023) 2541.

is now extending into complex data provenance issues at the intersection of the AIA and the GDPR.[126]

On the other hand, LLMs developers have begun to internalise legal expectations by translating them into technical benchmarks. In late 2024, the open-source benchmarking initiative COMPL-AI introduced a "hallucination resilience score" for GPAI systems,[127] designed to assess model robustness against hallucination risks. This metric reflects risk-based thresholds inferred from the AIA and offers a practical tool for aligning model performance with evolving regulatory standards. Together, these developments suggest a gradual closing of the governance gap through mutual adaptation between legal frameworks and technical implementations.

3.3 AI Regulation in China

3.3.1 Theoretical Framework: Regulatory Approach and Legal Provisions

3.3.1.1 Regulatory Approach in China

[126] Edith Hancock, 'Ireland's Privacy Watchdog Probes Musk's Grok AI Model' (*The Wall Street Journal*, 11 April 2025) <https://www.wsj.com/tech/irelands-privacy-watchdog-probes-musks-grok-ai-model-4779ba4e?gaa_at=eafs&gaa_n=AWEtsqeYTAWHI9WPcmQAd7tG8BfBWOiIzDVoz0N8ncp9-_KSFM3x_X4sUxScWqPosGc%3D&gaa_ts=69c5c097&gaa_sig=KOB-sEp87csYcoHjM0Xv9oUYlyHZixaUtXXHCuiuYz9HGSPAOIfZj7LOu9hvKN6LhXFyjVMRvDrr-W6SeW94bw%3D%3D> accessed 14 July 2025.

[127] Philipp Guldimann and others, 'COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act' (arXiv, 10 October 2024) <https://doi.org/10.48550/arXiv.2410.07959> accessed 10 October 2025.

The formulation and enforcement of law are rooted in their broader social context, particularly in prevailing cultural traditions. In China, regulatory practice is embedded in domestic social structures and has often reflected a more paternalistic governance style, whereas the EU regulatory model is more closely aligned with liberal constitutional values.

Firstly, China is often characterised by a government-led national system, with the central government playing a dominant role in setting strategic directions and policy objectives. While retaining authority, the central government also enables multiple departments, local governments, and policy entrepreneurs to actively contribute to policy formulation and implementation. This approach is underpinned by experimentalist governance, which allows for localised adaptation and iterative refinement within the framework of central guidelines.[128] In the context of AI governance, China has explicitly adopted an innovation-first approach, prioritising technological development as a key driver of national competitiveness, along with the goals to enhance national security and maintain social order. More recently, ethical considerations have been integrated into China's AI governance framework, functioning primarily as utilitarian tools to support and safeguard technological advancement, rather than as constraints on development.[129]

---

[128] Andrew Mertha, '"Fragmented Authoritarianism 2.0": Political Pluralization in the Chinese Policy Process' (2009) 200 The China Quarterly 995; Xufeng Zhu and Hui Zhao, 'Experimentalist Governance with Interactive Central–Local Relations: Making New Pension Policies in China' (2021) 49(1) Policy Studies Journal 13.

[129] Huw Roberts and others, 'Achieving a "Good AI Society": Comparing the Aims and Progress of the EU and the US' (2021) 27(6) Science and Engineering Ethics; Huw Roberts and others, 'Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes' (2023) 39(2) The Information Society 79; Viktor Tuzov and Fen Lin, 'Two Paths of Balancing Technology and Ethics: A Comparative Study on AI Governance in China and Germany' (2024) 48(10) Telecommunications Policy 102850.

Secondly, China has adopted a proactive and holistic approach to AI regulation, demonstrating strong regulatory attention to the technology. It has introduced supplementary and precautionary measures in specific domains and responded swiftly to emerging phenomena such as deep synthesis and algorithmic recommendation, reflecting a forward-looking regulatory posture. China is the first major jurisdiction to introduce a dedicated regulatory framework for Gen AI through the issuance of the Interim Measures. The New Generation Artificial Intelligence Development Plan further underscores the country's strategic commitment to AI, setting forth a national roadmap aimed at positioning China as a global leader in AI by 2030.[130] Beyond their role in information technology standardisation, national standardisation technical committees such as TC28/SC42 also play a critical role in shaping AI governance by formulating industry guidelines.[131] Another important contribution to China's regulatory ecosystem is the Ethical Norms for New Generation Artificial Intelligence[132] emphasising the

---

[130] The General Office of the State Council of the People's Republic of China, 'The Inaugural Meeting of the Artificial Intelligence Subcommittee of the National Information Security Standardization Technical Committee Held in Beijing' (全国信息安全标准化技术委员会人工智能分技术委员会成立大会在京召开) (*The Central People's Government of the People's Republic of China*, 20 July 2017) <https://www.gov.cn/xinwen/2017-07/20/content_5212064.htm> accessed 27 July 2025.

[131] National Public Service Platform for Standards Information, 'TC28/SC42 National Information Technology Standardization Technical Committee Artificial Intelligence Subcommittee' ('TC28/SC42 全国信息技术标准化技术委员会人工智能分技术委员会') (*National Public Service Platform for Standards Information, 2026*) <https://std.samr.gov.cn/search/orgDetailView?tcCode=TC28SC42> accessed 28 March 2026.

[132] National New Generation Artificial Intelligence Governance Professional Committee, 'Ethical Norms for New Generation Artificial Intelligence' (新一代人工智能伦理规范)(*Ministry of Science and Technology of the People's Republic of China*, 26 September 2021) <https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html> accessed 28 March 2026.

principles of trustworthiness and controllability, advocating for AI that promotes honesty, fairness, human well-being, and the protection of privacy and security. It also affirms the right of individuals to choose whether to engage with AI systems, allowing users to decline interaction at any point. In 2023, China further demonstrated its commitment to shaping global AI governance by introducing the Global AI Governance Initiative at the Belt and Road Summit, outlining eleven key proposals as part of China's approach to international AI regulation.[133]

Moreover, the regulation of hallucinations can draw upon existing approaches to risk-based governance. Although hallucinations are not always explicitly addressed in current AI legislation, the overarching principles of risk classification, proportionality, and accountability embedded in these frameworks provide a useful foundation for their regulation.[134] In this sense, existing AI laws offer indirect but meaningful guidance for mitigating hallucination risks across the LLMs lifecycle. The Interim Measures adopt a tiered and classified regulatory model based on the level of risk. Since hallucination risks can, to some extent, be evaluated within such a risk classification framework, this model provides a useful reference. Besides, as AI applications in China vary across foundational models, specialised models and vertically integrated applications, regulatory measures are correspondingly designed for each category to ensure alignment between risk levels and accountability mechanisms.[135] Given that foundational and specialised

[133] Central Cyberspace Affairs Commission, 'Initiative for Global AI Governance' (全球人工智能倡议) (*Cyberspace Administration of China*, 18 October 2023) <https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm> accessed 7 February 2024.

[134] Luciano Floridi and others, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28(4) Minds and Machines 689; Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22(4) Computer Law Review International 97.

[135] Zhihong Gao, 'Response and Transcendence: Legal Regulation of Generative Artificial Intelligence' (回应与超越：生成式人工智能法律规制—

models exhibit different patterns and degrees of hallucination-related harm, regulatory responses should reflect these distinctions. By integrating industry-specific rules, the governance framework can apply differentiated regulatory requirements to various types of hallucinations, thereby enhancing regulatory precision and effectiveness.

In addition, China adopts a targeted approach to AI regulation rather than a single, comprehensive legislative framework. Although this model is rich in regulatory content and responsive in administrative terms, it remains fragmented and relies chiefly on administrative regulations, policy documents, and sector-specific standards. These instruments provide flexibility, but often lack binding force and normative coherence, thereby limiting the emergence of a unified and legally robust system of AI governance. China's broader regulatory architecture is accordingly multi-layered, combining primary legislation, secondary legislation, policy documents, and industry standards. Yet no comprehensive AI law has been enacted by the National People's Congress or its Standing Committee. The governance of LLMs therefore remains indirect, mediated through related legal fields such as cybersecurity, data security, and personal information protection including the Cybersecurity Law[136], the Data Security Law[137], and the Personal Information Protection Law, thus offer only a partial and indirect foundation for regulating LLMs. At the level of local legislation, Chinese municipalities,

—以《生成式人工智能服务管理暂行办法》为视角) (2024) 5 Social Sciences Journal 121.

[136] Cybersecurity Law of the People's Republic of China (中华人民共和国网络安全法) (adopted 7 November 2016, effective 1 June 2017; amended 28 October 2025, effective 1 January 2026)

[137] Data Security Law of the People's Republic of China (中华人民共和国数据安全法) (adopted 10 June 2021, effective 1 September 2021).

particularly major high-tech cities such as Shenzhen[138] and Shanghai[139], have adopted measures that provide region-specific guidance on promoting AI innovation while addressing its associated risks. Policy documents play a complementary role, such as the "New Generation Artificial Intelligence Development Plan"[140], which outlines strategic goals for AI innovation and application. And the "Governance Principles for the New Generation Artificial Intelligence" which outlines strategic goals for AI regulation.[141] Industry standards further strengthen this framework by integrating technical and regulatory considerations. Examples include GB/T 42018-2022,[142] which outlines technical requirements for physical and virtual computing resources on AI platforms for machine learning, as well as testing methods for physical resources, and GB/T 41867-2022,[143] which defines common terms and terminology in the field of AI and information technology.

[138] Shenzhen Special Economic Zone Regulations on Promoting the Artificial Intelligence Industry (深圳经济特区人工智能产业促进条例) (Shenzhen Municipal People's Congress Standing Committee, 30 August 2022).

[139] Regulations of Shanghai Municipality on Promoting the Development of the Artificial Intelligence Sector (上海市促进人工智能产业发展条例) (Shanghai Municipal People's Congress Standing Committee, 22 September 2022).

[140] State Council of the People's Republic of China, New Generation Artificial Intelligence Development Plan (新一代人工智能发展规划) (20 July 2017).

[141] National New Generation Artificial Intelligence Governance Professional Committee, 'Governance Principles for the New Generation Artificial Intelligence: Developing Responsible Artificial Intelligence' (新一代人工智能治理原则——发展负责任的人工智能, 17 June 2019).

[142] State Administration for Market Regulation and Standardization Administration of the People's Republic of China, 'Information technology—Artificial intelligence—Platform computing resource specification'(信息技术 人工智能 平台计算资源规范, GB/T 42018-2022, 14 October 2022).

[143] State Administration for Market Regulation and Standardization Administration of the People's Republic of China, 'Information technology—

3.3.1.2 Legal Provisions Governing LLMs in China

There are several Chinese laws of general application relevant to AI. This thesis, however, focuses on instruments enacted specifically for AI governance, notably the Provisions on the Administration of Algorithmic Recommendations for Internet Information Services (Algorithmic Recommendations Provisions)[144], the Provisions on the Administration of Deep Synthesis in Internet Information Services (Deep Synthesis Provisions)[145], and, most importantly for LLMs, Interim Measures. The Interim Measures constitute the principal dedicated framework for public-facing Gen AI services in China, as they apply specifically to Gen AI services provided to the public within China. Article 4 sets out the general principles governing both the provision and use of such services and, by referring to algorithm design, training-data selection, model generation and optimisation, and service provision, extends across much of the LLM lifecycle.

For analytical purposes, the Chinese AI-specific framework may be understood as involving three principal categories of actors: technical supporters, service providers, and users. This structure is stated most clearly in the Deep Synthesis Provisions, which expressly define all three roles, while the Interim Measures define providers and users of Gen AI services. Their obligations differ in emphasis.

Artificial intelligence—Terminology' (信息技术 人工智能 术语, GB/T 41867-2022, 14 October 2022).

[144] Provisions on the Administration of Algorithmic Recommendations for Internet Information Services (互联网信息服务算法推荐管理规定) (Cyberspace Administration of China and others, 31 December 2021, effective 1 March 2022).

[145] Cyberspace Administration of China and others, Provisions on the Administration of Deep Synthesis in Internet Information Services (互联网信息服务深度合成管理规定, 25 November 2022, effective 10 January 2023).

Technical supporters are subject mainly to training-data, technical-management, filing and, in some cases, security-assessment obligations; service providers bear the core duties of filing, content governance, labelling, complaint handling and cooperation with supervision; and users are prohibited from using AI services to generate or disseminate unlawful content.

Specifically, for different stages of the lifecycle：

Firstly, in the model design stage, transparency and explainability are crucial for minimising the spread of hallucinations. While Article 11 requires providers to process individuals' requests to access, correct, or delete personal information, it only partially supports transparency by focusing on personal data rights rather than broader requirements critical for mitigating hallucinations. Furthermore, the Interim Measures lack explicit provisions to directly enhance transparency, leaving gaps in ensuring explainability and accountability.

Secondly, in the pretraining stage, training data quality and legality greatly influence LLMs output and hallucination risks. Therefore, Training data must come from legitimate sources, respect intellectual property, and process personal data legally. Article 6 supports Gen AI infrastructure, resource sharing, and high-quality public training data platforms. Article 7(4) emphasises improving data accuracy, relevance, objectivity, and diversity, while Article 8 mandates clear annotation guidelines, quality checks, accuracy verification, and personnel training to ensure data reliability.

Thirdly, in the fine-tuning and alignment stage, testing and evaluation are essential to assess performance and mitigate hallucinations. Article 8 of the Interim Measures sets rules for data annotation during research and development. This relates closely to the fine-tuning and alignment stage because the quality and

consistency of annotated data directly affect model adjustments and alignment results.

Fourthly, in the deployment and interaction stage, Article 13 requires corrective actions like adjusting model parameters or removing problematic content when hallucinations are found. Article 14 requires content evaluation and user feedback to improve models. This is designed to find the sources of hallucinations and lowers how often they happen.

Finally, in the monitoring and iteration stage, Article 9 says providers are seen as online content producers and must deal with illegal content by suspending, removing, fixing, or reporting it. Article 15 makes providers legally responsible for the content they generate. This encourages better quality control to stop hallucinations from causing harm or legal problems. Article 12 requires that AI-generated content is labelled. This lets users know and check such content. Even if hallucinations happen, labelling helps reduce the chance of misunderstanding.

In summary, China's Interim Measures broadly cover the five stages of the LLM lifecycle, but most of the provisions focus on models that have already been deployed. For example, Article 10 explains the appropriate user groups, scenarios, and purposes for model services, while Article 11 imposes obligations on providers to protect users' input data and usage records. These provisions tend to treat LLMs as products that are already capable of being deployed, rather than addressing the models in their training or tuning phases. As a result, the regulatory focus more on the later stages of the lifecycle. Chapters 3 and 4 primarily correspond to the monitoring and iteration phases, whereas the earlier stages are comparatively underregulated.

3.3.2 Practical Assessment: Effectiveness of China's Regulation

3.3.2.1 Implementation of China's AI Regulations

China has adopted a government-led, top-down approach to AI governance, in which enforcement plays a central role in shaping industry compliance. At the same time, the regulation of AI is characterised by a multi-agency framework involving bodies such as the State Administration for Market Regulation, the Cyberspace Administration of China (CAC), the Ministry of Industry and Information Technology, and the Ministry of Science and Technology. This arrangement reflects the cross-sectoral nature of AI-related risks, which require regulatory responses grounded in the expertise of different authorities. However, this multi-agency setup allows for joint oversight but can also cause overlapping duties[146], which might make the rules less clear in practice.

For enforcement, Chinese authorities mainly use administrative checks and rules that need approval before use. They do not depend much on courts. One important tool is the filing system. Providers of Gen AI services need to send their models for review before they make them public. This early review helps make sure models meet basic standards for data quality, control of outputs, and content safety, which are all connected to risks of hallucinations.

Another central feature of China's regulatory implementation is its reliance on technical audits and algorithmic assessments. The Interim Measures providers to establish internal auditing and content moderation systems. These systems are often composed of human-in-the-loop feedback loops, real-time output filtering,

[146] Wenxuan Bi, 'The Dilemma in the Risk Regulation of Generative Artificial Intelligence and Its Resolution: Taking ChatGPT as an Example' (生成式人工智能的风险规制困境及其化解：以 ChatGPT 的规制为视角) (2023) 3 Journal of Comparative Law 155.

and risk labelling of potentially misleading content. In practice, large platforms such as Baidu and Alibaba have developed automated screening tools and watermarking mechanisms to trace AI-generated content, with some models equipped to flag outputs with low confidence levels or high hallucination potential.[147]

Nevertheless, implementation challenges remain. One notable limitation is the lack of transparency in enforcement metrics and outcomes. Unlike the EU's growing emphasis on publishable compliance reports and independent oversight bodies, China's regulatory process remains opaque, with limited public access to enforcement data or audit outcomes. This lack of clarity makes it hard to know if the risks from hallucinations are truly being handled or just managed on the surface through rules.

3.3.2.2 Mitigation of Hallucination Harms

3.3.2.2.1 Mitigation of Harms on Micro Level

[147] Bytefeed, 'China Ramps Up Plans for Mandatory AI Watermarks Amid Rising Concerns' (*Bytefeed*, 27 September 2024) <https://bytefeed.ai/technology/china-ramps-up-plans-for-mandatory-ai-watermarks-amid-rising-concerns/> accessed 1 July 2025; Alibaba Cloud, 'AIGC and Forgery Detection Service of Image Moderation 2.0' (*Alibaba Cloud*, 24 Novermber 2025) <https://www.alibabacloud.com/help/en/content-moderation/latest/image-audit-enhanced-edition-detects-aigc-infringement> accessed 1December 2025; AIbase, 'Taobao Launches Platform-Wide AI-Generated Fake Image Governance' (*AI Base*, 27 March 2025) <https://www.aibase.com/news/16663> accessed 1 July 2025.

Firstly, for the cognitive misguidance and the acceptance of misinformation, Chinese models such as DeepSeek[148] and Qwen (by Alibaba)[149] have been publicly criticised for hallucinating historical or scientific facts. In response, research institutions proposed the HaluSearch framework, which integrates System-2 slow reasoning via tree search to reduce inference-stage hallucinations. [150] Experiments on Chinese-language knowledge bases have shown a significant drop in factual errors (over 25%) due to this slow reasoning mechanism.[151] In the legal domain, a team at Central China Normal University implemented an adapt-retrieve-revise pipeline: a smaller LLM drafts an answer, retrieves domain-specific evidence, and then revises it reducing hallucination rates by approximately 33% in zero-shot Chinese legal QA.[152] This mirrors international practices like retrieval-augmented generation used in the EU but tailored to Chinese information systems.

---

[148] Deepseek-ai, '[Hallucination Report] Model hallucinates biological validity in psychiatric analogies' (*Hugging Face discussion*, 11 January 2026) <https://huggingface.co/deepseek-ai/DeepSeek-R1/discussions/236> accessed 28 March 2026.

[149] Qwen, 'Qwen is loosing broad knowledge since Qwen2' (*Hugging Face discussion*, 29 April 2025) <https://huggingface.co/Qwen/Qwen3-235B-A22B/discussions/16> accessed 28 March 2026.

[150] Xiaoxue Cheng and others, 'Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking' in *Findings of the Association for Computational Linguistics: ACL 2025* (Association for Computational Linguistics 2025) 7979.

[151] Jiawei Chen and others, 'Benchmarking Large Language Models in Retrieval-Augmented Generation' (2024) 38(16) Proceedings of the AAAI Conference on Artificial Intelligence 17754.

[152] Zhen Wan and others, 'Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain' in *Findings of the Association for Computational Linguistics: ACL 2024* (Association for Computational Linguistics 2024) 5030.

Secondly, with regard to overreliance and declining trust, a Nature-indexed study[153] in the healthcare domain incorporated a dual-retrieval mechanism into a Chinese medical LLM. Through interaction with external knowledge bases, including medical ontologies, the model achieved a reported 35% reduction in clinically incorrect hallucinations. Following regulatory expectations set by the CAC, some providers now display output confidence levels and link directly to cited medical sources, which preliminary surveys report has restored user trust.

Thirdly, with regard to the loss of effective feedback mechanisms, Chinese researchers[154] have proposed several technical interventions to restore corrective feedback within LLM systems. In the context of Alipay Search, one study develops an optimised generative retrieval framework to mitigate retrieval hallucinations by integrating knowledge-distillation reasoning into model training and incorporating a decision agent to further improve retrieval precision. More broadly, Chinese scholars[155] have also proposed Dynamic Retrieval Augmentation based on Hallucination Detection (DRAD) as a novel method for detecting and mitigating hallucinations in LLMs.

3.3.2.2.2 Mitigation of Harms on Macro Level

---

[153] Qimin Yang and others, 'Dual Retrieving and Ranking Medical Large Language Model with Retrieval Augmented Generation' (2025) 15(1) Scientific Reports 18062.

[154] Yedan Shen and others, 'Alleviating LLM-Based Generative Retrieval Hallucination in Alipay Search' in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM 2025) 4294.

[155] Weihang Su and others, 'Mitigating Entity-Level Hallucination in Large Language Models' in *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (ACM 2024) 23.

Firstly, considering to the public information pollution and social cognitive disruption, the Interim Measures mandate that all public-facing AI providers implement source verification, content flagging, and user-complaint channels. Consequently, major LLM providers such as Baidu Ernie and iFLYTEK Spark now include built-in hallucination filters or adopt other mitigation strategies to reduce hallucinations, along with user-facing disclaimers in their public deployments.[156]

Secondly, in the aspect of erosion of public trust and acceptance toward the AI industry, according to a survey, 72% of Chinese respondents express trust in AI systems, which suggests a generally favourable public attitude toward AI in China.[157]

Thirdly, with respect to regulatory vacuums and governance deficits, although China's AI regulatory framework is proactive, significant challenges remain in assessing non-material harms. At the same time, China has continued to promote technological innovation and benchmarking, enabling domestic LLMs to narrow the gap with leading US models, and in some cases to outperform them on bilingual

---

[156] Baidu, Inc., 'Baidu Launches ERNIE 4.5 Turbo, ERNIE X1 Turbo and New Suite of AI Tools to Empower Developers and Supercharge AI Innovation' (*PR Newswire*, 25 April 2025) <https://www.prnewswire.com/news-releases/baidu-launches-ernie-4-5-turbo-ernie-x1-turbo-and-new-suite-of-ai-tools-to-empower-developers-and-supercharge-ai-innovation-302438584.html> accessed 14 July 2025; Ben jiang, 'Chinese online search giant Baidu's Ernie Bot joins iFlytek's Spark in Apple's mainland App Store for local AI' South China Morning Post (*South China Morning Post*, 5 July 2023) <https://www.scmp.com/tech/big-tech/article/3226659/chinese-online-search-giant-baidus-ernie-bot-joins-iflyteks-spark-apples-mainland-app-store-local-ai> accessed 14 July 2025.

[157] Ina Fried, "Exclusive: Trust in AI is much higher in China than in the U.S." (*Axios*, 13 February 2025) <https://www.axios.com/2025/02/13/trust-ai-china-us> accessed 10 October 2025.

benchmarks.[158] Through planning and policy measures, China also seeks to enhance LLM capabilities within lawful limits and to reduce the occurrence of hallucinations.[159] In parallel, China has supported sector-specific AI innovation through policies facilitating access to strategic datasets, particularly biodata.[160] This data-governance model may strengthen China's innovation capacity, but it also raises unresolved questions about data control, institutional accountability, and the legal assessment of downstream harms.

3.4 Conclusion

This chapter examined how the EU and China regulate hallucination risks in LLMs through both normative design and practical implementation. Drawing on the lifecycle-based framework established in Chapter 2, it showed that both jurisdictions seek to address hallucinations through AI-related regulation, yet they do so on the basis of markedly different regulatory logics. The EU treats hallucination risks as part of a broader system of legally structured risk governance, attempting to manage them through ex ante obligations concerning transparency, data governance, human oversight, and accountability across the lifecycle of LLMs. China, by contrast, approaches hallucination risks through a more interventionist and administratively driven model, combining high-level policy direction with

---

[158] Hodan Omaar, Information Technology and Innovation Foundation (*ITIF*, 26 August 2024) <https://itif.org/publications/2024/08/ 26/how-innovative-is-china-in-ai/> Accessed 23 September 2024.

[159] Daniel Castro, 'China's Annual Parliamentary Meeting Shows National Commitment to Advancing AI' (*Center for Data Innovation*, 18 March 2024) <https://datainnovation.org/2024/03/chinas-annual-parliamentary-meeting-shows-national-commitment-to-advancing-ai/> Accessed 23 September 2024.

[160] Caroline Schuerger, Vikram Venkatram and Katherine Quinn, 'China and Medical AI: Implications of Big Biodata for the Bioeconomy' (*Center for Security and Emerging Technology*, May 2024) <https://cset.georgetown.edu/publication/china-and-medical-ai/> accessed 26 July 2025.

sector-specific rules and ex post enforcement measures, particularly at the stages of deployment, dissemination, and content control.

The comparison also demonstrates that the divergence between the two regimes is not merely institutional, but conceptual. The EU's framework reflects a commitment to systemic legal ordering, in which hallucinations are addressed indirectly through generalised risk-management obligations embedded in a codified legislative structure. China's framework is more pragmatic and operational, prioritising governability, content security, and immediate regulatory responsiveness over doctrinal coherence. Each model therefore exhibits a distinct regulatory strength: the EU offers greater normative structure and lifecycle coverage, while China provides faster administrative implementation and more direct influence over provider behaviour.

At the same time, the chapter revealed important limitations in both jurisdictions. In the EU, the effectiveness of the AIA remains contingent on the development of technical standards, institutional capacity, and workable enforcement arrangements, such that legal ambition currently exceeds operational readiness. In China, although enforcement is comparatively swift and visible, the framework remains fragmented, heavily reliant on administrative instruments, and less attentive to the upstream technical causes of hallucinations. As a result, both systems continue to regulate hallucinations only indirectly, as part of broader AI governance, rather than as a distinct category of harm requiring targeted legal treatment.

A deeper implication of this analysis is that hallucinations expose a structural tension at the heart of contemporary AI governance. They are simultaneously technical failures, informational harms, and governance problems, yet existing regulatory frameworks tend to address only one or two of these dimensions at a time. This helps explain why gaps persist in both jurisdictions, particularly in

relation to harm evaluation, liability allocation, real-time mitigation, and the integration of technical detection methods into legal oversight. The chapter therefore suggests that effective regulation of hallucination risks cannot rest solely on general AI rules, nor solely on reactive administrative enforcement. It requires a more integrated governance approach capable of linking lifecycle-based technical intervention with legally coherent, enforceable, and context-sensitive regulatory design.

The next chapter builds on these findings by undertaking a more explicit comparative analysis of the EU and China, identifying their shared challenges, fundamental divergences, and the lessons each may offer for the development of more effective responses to hallucination risks in LLMs.

Chapter 4 - Comparison and Analysis of Reasons: The Similarities and Differences between the Regulation of Hallucination in EU and China

4.1 Introduction

This chapter undertakes a comparative and explanatory analysis of the regulatory approaches adopted by the EU and China in addressing hallucination-related harms arising from LLMs. The objective of the chapter is not merely to identify similarities and differences between the two jurisdictions, but to explore the underlying reasons that account for both regulatory convergence and divergence. National approaches to AI governance are shaped by a combination of legal traditions, political institutions, economic priorities, and historically embedded social values.

This chapter emphasis on the role of culturally informed risk perceptions and governance preferences in shaping regulatory outcomes by incorporating insights from Cultural Dimensions Theory of Hofstede[161], with particular reference to the framework developed by Geert Hofstede, as an analytical lens that complements doctrinal and institutional comparison. Hofstede's theory conceptualises culture as a set of shared value orientations that influence how societies perceive uncertainty, allocate responsibility, and legitimise authority. Certain cultural dimensions are especially relevant to the governance of epistemic risks generated by LLMs. These include uncertainty avoidance, which relates to societal tolerance for ambiguity and unpredictability; individualism and collectivism, which reflect the relative prioritisation of individual autonomy and collective interests; power distance, which concerns the acceptance of hierarchical authority and centralised decision-making; and long-term orientation, which captures the extent to which

[161] Geert Hofstede, 'Dimensionalizing Cultures: The Hofstede Model in Context' (2011) 2(1) *Online Readings in Psychology and Culture*.

future stability and sustained development are prioritised over short-term flexibility. In this chapter, these dimensions are not treated as deterministic explanations of legal outcomes. Rather, they are employed as contextual analytical tools that help illuminate why similar technological risks may be framed, prioritised, and regulated differently across legal systems.

From a structural perspective, both the EU and China have increasingly acknowledged the epistemic risks associated with the outputs of LLMs, including misinformation, fabricated content, and ungrounded responses. Although neither jurisdiction explicitly employs the technical term “hallucination” in its primary regulatory instruments, both have begun to recognise such outputs as legally and socially significant harms. In both contexts, hallucination-related risks are understood to operate on two interconnected levels. At the individual level, misleading outputs may distort decision-making in sensitive domains such as healthcare, legal advice, or education. At the societal level, large-scale dissemination of inaccurate or fabricated content may undermine the integrity of the public information environment and erode collective trust.

This shared recognition of risk has led to a degree of regulatory convergence. Both jurisdictions rely on regulatory mechanisms such as transparency obligations, accuracy-related commitments, requirements for human oversight, and ongoing monitoring or review of AI systems. However, beneath these apparent similarities lie substantial differences in regulatory rationales, institutional arrangements, and modes of enforcement. The EU’s approach is largely grounded in a rights-based and procedural regulatory tradition, while China’s regulatory framework reflects a more centralised and governance-oriented approach to information control. These differences cannot be fully explained by variations in technological development or market structure alone. Instead, they are closely linked to distinct legal cultures and governance philosophies that shape how epistemic risks are perceived and managed.

This chapter is structured as follows. Section 4.2 examines the areas of regulatory similarity between the EU and China in addressing hallucination-related harms. It identifies issues that have already been addressed in both jurisdictions and evaluates the extent to which existing regulatory mechanisms mitigate epistemic risks effectively or leave certain harms insufficiently regulated. This section also explores the legal, institutional, and cultural foundations that contribute to these shared regulatory patterns. Section 4.3 turns to the key differences between the two jurisdictions. It distinguishes between divergences at the theoretical level, including legislative approaches, regulatory models, and lifecycle coverage, and differences at the practical level, such as enforcement institutions, implementation effectiveness, and extraterritorial reach. Through this combined doctrinal and contextual analysis, the chapter provides a nuanced account of how hallucination risks are governed in practice in each system.

By situating regulatory similarities and differences within their broader cultural and institutional contexts, this chapter lays the analytical foundation for the subsequent chapter's examination of regulatory lessons. Understanding the reasons behind regulatory choices is essential to assessing the possibilities and limits of regulatory learning, adaptation, and transplantation across jurisdictions.

4.2 Similarities Between the EU and China

4.2.1 Issues Already Addressed in Both Jurisdictions

4.2.1.1 Regulating Hallucinations Through AI Regulatory Commitments

In both the EU and China, academic and policy discourse increasingly treats LLM hallucinations as a governance concern, especially regarding transparency, accuracy, and accountability, yet neither jurisdiction has enacted explicit rules addressing hallucinations.[162] Although neither jurisdiction has adopted explicit legal rules that directly target hallucinations as a discrete regulatory category, there is a shared understanding that the mitigation of such epistemic risks cannot rely solely on technical solutions. Instead, both EU and China acknowledge the necessity of embedding legal and normative commitments into the governance of Gen AI across the entire lifecycle of LLMs, including model design, data selection and training, deployment, user interaction, and post-deployment monitoring.

This convergence can be partly explained through Hofstede's cultural dimension of uncertainty avoidance. In both jurisdictions, hallucinations are perceived as a form of epistemic unpredictability that undermines trust in automated systems and distorts decision-making. In the EU, relatively high uncertainty avoidance has historically supported precautionary regulatory approaches emphasising transparency and accountability. In China, intolerance of uncertainty is more closely associated with concerns over information reliability and social order. Despite these different normative orientations, both systems converge on the view that uncontrolled hallucinations represent an unacceptable source of uncertainty that warrants regulatory intervention.

A further similarity lies in the strategic use of high-level AI policy instruments to articulate regulatory commitments. Both jurisdictions have adopted comprehensive AI strategies that, while not legally binding, play an important guiding role in shaping subsequent regulatory development. This approach reflects a shared long-term orientation in AI governance, in which early normative

[162] Qinyuan Cheng and others, 'Evaluating Hallucinations in Chinese Large Language Models' (arXiv, 25 October 2023) <https://doi.org/10.48550/arXiv.2310.03368> accessed 6 August 2025.

guidance is intended to influence regulatory trajectories over time rather than to address isolated technical failures. As Chapter 3 has shown, notwithstanding differences in legal context, these instruments reveal a significant degree of normative convergence, with ethics, safety, data protection, and accountability emerging as core pillars of AI governance.

Notably, both the EU and China have chosen to address hallucination-related risks indirectly by embedding them within broader legal frameworks, particularly those concerning personal data protection and online content regulation. This strategy reflects a pragmatic governance rationale that favours regulatory continuity and administrative efficiency. It also aligns with a shared preference for systemic and coordinated regulatory responses to emerging technological risks, even where the underlying cultural and institutional justifications differ.[163]

4.2.1.2 Grounding in Foundational Legal Frameworks

Policy makers in both the EU and China demonstrate notable convergence in their regulatory goals concerning ethical AI development, data protection, safety, and privacy. While their institutional approaches and political environments may differ significantly, both jurisdictions have grounded their AI governance efforts in pre-existing legal systems, particularly in relation to personal data protection and digital rights. This reflects a shared legal philosophy: that the governance of novel technologies should not be pursued through isolated instruments alone, but rather through integration with established legal doctrines that ensure continuity, legitimacy, and enforceability.

---

[163] Lilian Edwards, 'Regulating AI in Europe: Four Problems and Four Solutions' (*Ada Lovelace Institute*, 31 March 2022) <https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/> accessed 26 March 2026.

From the perspective of Cultural Dimensions Theory, this convergence can be partly explained by a shared preference for regulatory stability in the face of technological uncertainty. In societies characterised by relatively low tolerance for uncertainty, there is a strong inclination to manage novel risks through familiar legal structures rather than experimental or fragmented regimes. Embedding AI governance within existing legal frameworks allows regulators in both jurisdictions to reduce normative ambiguity and to frame hallucination-related harms as an extension of already recognised legal concerns, such as data accuracy, informational integrity, and accountability.

Regulatory clarity can enhance competition by levelling the playing field for innovative businesses that may lack resources to navigate complex legal landscapes.[164] However, concerns remain that overly rigid regulation may impede innovation by imposing disproportionate compliance costs, especially on developers and start-ups, which are often key drivers of technological change.[165]

In the EU, the GDPR has served as a foundational legal framework for AI governance. Many of the obligations proposed under the AIA such as transparency, data governance, accuracy, and human oversight are underpinned by GDPR principles, particularly the rights to data accuracy, explanation, and human intervention in automated decision-making processes.

---

[164] Reuben Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy' in *Proceedings of the 1st Conference on Fairness,* Accountability and Transparency (2018), 149.

[165] Richard Owen, Phil Macnaghten and Jack Stilgoe, 'Responsible Research and Innovation: From Science in Society to Science for Society, with Society' in Gary E Marchant and Wendell Wallach (eds), *Emerging Technologies: Ethics, Law and Governance* (Routledge 2020) 117.

Similarly, in China, the PIPL and the Cybersecurity Law provide the legal backbone for emerging AI regulation. The Interim Measures build on these pre-existing laws by incorporating principles of data minimisation, informed consent, security assessment, and risk-based supervision. These AI-specific rules are thus not autonomous instruments but are situated within an evolving legal system that reflects the broader administrative logic of China's digital governance model.

4.2.2 Issues Yet to Be Resolved in Both Jurisdictions

4.2.2.1 Balancing Minimisation and Reasonable Utilisation of Hallucinations

From the perspective of cultural background and regulatory logic, indulgence and restraint reflect different social attitudes towards the management of risk, autonomy, and social order. Indulgent cultures tend to allow greater space for freedom of expression and individual choice, whereas restrained cultures place greater emphasis on discipline, control, and conformity to social norms. In regulatory terms, this distinction may translate into a greater tolerance for experimentation and residual risk in indulgent settings, and a stronger preference for prevention and normative control in restrained ones. This difference is particularly relevant to the governance of hallucinations. Although both the EU and China currently adopt risk-based approaches that seek primarily to minimise or prevent hallucination-related harms as part of broader AI governance, such an approach may be incomplete. Hallucinations are not only a source of legal and social risk, but may, in some contexts, also contribute to creativity, exploratory reasoning, or innovation. The central regulatory question, therefore, is not simply how hallucinations can be eliminated, but how their harmful effects can be minimised while their potentially reasonable uses are recognised and appropriately governed. Precisely because hallucinations differ from many other

AI risks in both their causes and their effects, they should be treated as a distinct regulatory category requiring more tailored legal responses.

On one hand, the Cantor's diagonalisation argument proves that hallucinations themselves cannot be eliminated.[166] From a technical perspective, current methods are incapable of completely preventing hallucinations.[167] This is mainly because LLMs cannot learn all computable functions, and computable functions only cover a fraction of the real world. This is crucial to ensure their sustainable development and prevent their deployment in contexts that exceed their inherent capabilities. The complexity of the real world far exceeds the scope of their training data. As a result, hallucinations in real-world applications of LLMs are inevitable.

On the other hand, hallucinations are not inherently detrimental. As Professor Tiejun Huang observed, hallucination can be a manifestation of AI creativity.[168] If humans wish to develop AI agents that surpass them in creativity, they should not seek to eliminate hallucinations altogether; otherwise, AI will become no different from a mere search engine or information retrieval system. The preservation of AI's creative capacity is essential precisely because humans require forms of intelligence more imaginative than their own. According to Professor Huang, human capabilities are ultimately bounded—there is a cognitive

---

[166] Ziwei Xu, Sanjay Jain and Mohan Kankanhalli, 'Hallucination Is Inevitable: An Innate Limitation of Large Language Models' (arXiv, 22 January 2024) <http://arxiv.org/abs/2401.11817> accessed 17 October 2024.

[167] Manuel Cossio, 'A Comprehensive Taxonomy of Hallucinations in Large Language Models' (arXiv, 3 August 2025) <https://doi.org/10.48550/arXiv.2508.01781> accessed 21 October 2025.

[168] Xiaoting Ji, 'Prof. Tiejun Huang from the School of Computer Science at Peking University: We Should Not Simply "Kill" AI Hallucinations with One Stick' (北京大学计算机学院教授黄铁军：不能简单地将 AI 幻觉 "一棒子打死") *China Electronics News* (Beijing, 7 May 2024) 7.

ceiling to human intellect. Yet to survive and thrive in the vast universe, humanity must confront numerous questions that exceed the limits of human understanding and require a kind of imagination that humans alone may not possess. While LLMs can generate misinformation, they can also be leveraged to construct more realistic fake news propagation simulation frameworks, enabling in-depth research into the underlying patterns and trends of fake news dissemination. With the advancement of AI, LLMs, thanks to their powerful language comprehension and generation capabilities, offer new possibilities for studying the subtle dynamics of public opinion. LLMs can process and interpret complex natural language texts, capturing semantic, emotional, and logical relationships. This makes them effective tools for more accurately simulating the spread of fake news and shifts in public attitudes.

Moreover, hallucinations can be viewed as part of the continuum of human creative imagination, which can be particularly vulnerable to being influenced by traumatic experiences, leading to the amplification of negative emotional legacies. For example, ChatGPT shows astounding performance in understanding dialogue-related texts, and it tends to provide informal suggestions for medical tasks instead of definitive answers.[169]

4.2.2.2 Balancing Holistic Legislation and Specific Attention to Hallucination Harms

In both the EU and China, current regulatory frameworks do not treat hallucinations as a standalone legal category. Instead, hallucinations are addressed indirectly as one of several undesirable outcomes associated with high-

[169] Peter R Breggin, 'Understanding and Helping People with Hallucinations Based on the Theory of Negative Legacy Emotions' (2015) 43(1) The Humanistic Psychologist 70.

risk AI systems within broader legislative instruments. The potential harms arising from hallucinations are typically scattered across general provisions relating to data accuracy, user protection, transparency, misinformation, or algorithmic accountability. This reflects a legislative tendency to favour holistic and comprehensive regulation over fragmented, issue-specific rulemaking. Although general risk-based regulatory frameworks can encompass hallucination-related risks to some extent, they still lack targeted provisions specifically addressing hallucinations. Nonetheless, the existing legal instruments do cover key contexts in which hallucinations may result in harm particularly in relation to the spread of misinformation, the assurance of data quality, and the optimisation of post-deployment feedback mechanisms.

In addition, both the EU and China currently rely heavily on self-submitted technical and legal documentation as a basis for regulatory compliance. While this approach may contribute to overall AI risk reduction, it proves less effective in addressing the specific challenges posed by LLMs hallucinations. In practice, such reliance may increase the compliance burden on developers, reduce regulatory efficiency, and, to some extent, dampen innovation incentives. This raises concerns about the capacity of regulators to evaluate the technical accuracy and reliability of AI systems. Strengthening the technical expertise of enforcement authorities and integrating independent third-party verification may be necessary to move beyond formalistic compliance and achieve substantive oversight.

Moreover, neither the EU nor China has definitively classified LLMs as either products or technologies, a distinction that carries significant regulatory implications. If LLMs are seen as products, harms like hallucinations might be covered by product liability rules. This would mean developers or providers could be held responsible for safety problems or wrong information. But if LLMs are treated as general technologies or infrastructure, the focus would be more on rules for managing them, setting technical standards, and encouraging responsible

innovation. This makes it harder to decide who is responsible and how to create clear legal rules for problems like hallucinations.

4.2.2.3 Balance Innovation and Regulation

From a sociological perspective, this issue is closely linked to uncertainty avoidance, which measures the extent to which a society tolerates ambiguity and risk. High uncertainty avoidance cultures tend to prefer clear rules and structure, while low uncertainty avoidance cultures are generally more open to change and experimentation. This difference is reflected in regulation: the former emphasises guidance and control, whereas the latter is more likely to support innovation.

Research and public policy have long suggested that the regulatory environment the rules that define the range of permissible behaviours[170] is an important determinant of innovation among firms. Restrictiveness can have both a negative and positive relationship with innovation output depending on the level of regulatory uncertainty and the innovation type in question.[171] Pelkmans and Renda[172] suggest that highly detailed rule-based regulation often impedes innovation, while more flexible principle-based regulation may facilitate it by lowering compliance costs. Likewise, Castro and McLaughlin[173] argue that

---

[170] Lauren B Edelman and Mark C Suchman, 'The Legal Environments of Organizations' (1997) 23 Annual Review of Sociology 479.

[171] Michael Park, Shuping Wu and Russell J Funk, 'Regulation and Innovation Revisited: How Restrictive Environments Can Promote Destabilizing New Technologies' (2025) 36(1) Organization Science 240.

[172] Jacques Pelkmans and Andrea Renda, 'Does EU Regulation Hinder or Stimulate Innovation? '(*CEPS Special Report*, November 2014) <https://cdn.ceps.eu/wp-content/uploads/2015/01/No%2096%20EU%20Legislation%20and%20Innovation.pdf> accessed 17 October 2024.

[173] Daniel Castro and Michael McLaughlin, 'Ten Ways the Precautionary Principle Undermines Progress in Artificial Intelligence' (*Information*

precautionary AI regulation can raise research and development costs, discourage innovation, and weaken long term economic growth and technological competitiveness.

Both the EU and China legally affirm their commitment to supporting technological innovation. In the EU, the AIA expressly promotes the development, uptake, and use of AI in Article 1, provides for regulatory sandboxes to facilitate innovation in Article 53, and requires institutional support for AI research and development in Article 55. In practice, however, these objectives have not yet been fully realised. As LLMs become more advanced in reasoning and problem-solving, both jurisdictions have sought to promote technological progress while also responding to the risks posed by hallucinations. Even so, critics argue that existing regulatory frameworks may be overly restrictive and could inhibit innovation.

Moreover, both systems say they support innovation, but they do not clearly explain how much is allowed. Because of this, developers know they have some freedom under the rules, but they are not sure where the line is. This makes it hard to know how far they can go before the law steps in. Legally, both the EU and China use risk-based rules. These include hallucinations as part of bigger problems like wrong content or system-wide risks, but they do not give clear rules just for hallucinations. For example, the AIA says high-risk systems must be strong and accurate, but it does not say how much hallucination is too much. In the same way, China's Interim Measures say outputs must be "true and accurate," but they do not set clear error limits.

Technology and Innovation Foundation, 4 February 2019) <https://itif.org/publications/2019/02/04/ten-ways-precautionary-principle-undermines-progress-artificial-intelligence/> accessed 17 October 2024.

A bigger problem is that even though the rules slow down innovation, they have not done a good job of stopping hallucinations. This way of making rules causes two problems at once: it holds back progress in technology and fails to deal with the real dangers caused by hallucinations.

On one hand, developers are incentivised to improve overall model performance through scaling, RAG, and reasoning enhancements. On the other hand, without explicit legal benchmarks for hallucination rates or structured validation mechanisms, these improvements lack accountability and do not reliably address the residual hallucination harms.[174] Consequently, both jurisdictions see similar patterns. Technological progress reduces but does not eradicate hallucinations, while regulatory ambiguity sustains a persistent compliance gap.

To resolve this, governance must evolve beyond general risk-based flexibility and move toward hallucination-aware legal frameworks. This may involve the establishment of quantitative hallucination metrics, tiered accuracy requirements for high-impact or safety-critical domains, and transparent reporting obligations. Crucially, it also requires the development of operationalised provisions that clearly define the permissible boundaries of innovation. By coupling sustained technological innovation with measurable legal standards, both the EU and China can more effectively align normative expectations with technical realities, thereby addressing one of the most socially consequential shortcomings of modern LLMs systems.

4.2.2.4 Balance individual rights and social rights

[174] Sandra Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law' (2022) 97(2) Tulane Law Review 149.

From the perspective of individualism and collective interest, the key issue is the degree of connection between the individual and the group. Individualist cultures tend to emphasise personal achievement and autonomy, which in the regulation of Gen AI is reflected in a stronger focus on the protection of individual rights. By contrast, collectivist cultures place greater value on group harmony and the pursuit of collective interests, including broader social rights. The protection of individual rights in individualist systems and the preservation of social harmony in collectivist systems therefore require careful balancing. Similarly, other scholars[175] argue that clear guidelines and principles are essential to ensure the responsible development, deployment, and use of AI technologies. This further highlights the importance of cultural values in shaping how different countries understand, adopt, and regulate emerging technologies.

One key reason for distinguishing between micro- and macro-level harms in this thesis is that the same hallucinated output may manifest differently depending on whether it affects individuals or broader social structures. Legal and regulatory responses must therefore strike a careful balance between protecting individual rights and promoting collective societal interests. Focusing too much on reducing social risks can cause personal rights to be unfairly limited. But protecting individual freedom too much can slow down technology or make following rules too hard.

Notably, neither the EU nor China has paid sufficient attention to the issue of individual overreliance on LLMs, nor have they developed effective strategies to address this emerging risk. This gap shows the need for a clearer and more detailed governance system that looks at both personal risks and wider effects.

[175] José Manuel Simões and Wilson Caldeira, 'Ethics concerns in the use of computer-generated images for human communication' (2024) 4 Journal of Ethics in Higher Education 169.

Both the EU and China include rules to protect individual rights like privacy, personal dignity, and control over personal information in their laws for Gen AI. At the same time, both also focus on protecting public interests such as public order, truthfulness of information, and trust in society. This shows they agree hallucinations from LLMs can harm not just individuals but society.

In the EU, individual rights form the foundation of AI governance. The AIA, together with the GDPR and the EU Charter of Fundamental Rights, requires that AI systems protect personal data (Article 8 of EU Chapter 2), support freedom of expression and access to information (Article 11 of EU Chapter 2), and prevent discrimination (Article 21 of EU Chapter 3). Although hallucination is not separately defined, its effects such as reputational damage, misinformation, and disempowerment fall within these protective domains. For instance, an LLM that fabricates defamatory content or misrepresents legal obligations may infringe the rights of individuals to truthful and accurate information and compromise the integrity of the digital public sphere.

China has also adopted regulatory measures that reference individual rights, particularly in the areas of privacy and data security. For example, Article 11 of the Interim Measures explicitly emphasises the protection of personal information, underscoring the importance of safeguarding individual rights within the broader framework of AI governance. While these obligations are often framed through the lens of maintaining social stability, they nonetheless extend protection to users who may suffer material or moral harm from hallucinated outputs. In practice, however, the prioritisation of collective interests such as national informational sovereignty and online content control means that the protection

of individual rights often occurs within a broader logic of state-led risk management.[176]

Despite these structural differences, both jurisdictions converge in their attempt to reconcile the micro-level effects of hallucinations on individuals with the macro-level risks they pose to society. Yet neither the EU nor China has fully clarified how conflicts between individual rights and collective interests should be resolved in specific regulatory contexts. This is particularly evident where the suppression of hallucinated content may restrict freedom of expression, or where platform obligations to protect the public may outweigh the interests of an individual user. In the EU, such tensions are likely to be mediated through the principles of necessity and proportionality. In China, by contrast, they are more commonly addressed through a public-interest standard exercised within a framework of administrative discretion.

A more explicit balancing framework would benefit both systems. This might include formalised tests for weighing individual harms against collective benefits, procedural safeguards ensuring redress, and transparency obligations requiring platforms to disclose the rationale for content interventions. Without such tools, there is a risk that rules for handling hallucinations, even if made with good intentions, could end up going too far one way or the other.

4.3 Key Differences Between the EU and China

4.3.1 Differences at the Theoretical Level

[176] Huw Roberts and others, 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation' in Luciano Floridi (ed), *Ethics, Governance, and Policies in Artificial Intelligence* (Springer 2021) 47.

4.3.1.1 Differences in Legislative Approaches to Hallucination-Related Risks

The divergence in overall legislative approach ultimately reflects deeper differences in social background and culture. First, the two jurisdictions differ in terms of power distance, that is, the degree to which a society accepts unequal distributions of power, such as hierarchical or pyramid shaped structures. China is often regarded as a relatively high-power distance culture, in which hierarchy is more readily accepted, subordinates are expected to defer to superiors, ordinary people are more inclined to comply with authority, and decision making tends to be concentrated. A meaningful comparison should therefore go beyond legal provisions and regulatory mechanisms alone. It should also take into account broader background differences, including the historical experience, legal traditions, and social foundations of the EU and China. In addition, market structure and economic incentives are highly relevant. The EU has often been positioned more strongly as a user market and content provider, whereas China has also sought to act as a developer and market leader. This affects both jurisdictions' regulatory motivations and their perceptions of risk. Finally, the comparison must engage with key normative debates, including the balance between innovation and regulation, and the relationship between rights protection and state governance. Against this background, the EU and China adopt notably different legislative approaches to the hallucination risks posed by LLMs, reflecting divergent legal traditions and institutional structures.

Against this background, the EU adopts a broad, top-down, and risk-based regulatory approach, with the AIA serving as its central legal instrument. This comprehensive framework is designed to cover most AI systems operating within the EU, thereby ensuring uniform standards across diverse applications and sectors. The AIA tries to create one set of rules for the whole life of AI systems, from how they are built and put on the market to how they are used. Together with other laws like the DA and DSA, it is part of a broad system of rules that covers multiple

sectors, actors, and technological forms. Central to the AIA is its broad definition of "AI systems" as “machine-based systems designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, by inferring from input data to produce outputs such as predictions, content, recommendations, or decisions influencing physical or virtual environments.” The most relevance is the Act’s explicit treatment of GPAI models, including foundational models that can be integrated across a wide range of downstream applications.

This horizontal, comprehensive framework not only defines legal obligations across actors and sectors, but also incorporates key concepts such as risk classification, conformity assessments, and governance obligations under a single legislative instrument. In practice, this means that developers and deployers must assess not only whether their activities fall under the scope of the AIA, but also how they intersect with other legal regimes, including data protection, intellectual property, and sector-specific regulations in fields such as healthcare or automotive technology. However, the AIA offers a clear and complete system, bringing together many key responsibilities in a single main law.

In contrast, China has adopted a more centralised, vertical, fragmented approach to AI regulation. Instead of a single comprehensive law, Chinese authorities have advanced governance through a two-track strategy: (1) sector-specific regulations (such as the Algorithmic Recommendations Provisions, the Deep Synthesis Provisions, and the Interim Measures); and (2) the issuance of technical standards and pilot governance projects aimed at accumulating practical experience and best practices. While there is some overlap, each rule usually applies to a specific AI service or function. As a result, China has not yet provided a unified legal definition of “AI” within binding legislation, although definitions can be found in several technical standards.

This regulatory fragmentation means that compliance in China is primarily determined by whether a given service falls within the scope of an individual regulation, rather than whether it meets a general definition of an "AI system." Moreover, China's legislative technique characterised as "small, fast, and flexible" rulemaking prioritises rapid institutional responses to technological uncertainty over the construction of a fixed legal taxonomy. On the surface, this reflects a pragmatic response to rapid AI development, but it also reveals deeper institutional features: the Chinese legal system often adjusts its normative stance toward technology through iterative administrative experimentation rather than codified legal abstraction.

The root divergence lies in the distinction between horizontal and vertical regulation. While the EU's horizontal approach seeks to harmonise and integrate AI governance under a coherent transnational legal regime, China's vertical model allows more dynamic, sector-driven responses. Nonetheless, both models face limitations. In the EU, horizontal integration may result in regulatory rigidity or conflicts between overlapping laws. In China, vertical fragmentation can lead to definitional uncertainty and gaps in cross-sector oversight. At the time of writing, reports suggest that China may be moving towards developing a unified AI law, which may signal a shift in strategy and affect future convergence with the EU's regulatory model.[177]

4.3.1.2 Differences in Regulatory Approaches to Hallucination Risk

The EU and China adopt fundamentally different approaches to hallucination risks in AI systems, shaped by contrasting regulatory philosophies. The EU follows a

[177] Geotechnopolitics, China's AI Law: Recent Developments and Legislative Proposals" (*Geopolitechs*, 25 June 2025) <https://www.geopolitechs.org/p/chinas-ai-law-recent-developments> Accessed 12 October 2025.

rights-based and risk-based model grounded in the protection of human rights, with particular emphasis on user protection, risk classification, and fundamental rights. As Karen Yeung and others argued,[178] human rights standards provide one of the most coherent normative foundations for AI ethics and governance. The EU's regulatory approach therefore also reflects a strong concern for the protection of human rights. China, by contrast, adopts a more state led approach, prioritising national security, social stability, and industrial development. These differences help explain the distinct legal and regulatory responses of the two jurisdictions.

The EU adopts a balanced approach between individual rights and collective well-being. As a low power distance and collectivist region, the EU emphasises transparency, ethical considerations, and collective interest, prioritises harm prevention through transparent regulation without completely banning deepfakes, reflecting the region's values of privacy protection and societal welfare.[179] On top of this, it can be read that in accordance with the human-centricity principle the AIA emphasises that AI development should be orientated towards societal benefits, not only technological advancement for its own sake.[180]

In addition, the European ethical principles emerge from a more individual-focused and rights-based approach. They express a different aspiration, rooted in

[178] Karen Yeung, Andrew Howes and Ganna Pogrebna, 'AI Governance by Human Rights-Centered Design, Deliberation, and Oversight: An End to Ethics Washing' in Markus D Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (OUP 2020) 76.

[179] Zhang Long, John A Parnell and Eric B Dent, 'Individualism, Collectivism and Management in China: Does Atlas Shrug in China?' (2019) 20(3) Journal of Asia-Pacific Business 166.

[180] Nathalie A Smuha, 'From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence' (2021) 13 Law, Innovation and Technology 57.

the Enlightenment tradition, and coloured by the European history. Their primary goal is to protect individuals against well-identified harms. Whereas the Chinese principles emphasise the promotion of good practices, the EU focuses on the prevention of evil consequences. The former draws a direction for the development of AI, so that it contributes to the improvement of society. The latter sets limitations to its uses, so that it does not happen at the expense of certain categories of people. This distinction is clearly illustrated by the presentation of fairness, diversity and inclusiveness. While the EU emphasises fairness and diversity with regard to individuals from specific demographic groups (specifying gender, ethnicity, disability, etc.), Chinese guidelines urge for the upgrade of "all industries", reduction of "regional disparities" and prevention of data monopoly. While the EU insists on the protection of vulnerable persons and potential victims, China prescribes "inclusive development through better education and training, support".

Under the AIA, various roles in the LLMs lifecycle are precisely defined, with detailed compliance obligations corresponding to each role. Providers of AI systems or GPAI models are required to implement technical and organisational safeguards, maintain quality management systems, and conduct post-market monitoring. Deployers must ensure that the system's actual use aligns with its intended purpose and regulatory expectations. Importers and distributors are obligated to verify product compliance and ensure ongoing system reliability when placing it on the Union market. This clearly articulated allocation of responsibilities is designed to prevent accountability gaps in addressing potential harms, including those caused by hallucinations.

In contrast, China's regulatory structure is more service-specific and function-oriented. Instead of regulating AI as a lifecycle-wide ecosystem, Chinese rules target distinct service providers such as those offering recommendation algorithms, deep synthesis technologies, or Gen AI applications. These entities are

obligated to ensure content legality, protect user privacy, conduct security assessments, and submit algorithmic filings to government authorities. While this approach helps address specific harms, including hallucinations, it does not set out role-based duties across the LLMs lifecycle in the way the AIA does.

In China, a high sense of social responsibility and self-discipline is also expected from individuals to harmoniously partake in a community while promoting shared responsibilities and open collaboration. The emphasis is explicitly informed by the Confucian value of “harmony” as an ideal balance to be achieved through the control of extreme passions to avoid conflicts. Other than a stern admonition against “illegal use of personal data”, such value leaves little room for constraining rules. These principles are not paths to regulation, what would be detrimental to the development of research and business opportunities in a highly competitive environment where innovation is crucial. Rather, they are framed to guide AI providers in a way that would promote collective good for the Chinese society.

The difference in regulatory design also stems from differences in risk classification frameworks. Since both the EU and China treat hallucinations as part of broader AI-related risks rather than as a distinct regulatory category, their respective risk classification systems indirectly inform the governance of hallucinations. In this sense, although hallucinations are not explicitly singled out, the general risk-based structure provides a partial foundation for their regulation. The AIA uses a hierarchical risk-based system, dividing AI into prohibited, high-risk, and transparency-required categories. AI systems seen as too dangerous, such as those used for social scoring, emotion recognition at work, or untargeted biometric surveillance, are banned. High-risk AI systems must follow strict legal rules, including risk management, technical documentation, data governance, human oversight, and transparency. In contrast, China's classification of AI risks is less formalised and more pragmatic, focusing on the functional characteristics of

specific services. Risk assessment is often carried out through administrative guidance, sectoral guidelines, and real-world pilot governance projects, rather than codified statutory tiers.

The underlying reasons for these differences are multi-layered. Firstly, the EU's approach reflects its supranational legal tradition, which emphasises rule-based harmonisation and precautionary regulation, particularly in fields that implicate fundamental rights. Hallucination risks, although not explicitly defined in the AIA, fall within this broader logic of ex ante control and structured liability allocation.[181] The EU's horizontal governance model favours comprehensive and anticipatory legal intervention to prevent harms before deployment.

Secondly, China's system is shaped by a governance model that prioritises administrative responsiveness, sectoral flexibility, and state-led coordination. Rather than prescribing abstract legal duties, regulators prefer to iterate policy through guidance documents, algorithm registries, and multi-stakeholder governance pilots. This approach is partly institutional, as China lacks a tradition of unified horizontal legislation in emerging technologies, and partly strategic, allowing regulators to address fast-evolving AI harms such as hallucinations with sector-specific interventions that can be updated as needed.

Finally, the technical environments and market incentives in each jurisdiction also play a role. The EU's fragmented digital market benefits from unified legal certainty to enable cross-border AI innovation, whereas China's highly centralised digital economy facilitates direct state oversight and vertical intervention. In both places, the risk of hallucination is handled in an indirect way. The EU uses rules

[181] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' (2021) 41 Computer Law & Security Review 105567.

based on different risk levels, while China focuses on who is responsible for content and making algorithms clear. But neither system treats hallucination as its own rule category, which makes it harder to deal with the problem directly.

4.3.1.3 Differences in Lifecycle Coverage of LLMs

A key difference between the EU and China is how much their rules cover the full life of LLMs. The AIA in the EU takes a full lifecycle approach. As explained in Section 3.2, its rules cover most major steps in the life of an LLM, giving complete protection. This shows the EU's goal of making sure systems can be tracked and held responsible from the design stage to real-world use, helping reduce hallucination risks at different points.

By contrast, China's framework explicitly covers fewer stages of the LLM lifecycle. Article 4(2) of the Interim Measures refers to algorithm design, training data selection, model generation and optimisation, and service provision, thereby dividing the lifecycle into four broad stages. However, "model generation and optimisation" merges stages that this thesis treats separately, namely fine tuning and alignment, and deployment and interaction. Since each stage gives rise to distinct risks and regulatory concerns, this compression may weaken the identification and regulation of hallucination related harms. More broadly, China's AI rules focus on providers of specific services, such as algorithmic recommendation, deep synthesis, and Gen AI. This reflects a dual track strategy of service specific regulation, supplemented by technical standards and pilot governance projects. While pragmatic, this fragmented model does not fully cover the entire LLM lifecycle.

Furthermore, as noted in Section 3.2, China's current regulatory focus particularly under the Interim Measures, remains concentrated on the monitoring stage, which

limits the broader efficacy of lifecycle-based governance. This fragmented approach may hinder the development of a more precise and responsive regulatory framework capable of addressing hallucinations across the full model lifecycle.

This structural divergence can be attributed to several underlying factors. Firstly, the EU's regulation covers the whole lifecycle and is based on its legal tradition of setting rules before problems happen, especially when fundamental rights are involved. The AIA tries to manage risks like hallucinations early on, during development and testing, before AI is used by the public.[182] Moreover, the EU's internal market logic necessitates harmonised obligations to reduce regulatory fragmentation across Member States, justifying a lifecycle-wide regulatory scope.

Conversely, China's approach is shaped by a function-oriented regulatory logic, rooted in administrative pragmatism and sector-specific intervention. The focus on output and content control aligns with the Chinese state's core priorities: maintaining social stability, managing online information flows, and responding flexibly to technological developments. This leads to a governance model that prioritises accountability for service providers at the point of user interaction, rather than imposing upstream obligations on infrastructure providers. Because of this, China can respond quickly to problems like hallucinations, but it may not have strong tools to manage risks across the whole LLMs lifecycle.

This difference matters a lot. The EU's rules cover every stage of AI development, helping to stop hallucination risks from the start. China, on the other hand, focuses on service providers. This can make responses faster but might cause

[182] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019(2) Columbia Business Law Review 494.

fewer steady policies. This shows that both different systems and different ideas about who should be responsible for hallucination harms play a role in how the rules are made.

4.3.2 Differences at the Practical Level

4.3.2.1 Differences in Enforcement Institutions and Legal Environments

Power Distance Index (PDI) refers to the extent to which less powerful members of a society accept and expect unequal distributions of power. In high power distance cultures, hierarchical order is more readily accepted and less frequently questioned, whereas low power distance cultures place greater emphasis on equality and expect power disparities to be justified. This distinction helps explain important differences between China and the EU in the enforcement of AI regulation. China's governance model is more state led and reflects a relatively high power distance structure, characterised by top down authority, hierarchical compliance, and concentrated decision making. By contrast, the EU tends to favour a more pluralistic and decentralised model, with flatter institutional structures, dispersed responsibility, and the involvement of multiple actors and regulatory bodies. These differences in political and legal culture shape how AI rules are enforced in practice and, in turn, how hallucination risks and other AI related problems are addressed.

In the EU, the AIA sets wide rules for high-risk AI systems. These rules cover many areas like managing risks, handling data, keeping technical records, being open, cybersecurity, and human checks. The AIA also needs tests to make sure AI systems follow safety laws. Each Member State gives enforcement work to national authorities. These authorities watch for rule-breaking, investigate problems, and fix issues. People are still talking about whether data protection offices, for

example, this function could be performed by Germany's Federal Commissioner for Data Protection and Freedom of Information (Bundesbeauftragter für den Datenschutz und die Informationsfreiheit, BfDI) or France's Commission Nationale de l'Informatique et des Libertés (CNIL). To keep rules the same in all EU countries, the EU made a main AI Office. This office checks GPAI models to make sure rules are followed the same way in the EU. It also helps with using voluntary rules and trying new ideas in special test areas called sandboxes.

In contrast, China, a high-power distance and collectivist society, has implemented stringent regulations to control deepfake creation and dissemination. China's AI rules are more centralised and controlled from the top. Several government agencies share power over AI systems. It depends on the sector and rules. The Chinese government's top-down approach reflects its cultural emphasis on centralised authority and collective good over individual rights. These regulations aim to protect social harmony and maintain control over the digital landscape, aligning with China's values of order and unity.[183] The CAC serves as the lead regulator in areas such as Gen AI, algorithm recommendation, and deep synthesis, while MIIT takes the lead on telecoms, IoT, and mobile AI applications. These agencies have local offices that follow central government policies in different regions. The rules focus mostly on security checks, clear algorithms, user rights, and content control. Providers must register their algorithms and set up internal compliance systems.

At first glance, Chinese ethical principles appear similar to those of the EU in several respects. Both notably promote fairness, robustness, privacy, safety, and transparency. Their prescriptive approaches, however, reveal different cultural perspectives associated with different objectives. Influenced by Chinese

[183] Marijana Krkić, 'Cultural Perspectives on AI Usage and Regulation in Deepfake Creation: How Culture Shapes AI Practices' (2025) 12 International Communication of Chinese Culture 225.

traditional cultural heritage.[184] Confucian philosophy has shaped the governing system in China and the rest of East Asia for centuries. It emphasises the "rule for the people", rather than rule by the people", and favours an elitist leadership, associating political mandates with competence and merit. The Chinese government's belief in "doing the right thing" for its citizens is informed by the Confucian ideas of virtuous authority and exemplary person, grounded in ren (humaneness), yi (appropriateness), li (rite), and zhi (wisdom). This philosophical tradition explains the community-focused and goal-oriented perspective, from which the Chinese guidelines derive, together with the promotions of principles, such as "harmony and friendship", "shared responsibilities", "tolerance and sharing", and "open collaboration". The European ethical principles, in contrast, emerge from a more individual-focused and rights-based approach. They express a different aspiration, rooted in the Enlightenment tradition, and coloured by the European history. Their primary goal is to protect individuals against well-identified harms. Whereas the Chinese principles emphasise the promotion of good practices, the EU focuses on the prevention of evil consequences. The former draws a direction for the development of AI, so that it contributes to the improvement of society. The latter sets limitations to its uses, so that it does not happen at the expense of certain categories of people. This distinction is clearly illustrated by the presentation of fairness, diversity and inclusiveness. While the EU emphasises fairness and diversity with regard to individuals from specific demographic groups (specifying gender, ethnicity, disability, etc.), Chinese guidelines urge for the upgrade of "all industries", reduction of "regional disparities" and prevention of data monopoly. While the EU insists on the protection of vulnerable persons and potential victims, China prescribes "inclusive development through better education and training, support."

[184] Pascale Fung and Hubert Etienne, 'Confucius, Cyberpunk and Mr. Science: Comparing AI Ethics Principles between China and the EU' (2023) 3(2) AI and Ethics 505.

The difference between the EU and China comes from basic differences in administrative law and how they think about rules. The EU uses a system where national authorities have the main power to enforce rules. Central bodies like the AI Office help and guide but do not give orders. In China, the system is more top-down. The central government makes the rules and checks how they are followed. Local offices follow these rules strictly. This setup lets China act fast and keep rules the same everywhere. But it also means one central group controls how the rules are understood and used. This can limit different opinions on enforcement.

These differences in structure change how companies work with regulators. In the EU, enforcement is spread out. An AI company working in different Member States might see different views on how to follow the AIA. This can cause confusion, but it also lets each country adjust rules to local needs. In China, many agencies are involved, but the top-down system means local offices follow central rules closely. So, companies may deal with more than one regulator, but the rules and how they apply are usually the same across the country.

Looking ahead, these differing enforcement structures may produce different outcomes in cross-border regulatory scenarios. In this respect, a European provider seeking to market AI service in China may be required to comply with centralised filing and security review procedures. By contrast, a Chinese AI provider entering the EU market must navigate a more decentralised enforcement landscape involving multiple national authorities, depending on the Member State concerned. Understanding the operation of each system is therefore essential for global companies seeking to ensure compliance across both jurisdictions.

4.3.2.2 Differences in Implementation Effectiveness

A comparison of how the EU and China seek to mitigate the harms caused by hallucinations shows that, although they share certain regulatory concerns, they address these issues in materially different ways at both the micro and macro levels. These divergences are also shaped by their broader political economy and regulatory priorities. China is more strongly positioned as an aspiring market leader in AI development, whereas the EU operates within a different incentive structure that places greater emphasis on the regulation and governance of AI deployment and use.

At the micro level, both jurisdictions recognise that hallucinations generated by LLMs can cause cognitive misguidance and foster undue user reliance. In each regulatory environment, users may overestimate the epistemic authority of LLMs, particularly in high-risk contexts such as legal and medical advice. At the same time, although both systems formally encourage corrective feedback, neither has yet developed robust and user-friendly mechanisms through which individuals can report hallucinations in ways that meaningfully improve model performance. Important differences nevertheless remain. Chinese platforms more frequently display warning messages, disclaimers, or responses indicating uncertainty when the model lacks confidence, reflecting a stronger emphasis on content governance and preventive control. This follows government content rules. In the EU, platforms follow softer rules focused on openness, but many actions are optional. This means no strong rules make providers clearly show when a hallucination might happen.[185]

At the macro level, both the EU and China face serious challenges arising from AI-generated false or fabricated content, particularly in relation to public trust and

[185] Błażej Sajduk and Dominika Dziwisz, 'Comparative Analysis of AI Development Strategies: A Study of China's Ambitions and the EU's Regulatory Framework' (EuroHub4Sino Policy Paper 2024/12, 20 September 2024) <https://storage.eh4s.eu/vitrin/files%2FComparative-Analysis-of-AI-Development-Strategies.pdf> accessed 26 July 2025.

the integrity of the information environment. Their governance responses, however, differ significantly. In the EU, responsibility is distributed across multiple institutional actors, and the regulatory framework seeks to balance fundamental rights, transparency, and provider accountability. Yet this plural and rights-based structure can also make enforcement slower and more complex. In China, by contrast, governance is more centralised and state-led.[186] This enables relatively rapid intervention through mechanisms such as model filing, content filtering, and rectification orders. Although these tools may contain risks more quickly, they are directed primarily towards maintaining political order, ethical conformity, and information control, rather than improving the epistemic reliability of AI-generated content as such. More broadly, whereas the EU seeks to build public trust through participatory governance and the gradual strengthening of digital literacy, China relies more heavily on top-down regulatory coordination aimed at preserving social stability and controlling the circulation of information.

4.3.2.3 Differences in Extraterritorial Reach

One clear difference between the EU and China in LLM regulation is how far their rules reach beyond their own countries. While both jurisdictions are actively developing AI governance structures, the EU leverages its regulatory power not only to manage domestic risks but also to project normative values internationally. This phenomenon, commonly referred to as the "Brussels Effect", captures the EU's ability to shape global markets through unilateral regulation without relying on international cooperation.

The extraterritorial reach of the AIA significantly enhances its global impact. Under Article 2, the AIA applies not only to AI systems placed on the EU market or

[186] Jia Wu, 'The Choice of Legislative Regulatory Model of Generative Artificial Intelligence in China'(中国生成式人工智能立法监管模式选择) (2024) 4 China Legal Science 133.

used within the EU, but also to outputs generated by systems deployed outside the EU if such outputs are used within the Union. Consequently, companies operating globally must ensure compliance with EU standards if their AI systems or outputs are accessible to users within the Union. In contrast, China's AI rules mainly apply within its own borders. Only AI services that officially enter the Chinese market are covered by these rules. If the content is made outside China, the current rules do not apply.

The impact of this divergence in regulatory reach is also evident in the nature and complexity of compliance obligations. The AIA imposes a detailed set of technical and organisational requirements for high-risk AI systems throughout their lifecycle. These include risk management protocols, data governance measures to ensure training and testing data quality, documentation, and record-keeping duties to demonstrate compliance and facilitate post-market monitoring, transparency obligations requiring systems to be understandable and accompanied by user instructions, safeguards to ensure accuracy, robustness and cybersecurity, and human oversight mechanisms tailored to system autonomy and use context. These obligations collectively form a comprehensive compliance architecture. In practice, several technology firms have already begun to align their internal compliance processes with the anticipated AIA requirements, such as OpenAI[187].

Chinese regulatory requirements touch upon similar concerns but diverge in regulatory technique. For instance, China imposes algorithm filing obligations requiring providers to submit detailed information to authorities. It also mandates pre-deployment security assessments for AI services with significant public

[187] PYMNTS, 'AI Regulations: OpenAI Calls on EU to Review, Simplify AI Rules' (*PYMNTS*, 17 April 2025) <https://www.pymnts.com/artificial-intelligence-2/2025/ai-regulations-openai-calls-on-eu-to-review-simplify-ai-rules/> accessed 20 June 2025.

influence. These assessments cover risk prevention, user protection, content moderation, model integrity and data security.

Three structural distinctions between the EU and China's regulatory systems are particularly noteworthy:

Firstly, conformity assessments in the EU are integrated into existing product safety regimes and involve third-party evaluations. This procedure is rare in China, where AI regulation relies more on government-led review mechanisms. Although local regulations like the Shanghai AI Ordinance have introduced elements of structured compliance, these are not yet widespread.

Secondly, algorithmic filing in China is centralised and mandatory for specific services, with dedicated portals for submission and tracking. The EU does require certain forms of registration under the AIA but has not yet established a centralised filing database for public use.

Thirdly, security assessments in China focus on the potential for AI systems to influence public opinion or mobilise collective behaviour. These assessments must be completed before services are launched. In contrast, the EU's pre-market scrutiny revolves around technical conformity and product safety, often certified by external bodies rather than state regulators.

Fundamentally, the EU aims to globalise its regulatory model and embed its normative standards in international commercial and legal frameworks. China, while increasingly assertive in proposing a "Chinese approach to AI governance," focuses more on domestic stability, content control, and incremental regulatory experimentation.

As a result, the AIA is poised to shape international compliance norms, particularly for multinational technology companies. China's regulatory model, although rigorous and rapidly evolving, is more likely to retain influence within its national jurisdiction or within regions where Chinese digital infrastructure plays a dominant role.

4.4 Conclusion

This chapter examined the similarities and differences between the EU and China in regulating the risks of hallucinations generated by LLMs, focusing on their legal frameworks, institutional arrangements, regulatory scope, enforcement practices, and international influence. Despite the significant cultural, political, and legal divergence between these two jurisdictions, both have converged on several fundamental principles. Notably, both recognise the importance of transparency, accountability, and lifecycle governance in mitigating hallucination harms, and both acknowledge the necessity of aligning AI regulation with existing laws, especially those concerning data protection and cybersecurity.

However, the legal structures adopted to achieve these ends vary markedly. The AIA exemplifies a horizontal and comprehensive legislative approach, striving to establish an integrated risk-based framework that applies uniformly across Member States. It adopts a tiered risk classification model and imposes detailed obligations on providers, deployers, importers, and distributors throughout the LLMs lifecycle. This structure enables proactive governance and anticipatory risk mitigation, particularly for high-risk systems, including LLMs. In contrast, China has adopted a more vertical and function-specific regulatory approach, issuing service-type-focused rules targeting providers of algorithmic recommendation, deep synthesis, and Gen AI services. Instead of defining AI or LLMs generally, China focuses on controlling the results and how AI is used, using tools like algorithm filing, content monitoring, and security checks.

These structural differences come from deeper legal and political traditions. The EU's system is based on liberal-democratic values and the goal to keep rules the same across its internal market. China's system is based on practical administration, keeping social stability, and rule by policy. So, the EU focuses on clear accountability and consistent rules across countries, while China focuses on getting results quickly and dealing fast with new risks.

Looking at the whole lifecycle, as discussed in Section 4.3.1.3, the EU covers most stages of model development, including pre-training, fine-tuning, deployment, and post-market monitoring. It focuses on early prevention through rules, record-keeping, and risk checks. In contrast, China displays a narrower lifecycle reach. Regulatory efforts tend to concentrate on the downstream stages particularly model deployment and output supervision while upstream activities such as data sourcing, pre-training architecture, and alignment techniques are subject to comparatively less scrutiny. This difference matters in practice: the EU's wide coverage allows it to act earlier but can slow down regulation, while China's focus on later stages helps it react faster to visible problems but may miss issues that start earlier in development.

From a practical implementation perspective, both systems exhibit partial convergence in their recognition of hallucination-induced harms at both the micro and macro levels. At the micro level, both seek to mitigate individual risks such as misinformation acceptance, overreliance, and the erosion of user feedback loops. However, their strategies are different. The EU prefers to rely on ex ante governance mechanisms, including conformity assessments, documentation requirements, and risk management protocols, which measures that aim to prevent hallucination risks before deployment. In contrast, China focuses more on rules after problems occur, including mandatory disclaimers, filtering systems, and administrative penalties for violating content standards. Chinese platforms

often add hallucination alerts and fast-response systems as the government asks. EU providers rely more on voluntary openness and internal checks.

At the macro level, both the EU and China face challenges from hallucinations like spreading false information and weakening public trust in Gen AI and they take different measures to solve this problem. The EU tries to extend its influence globally through the “Brussels Effect,” encouraging other countries to adopt similar rules. China’s influence is mostly regional but growing, especially through partnerships with countries in the Global South, promoting its own style of digital governance. However, neither the EU nor China currently treats hallucinations as a separate regulatory risk. Instead, hallucinations are grouped under broader categories like output harm, misinformation, or inaccuracies or even risks of Gen AI. This makes it harder to address the unique and complex problems hallucinations.

Both the EU and China are moving toward fuller rules that cover all stages of LLMs, but each has its own strong points and weaknesses. The EU has clear, comprehensive, and risk-based laws that can be slow to change. China acts faster and focuses on enforcement but has less focus on legal principles. These differences show chances to learn from each other: the EU could try China’s quick testing of rules, and China could build stronger legal foundations like the EU. In the future, mixing these approaches, working across borders, and keeping up talks will be key to making rules that can change with the times and handle the changing risks of hallucinations in LLMs.

Chapter 5 - Lessons: Experiences and Insights from the EU and China

5.1 Introduction

As hallucinations in LLMs become more serious, the EU and China have adopted different regulatory approaches shaped by their political systems, institutional arrangements, and regulatory philosophies. Even so, both seek to address common challenges arising from the spread of AI, including safety, accountability, innovation, and rights protection. This chapter examines what can be learned from both jurisdictions and how these lessons may inform the regulation of hallucinations and the wider development of global AI governance.

China may learn from the coherence of the EU AI AIA, which aims to balance innovation, legal certainty, and competitiveness within a unified framework. At the same time, China's own regulatory model has drawn growing international attention as an example of how a major AI market seeks to balance innovation, social order, and state control through sector specific regulation, administrative coordination, and flexible governance. Although this approach differs from the EU's more rights-based model, it may still offer useful lessons for other jurisdictions.

At the same time, the EU is a global leader in setting rules for AI. The AIA is the world's first full legal framework just for AI regulation, using a risk-based approach to support both innovation and protect people. The EU wants not only to regulate its own market but also to use "Brussels Effect" to spread its values to other countries.

Against this background, comparing the EU and China provides a rich practical foundation to derive broader lessons about effective LLMs governance. The EU and

China have different ways to regulate, but both agree it is very important to reduce the harms of LLMs. These harms include algorithm bias, wrong information, hallucinations in LLMs, and loss of public trust. Both also agree that AI should be regulated throughout its lifecycle, though they differ in how much the law covers this and how they put it into practice.

This chapter aims to bring together and evaluate the main lessons from the EU and China. It focuses on two points: what China can learn from the EU and what lessons China can offer to the world.

5.2 Lessons for China in Regulating LLMs Hallucinations

5.2.1 Enhancing Global Regulatory Influence

The international landscape is characterised by diverse regulatory approaches and standards, which can create inconsistencies and complicate compliance for multinational corporations. Hence, the EU effort to establish pioneering standards deserves to be appreciated. As mentioned above, without a coordinated global effort, disparate regulations may lead to regulatory arbitrage where companies exploit more lenient jurisdictions.[188]

Given the transnational character of AI technologies and the universality of challenges such as hallucinations, regulatory efforts must be designed with an international perspective. AI regulation has become a key component of China's broader legal diplomacy and external rule-of-law agenda. Because AI development and its associated harms transcend national borders, China's regulatory framework

[188] Christopher Kuner, 'The European Union and the Search for an International Data Protection Framework' (2014) 2(2) Groningen Journal of International Law 55.

must reflect not only domestic governance goals but also its strategic positioning in global competition. Instead of wholesale transplantation of EU-style horizontal legislation, China should adopt a more prudent and context-specific approach that balances global regulatory convergence with the preservation of sovereign interests. This includes actively engaging in international AI rule-making processes and promoting a uniquely Chinese governance model that reflects domestic political and institutional realities while remaining interoperable with international norms.

5.2.2 Balancing Innovation and Risk

Critics from within the EU, but also from non-EU competitors, the latter not without a drop of satisfaction, argue that overly burdensome requirements deter investment in AI technologies or lead companies to relocate operations outside the EU where regulations are less stringent.[189] The concerns regarding legal and regulatory compliance costs are echoed by scholars who argue that high regulatory burdens may disproportionately affect smaller firms that lack the financial resources to assure systems, advisors, or internal qualified staff to comply with extensive regulatory frameworks.[190]

The UK Information Commissioner's Office (ICO) has guided on this matter, clarifying that the accuracy principle applies to both the input data fed into an AI system and the outputs it generates. "However," the ICO notes, "this does not mean that an AI system needs to be 100 % statistically accurate to comply with

[189] Richard Owen, Phil Macnaghten and Jack Stilgoe, 'Responsible Research and Innovation: From Science in Society to Science for Society, with Society' (2012) 39(6) Science and Public Policy 751; Mark Ryan and Bernd Carsten Stahl, 'Artificial Intelligence Ethics Guidelines for Developers and Users: Clarifying Their Content and Normative Implications' (2021) 19(1) Journal of Information, Communication and Ethics in Society 61.

[190] Joshua Gans, *The Disruption Dilemma* (MIT Press 2016) 72.

the accuracy principle. In many cases, the outputs of an AI system are not intended to be treated as factual information about the individual."[191] Hence, minor inaccuracies, such as a wrong birthday date, may not constitute a breach of the accuracy principle unless they have significant consequences.[192] Similarly, under the GDPR's accuracy principle, significant false information generated by AI, such as wrongly attributing a movie role to an actor or incorrect supposed statements of public interest (e.g., election contexts), may require correction, especially when it impacts the individual's professional reputation or public record.[193] Although new tools are being developed to detect hallucinations, they remain probabilistic in operation and are unlikely to identify or eliminate all hallucinations, especially in high stakes contexts.[194] Their underlying principles may offer useful guidance, but the applicable standards remain insufficiently clear. It is therefore important not to sacrifice innovation in the name of compliance. What is required instead is a balanced approach that preserves long term competitiveness while ensuring that AI develops in a manner consistent with ethical principles and societal wellbeing. Striking this balance is essential if AI is to serve humanity rather than intensify authoritarianism, inequality, and geopolitical tension.

---

[191] Information Commissioner's Office, Guidance on AI and Data Protection (*Information Commissioner's Office*, 15 March 2023) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/> accessed 26 March 2026.

[192] Claudio Novelli and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (2024) 55 Computer Law & Security Review 106066.

[193] Philipp Hacker and others, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' in Philipp Hacker, Andreas Engel, Sarah Hammer and Brent Mittelstadt (eds), *The Oxford Handbook of the Foundations and Regulation of Generative AI* (OUP 2025).

[194] Sebastian Farquhar and others, 'Detecting Hallucinations in Large Language Models Using Semantic Entropy' (2024) 630 Nature 625.

By establishing the rules for AI development and deployment, the AIA was claimed to be reducing administrative burdens on businesses, particularly small and mediumsised enterprises.[195] This ultimately has been intended to foster an environment conducive to investment and innovation since companies could operate with greater certainty regarding compliance requirements. So much for the declarations. The practice will show to what extent the regulatory framework created by the AIA has helped to attract investments and foster innovations and not to overburden SMEs with a complex set of duties and responsibilities.

Whereas the EU framework is rooted in the core Enlightenment values of individual freedom, equal rights and serves to protect against State abuses, the Chinese guidelines are based on the Confucian values of virtuous government, harmonious society, and targets to protect against commercial exploitation.

The AIA represents a cautious yet structured first step in AI governance. One of its strengths lies in categorising AI systems based on risk level and assigning corresponding legal obligations. This risk-tiered approach avoids blanket prohibition and instead builds flexible "guardrails" that allow innovation to proceed within defined safety parameters. From China's perspective, this calibrated method offers valuable lessons. In regulating hallucinations, a risk-based approach offers valuable guidance in balancing the positive use of hallucinations with efforts to mitigate their harmful effects. Specifically, a tiered and classified governance framework can be used to manage hallucinations of LLMs, which appear in different forms. Governance should be tailored to the type of content involved and the application context. Users see and accept hallucinations in different ways, depending on how serious they are and what kind they are. For factual or knowledge-based queries, hallucinations typically equate

[195] Athanasios Polyportis and Nikolaos Pahos, 'Navigating the Perils of Artificial Intelligence: A Focused Review on ChatGPT and Responsible Research and Innovation' (2024) 11 Humanities and Social Sciences Communications 107.

to errors, and their negative impact is significantly higher due to the low margin for error. In such cases, immediate and stringent regulatory measures are necessary. On the other hand, in creative work, hallucinations can show the model's ability to imagine new things and even do more than what humans can think of. So, it makes sense to set clear limits and make rules that treat different types of hallucinations based on how risky or useful they are to society.

5.2.3 Covering the Full Lifecycle of LLMs

A strong point of the EU system is that it covers the whole lifecycle and clearly divides responsibilities among different people involved in LLMs.[196] The AIA gives different legal duties to providers, deployers, importers, and distributors, which matches the complex supply chains and makes sure someone is responsible at each step. In China, the current rules mostly focus on service providers during the application stage. Learning from the EU, China's AI rules could improve by making roles and duties clear for every stage of LLMs, from training and using the model to watching it and controlling it after release. This would make the rules more accurate and help avoid missing important areas, like data handling and risks from basic models.

5.2.4 Expanding the Toolbox of Regulatory Mechanisms

This widespread adoption has required a proactive regulatory approach that seeks to anticipate potential risks rather than merely respond to them after they

[196] Luciano Floridi and others, 'capAI: A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act' (SSRN, 23 March 2022) <https://ssrn.com/abstract=4064091> accessed 26 March 2026.

materialise.[197] The AIA adopts a risk-based classification system that distinguishes between unacceptable, high, limited, and minimal risk applications. Although this provides a structured regulatory framework, it also creates difficulties in assessing risk accurately in light of the rapid and evolving nature of AI technologies. The relatively late inclusion of provisions on Gen AI further illustrates the challenge of maintaining regulatory relevance in a fast changing technological environment.[198] As a result, there remains a risk that the framework may fail to address emerging harms effectively.[199]

At present, China's main rules have little flexibility to handle new AI risks. At the same time, the EU has new ways to control risks, like conformity checks and regulatory sandboxes. These give more options to manage risks without stopping new ideas. For example, the AIA has a marking system for high-risk AI. If a system passes certain tests and meets legal rules, it gets a certification mark. This gives companies, especially small and medium businesses and start-ups, confidence that their products follow the rules. This system not only gives legal clarity but also lowers worry about following rules for smaller players.

5.3 Broader Lessons for Hallucinations Regulation in the EU and Other Jurisdictions

[197] Christiaan Stuurman and Eric Lachaud, 'Regulating AI: A Label to Complete the Newly Proposed Act on Artificial Intelligence' (2022) 44 Computer Law & Security Review 105657.

[198] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) 39—42.

[199] Mark Dempsey, Keegan McBride, Meeri Haataja and Joanna J Bryson, 'Transnational Digital Governance and Its Impact on Artificial Intelligence' in Justin B Bullock and others (eds), *The Oxford Handbook of AI Governance* (OUP 2024) 253.

China's experience in managing LLMs, especially regulating hallucinations, gives useful lessons for places that want to control fast-changing technologies and support local innovation. Its regulatory design has a lot of legislative flexibility and administrative centralisation. This helps make rules that match industry practice and can be enforced by steady government work. The EU uses a risk-based approach that focuses on a broad and preventive regulatory framework. China's strategy focuses more on quick response and specific sectors. It often lets AI policy be tested and changed in pilot projects before applying it across the country.[200] This difference gives important lessons for the EU, China, and global efforts to create strong and flexible AI governance frameworks.

5.3.1 Enhancing Flexibility in Governance

The EU has made important progress in setting rules for LLMs. The main part of this effort is the AIA, which builds a system of rules based on the risk level of AI systems. These rules include using data correctly, managing risks, being clear about how the system works, checking the system after use, and making sure it meets set standards. But the AIA is based on the idea that LLMs can be grouped and controlled using a fixed risk order. This idea might not work well for new technologies that have unpredictable uses and ways they are put to work.

By putting safety first, the EU approach may show "over-securitisation," where political worries take more focus than the need for quick innovation. For LLMs, this can cause problems because their use depends on the situation and changes fast. This makes fixed categories likely to become outdated soon after they are made. Since the rules for LLMs hallucinations come from the rules for LLMs

[200] Diyin Zhu and Bhaumik Amiya, 'The Impact of Artificial Intelligence on Business Strategy: A Review of Theoretical and Empirical Studies in China' (2025) 2(3) International Journal of Advances in Business and Management Research 9.

themselves, any weaknesses in the LLMs rules will also cause weaknesses in the rules for hallucinations. Moreover, the AIA's focus on ex ante documentation and conformity procedures can impose substantial burdens on smaller firms and start-ups, possibly chilling experimentation, and iteration critical to innovation.

China offers an alternative with its measured flexibility and adaptive strategy, combining sectoral rules such as those for algorithmic recommendation, deep synthesis, and Gen AI with real-time experimentation. These measures are implemented through administrative guidance, algorithmic filing, and pilot zones, which enable regulators to monitor technological developments and adjust policy dynamically. Rather than pre-emptively limiting innovation, this approach establishes "guardrails" that can evolve in pace with technology, blending administrative oversight with experimental autonomy. This more flexible multi-regulatory approach has advantages in regulating hallucination, which are relatively virtual, scattered, and appear in multiple scenarios, as well as the various harms they cause.

5.3.2 Supporting Implementation through Secondary Instruments

Despite the promise of the AIA, its impact remains muted in practice. In July 2024 EU introduced new legislation laying the foundations for regulating AI in AIA. It provides AI providers and deployers with requirements and obligations regarding specific uses of AI. The regulation seeks to reduce administrative and financial burdens for businesses, particularly SMEs, however, the stakeholders should be aware of the requirements and how and when they will affect their operations.[201] This highlights a gap between legislative intention and implementation readiness.

[201] European DIGITAL SME Alliance, 'New EU Initiatives to Ease the Transition of AI Act Regulatory Compliance for SMEs' (*European DIGITAL SME Alliance*) <https://www.digitalsme.eu/new-eu-initiatives-to-ease-the-transition-of-ai-act-regulatory-compliance-for-smes/> accessed 15 July 2025.

Good governance needs more than just main laws. It demands secondary instruments that clarify, operationalise, and contextualise legal obligations. Technical manuals, sector-specific benchmarks, and rolling guidance can demystify compliance for firms. Regulatory engagement must also be sustained: ongoing capacity-building, training initiatives, and consultation mechanisms can ensure the regulatory system evolves in parallel with technology and market developments.

China already uses many of these tools. Its centralised algorithmic filing system and mandatory risk assessments provide structured yet navigable compliance routes. Authorities regularly publish technical standards and guidance documents, enabling firms to understand and adjust to regulatory expectations. Test areas allow changes to be tried out before making official rules. This way, the rules can change with feedback and become a two-way process, not just one-sided orders.

5.3.3 Strengthening Technical and Innovation Infrastructure

A further lesson lies in the contrast between academic strength and industrial implementation. The EU boasts a strong research base in AI science. However, this foundation has not translated into scalable industrial applications, partly due to structural gaps: risk-averse investment culture, fragmented national policies, and cautious regulatory frameworks hamper the establishment of deep-tech enterprises.

China closes this gap by focusing on engineering and quickly using new AI models. Companies like DeepSeek do well in tasks for certain industries. They use their

large number of users and a lot of data to improve performance.[202] This approach, which emphasises initial AI system deployment followed by regulatory adjustments within existing sectoral governance frameworks, helps speed up commercialisation and allows developers to improve products continuously based on real-world feedback. To prevent the harm caused by the proliferation of "hallucination data" generated by large models on the internet, building a secure and trustworthy data labelling system to enhance content reliability is necessary. This includes establishing a secure and trustworthy, dynamically updated source of information and data knowledge base, creating a labelling system for the credibility and harmfulness of different types of data to reduce the probability of AI generating hallucinations and improve the reliability of generated content. Research and development of AIGC hallucination governance technology and platform, regular cleaning of hallucination data, study of technology and software platforms for automatic analysis of hallucinations, conducting automatic analysis of hallucinations, deep authentication of AIGC, detection of false information, identification of harmful content, and tracing the origin of internet dissemination. The Central Cyberspace Administration and the National Data Bureau, among other departments, will regularly clean up hallucination data and provide the public with AIGC hallucination information detection tools and services.[203]

Closing the “theory application” gap in the EU will require more than legislative finesse. It calls for investment in engineering capability, regulatory sandboxes that facilitate high impact pilots, and a catalytic venture capital environment that supports early stage risk taking. Harmonisation across EU member states is

---

[202] DeepSeek-AI and others, ‘DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model’ (arXiv, 7 May 2024) <http://arxiv.org/abs/2405.04434> accessed 19 June 2025.

[203] GMW.cn, ‘“Artificial Intelligence” Becomes a Hot Topic at the Two Sessions, with Deputies and Members Discussing Security Governance’ (“人工智能”成两会热词，代表委员热议安全治理) (*GMW.cn*,14 March 2025) <https://m.gmw.cn/toutiao/2025-03/14/content_37907541.htm> accessed 1 July 2025.

essential for reducing fragmentation in innovation policy. This approach can catalyse a virtuous cycle, where strong research drives applied innovation, and applied innovation in turn refines governance practice.

5.4 Conclusion

This chapter comparatively analysed the EU and China's approaches to regulating hallucination risks in LLMs across their lifecycle, drawing out valuable lessons that each jurisdiction can offer the other. The comparative insights reveal not only divergent legal cultures and institutional architectures, but also areas of potential regulatory convergence and mutual enrichment.

On one side, the EU's experience gives China useful lessons on how to build a clear and well-organised system to manage AI. The AIA is a good example of a law that covers all stages of AI development and focuses on protecting basic rights and being open about how AI works. This matches well with the complex risks of LLM hallucinations. It includes risk levels, required records, and rules that make AI makers take responsibility. It also works well with other laws, like data protection and product safety, to keep the rules connected. For China, using this kind of step-by-step and rights-based system can make the law easier to follow, more complete, and help build trust in AI. Moreover, the EU's combination of shared governance and flexible regulatory instruments, including Codes of Practice, offers a way to respond to technological change without placing disproportionate burdens on companies. China may draw on this approach in shaping its future regulatory framework.

On the other side, China's way of making rules also gives helpful ideas to the EU and other places. China uses a system that focuses on each sector and reacts quickly, with tools like the Interim Measures and rules for filing algorithms. This

shows that China can act fast and enforce rules well. Even though China does not have one full AI law, its flexible system allows quick steps, like filing, safety checks, and setting technical standards. While the EU takes a long time to pass new laws, China's step-by-step rulemaking shows a practical way to manage AI. This could give the EU ideas for using more flexible and trial-based methods, especially now as the AIA starts to be put into action. Moreover, China's experience with platform governance, verified-identity requirement, and integration of technical audit requirements into compliance regimes provides concrete enforcement tools that the EU can consider adapting, especially in contexts involving high-volume AI deployment or rapidly evolving misinformation risks.

In the end, what the EU and China can learn from each other shows that hallucination risks are a global issue. Both sides need to find a balance between encouraging new ideas, making clear rules, and keeping public trust. This thesis's lifecycle approach shows that hallucinations appear in different ways at each stage, so the law needs different answers at each point. Neither a static, rule-based model nor a purely reactive, fragmented one is sufficient on its own. The EU's strong structure and China's quick actions show that a mix of both ways works best. This kind of system should include legal rules for each stage, strong checks on the technology, and a focus on what works in real life.

This comparative analysis thus not only provides a roadmap for China to refine its AI legislation by selectively incorporating EU practices but also invites the EU to consider China's institutional innovations and practical enforcement strategies. More broadly, these findings contribute to the global discourse on AI regulation, suggesting that effective regulation of LLMs hallucinations must transcend jurisdictional boundaries and rely on collaborative, real evidence policy development.

Chapter 6 - Conclusion

6.1 Summary of Key Findings

This thesis has explored in depth the phenomenon of hallucinations in LLMs, mapping out their lifecycle across model design, pretraining, fine-tuning, deployment, and monitoring stages, and analysing the regulatory approaches of both the EU and China. It shows that both the EU and China know hallucinations in LLMs can harm people and society. But their laws and how they enforce them are different in how they are created and applied.

At the theoretical level, both the EU and China seek to establish regulatory mechanisms that prioritise transparency, accountability, and risk mitigation. They understand that hallucinations are not just technical mistakes but also cause bigger problems including the spread of misinformation, erosion of public trust, and regulatory enforcement gaps. Both jurisdictions have rules that cover the whole AI process, from early development to use and monitoring after release.

However, their ways of making rules are very different. The EU uses a central and broad law called the AIA. This law sorts of AI systems by risk and sets strict rules based on that. China, on the other hand, uses a vertical, industry-focused approach. It relies on test programs and specific rules made for certain uses, like recommendation algorithms and Gen AI services. Each model presents distinct advantages and limitations. While the EU provides more comprehensive lifecycle coverage and stronger legal authority, it has not yet specifically addressed Gen AI or LLMs. In contrast, China's approach places greater emphasis on the monitoring and enforcement stages of the LLMs lifecycle. Its regulatory measures are more targeted and flexible but remain relatively fragmented and of lower legal status.

At the practical level, this thesis employs a tech and law method to comparatively evaluate the practical effectiveness of hallucination regulation in both jurisdictions. This framework enables a contextualised understanding of how each regulatory system performs in mitigating real-world hallucination risks and what lessons may be drawn for future refinement.

In practice, although the EU and China differ in terms of the institutional actors responsible for enforcing AI-related laws, both exhibit a pattern in which regulatory authorities and regulated entities actively collaborate to implement legal provisions. The regulation of hallucinations is also carried out through the enforcement of indirect legal norms.

On the micro level, this thesis identifies key risks including user cognitive misguidance, overreliance on model outputs, and a breakdown in feedback mechanisms necessary for system improvement. Both the EU and China try to reduce these risks by making rules for clear information, output labels, and human checks. But they have not done enough to deal with the problem of people relying too much on what the models say.

On the macro level, hallucinations can contribute to public information pollution, the erosion of trust in AI, and broader regulatory deficits. The EU's AIA attempts to pre-empt such harms by requiring risk assessments, data governance audits, and conformity procedures for high-risk systems. China acts by requiring algorithm registration and safety checks, but it does not yet have one clear law for all of this.

Furthermore, the thesis highlights structural differences. In the EU, the AI Office and national regulators work in a system that is spread out but still works together. In China, the government controls things more directly from the top. This leads to

different ways of enforcing rules, different levels of government choice, and how clear the rules are for companies. The EU has laws with clear rule types. In China, the rules often come after something happens and depend on how the government sees the situation.

Despite the systems are different, both face similar problems. The fast changes in LLMs are hard to keep up with, and it is difficult to fit new and changing risks into fixed legal rules. Also, both the EU and China deal with hallucinations in an indirect way. So far, neither has made a specific law that treats hallucinations as a separate issue, even though they cause special problems with truth and information.

Overall, the findings show that the EU uses detailed laws, and China takes a flexible and practical approach. But both agree that hallucinations are a serious problem that need action from many levels and different groups.

6.2 Policy and Legislative Recommendations

Based on the comparison, some important suggestions can help improve how the EU and China deal with hallucinations in LLMs. These ideas can also give useful lessons for other countries around the world.

To begin with, hallucinations should be treated as a separate harm of Gen AI, not just a type of AI risks. Because they affect how people think, spread misinformation, and cause social problems, hallucinations deserve focused attention in the same way that disinformation or automated decision-making risks are addressed. Both the EU and China should consider clearly defining hallucination-related harms in legal terms and including them in their risk classification systems.

Based on this, regulators should make rules that cover the whole life of LLMs. Instead of having separate or step-by-step rules, there should be one clear set of rules for all parts of LLM development and use. This includes checking training data, testing the model, reviewing how it works after release, and collecting feedback later. The AIA already follows this idea in how it is set up. In the same way, China's rules in different areas can also change to watch over the full life of LLMs.

In addition, the need for algorithm registries and transparency mechanisms has become increasingly apparent, which is useful for hallucination regulation. China's algorithm filing system provides a practical model of ex ante regulatory visibility, while the EU's conformity declarations contribute to ex post accountability. These could be synthesised into a transnational AI registry that consolidates risk levels, intended applications, performance indicators, and accountability channels, thus enhancing both public understanding and institutional oversight.

Moreover, a more contextualised approach to risk assessment is required. Because hallucination risks are highly use-dependent varying significantly between fields such as healthcare, finance, education, and entertainment. It is essential that legal frameworks allow for domain-specific regulation. Here, China's service-oriented categorisation provides a useful complement to the EU's product-based conformity assessments, suggesting a hybrid model could be developed.

Equally important is the emphasis on public education and user-side transparency. As LLMs become increasingly embedded in search engines, legal services, and educational tools, the public must be informed about the limitations of such systems. Labels like "AI-generated," "may contain mistakes," or "not checked for facts" should be easy to see. At the same time, public education can help people

better understand how large language models work and when they might give wrong or unclear answers.

Furthermore, rules should use different methods, like shared control and non-binding rules. In China, using testing areas for algorithms, setting technical rules, and trying out new programs shows that flexible and trial methods can help. The EU can also grow its use of voluntary rules and technical guides. This lets companies follow rules in a way that supports new ideas but still makes sure they act responsibly.

Finally, given the transnational deployment of LLMs, more effective international coordination is essential. Through multilateral fora such as the OECD and UNESCO,[204] the EU and China may play an important role in shaping emerging global standards for AI governance. By drawing on each other's regulatory experience, they may contribute to a more balanced and inclusive framework that combines China's adaptive and experimental governance style with the EU's more structured and legally systematised approach.

6.3 Future Research Directions

As rules for LLMs and other AI systems keep developing, there are still important topics that need more research. Future research should address not only the normative questions raised by LLMs hallucinations but also the ways institutions, technology, and society affect how these risks are managed.

[204] OECD, Recommendation of the Council on Artificial Intelligence (*OECD*, 22 May 2019) https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 accessed 17 July 2025; UNESCO, Recommendation on the Ethics of Artificial Intelligence (*UNESCO*, 23 November 2021) <https://unesdoc.unesco.org/ark:/48223/pf0000381137> accessed 17 July 2025.

To begin with, a clear taxonomy of LLMs hallucinations is needed. While current literature often distinguishes between intrinsic and extrinsic hallucinations, future research should investigate subcategories in relation to different contexts such as hallucinations in legal advice, medical diagnostics, or academic search. Understanding these domain-specific risks is crucial for developing tailored safeguards, especially as LLMs are increasingly embedded in high-stakes decision-making environments.

Another critical area lies in the assessment of regulatory effectiveness. Although both the EU and China have enacted or proposed regulatory instruments targeting the general AI harms, little is known about how these measures are implemented in practice and whether they succeed in mitigating hallucinations. Comparative studies on enforcement practices, audit regimes, and compliance behaviour across different jurisdictions would provide valuable insight into the real-world functioning of AI regulation.

In parallel, further interdisciplinary inquiry is needed to understand the socio-cognitive impact of interacting with hallucinating LLMs. Research in behavioural science, media studies, and psychology could enrich our understanding of how users perceive and respond to AI-generated misinformation, particularly in long-term usage scenarios. This would enable policymakers to design more effective feedback loops, trust calibration mechanisms, and user-facing warning systems.

Moreover, future studies should investigate the intersection between hallucination governance and global digital inequality. The development and regulation of LLMs are currently dominated by a small number of jurisdictions with significant technological and regulatory capacity. However, the deployment of such technologies often affects regions with weaker institutional oversight and

lower AI literacy. Looking at how AI causes harm in different countries and how each country can deal with it could help reduce the global gap in rules.

More legal research is needed to see if it is possible to go beyond hallucinations and bring out more legal questions about how to control AI. For example, how will open-source AI change the way we make rules?[205] Who should be responsible when Gen AI causes harm?[206] What laws are needed to manage the extra risks from AI-generated content?[207] And how should we deal with AI agents that act like they have their own role or status?[208] These are important questions that legal experts have just started to look into.

---

[205] Harry Law and Sébastien Krier, 'Open-Source Provisions for Large Models in the AI Act' (2023) 4 (1) Cambridge Journal of Science and Policy <https://doi.org/10.17863/CAM.100083> accessed 26 July 2025; Fatih Bildirici, 'Open-Source AI: An Approach to Responsible Artificial Intelligence Development' (2024) 5(1) Reflektif Sosyal Bilimler Dergisi 74; Elizabeth Seger and others, 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives' (arXiv, 29 September 2023) <https://doi.org/10.48550/arXiv.2311.09227> accessed 26 July 2025.

[206] Hilda Hadan and others, 'Who Is Responsible When AI Fails? Mapping Causes, Entities, and Consequences of AI Privacy and Ethical Incidents' (arXiv, 28 March 2025) <https://arxiv.org/abs/2504.01029> accessed 26 July 2025; Henry L. Fraser and Nicolas P. Suzor, 'Locating Fault for AI Harms: A Systems Theory of Foreseeability, Reasonable Care and Causal Responsibility in the AI Value Chain' (2025) 17(1) Law, Innovation and Technology 103.

[207] Bram Rijsbosch, Gijs van Dijck and Konrad Kollnig, 'Adoption of Watermarking Measures for AI-Generated Content and Implications under the EU AI Act' (arXiv, 23 March 2025) <https://doi.org/10.48550/arXiv.2503.18156> accessed 26 July 2025; Xiangwei He and Lijuan Fang, 'Regulatory Challenges in Synthetic Media Governance: Policy Frameworks for AI-Generated Content Across Image, Video, and Social Platforms' (2024) 9(12) Journal of Robotic Process Automation, AI Integration, and Workflow Optimization 36.

[208] Noam Kolt, 'Governing AI Agents' (arXiv, 14 January 2025) <https://doi.org/10.48550/arXiv.2501.07913> accessed 26 July 2025; Deven R Desai and Mark O Riedl, 'Responsible AI Agents' (arXiv, 25 February 2025) <https://doi.org/10.48550/arXiv.2502.18359> accessed 26 July 2025.

Finally, working together on technical and legal design is still a new but important area. Future research should look at how legal rules can be added into the way LLMs are built. For example, models could be asked to show how they make choices, explain where their answers come from, or be checked for mistakes like hallucinations. This work will need close teamwork between legal experts, engineers, and people who study ethics.

In sum, as LLMs play a bigger role in shaping online information, future research should focus on both the harm they might cause and the challenges in making rules for them. Using different methods, working across fields, and knowing how local laws work will be significant to make sure rules can keep up with new technology.

Bibliography

Books

Broussard M, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press 2018)

Calo R, *Law and Technology: A Methodical Approach* (OUP 2025)

Gans J, *The Disruption Dilemma* (MIT Press 2016)

Macpherson F and Platchias D (eds), *Hallucination: Philosophy and Psychology* (MIT Press 2013)

Turner J, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019)

Book Sections

Ard BJ and Crootof R, 'Legal Responses to Techlaw Uncertainties' in Bartosz Brożek, Olia Kanevskaia and Przemysław Pałka (eds), *Research Handbook on Law and Technology* (Edward Elgar 2023)

Belloir N, Ouerdane W and Pastor O, 'Characterizing Fake News: A Conceptual Modeling-based Approach' in Jolita Ralyté and others (eds), *Conceptual Modeling: 41st International Conference Proceedings* (Springer 2022)

Dempsey M and others, ‘Transnational Digital Governance and Its Impact on Artificial Intelligence’ in Justin B Bullock and others (eds), *The Oxford Handbook of AI Governance* (OUP 2024)

Hacker P and others, ‘Generative Discrimination’ in Philipp Hacker and others (eds), *The Oxford Handbook of the Foundations and Regulation of Generative AI* (OUP 2025)

Maynez J and others, ‘On Faithfulness and Factuality in Abstractive Summarization’ in Dan Jurafsky and others (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020) 1906

Owen R, Macnaghten P and Stilgoe J, ‘Responsible Research and Innovation: From Science in Society to Science for Society, with Society’ in Gary E Marchant and Wendell Wallach (eds), *Emerging Technologies: Ethics, Law and Governance* (Routledge 2020)

Roberts H and others, ‘The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation’ in Luciano Floridi (ed), *Ethics, Governance, and Policies in Artificial Intelligence* (Springer 2021)

Smits JM, ‘What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research’ in Rob van Gestel and others (eds), *Rethinking Legal Scholarship* (CUP 2017)

Yeung K, Howes A and Pogrebna G, ‘AI Governance by Human Rights-Centered Design, Deliberation, and Oversight’ in Markus D Dubber and others (eds), *The Oxford Handbook of Ethics of AI* (OUP 2020)

Journal Articles

Albrecht D, 'Chinese First Personal Information Protection Law in Contrast to the European GDPR' (2022) 23(1) Computer Law Review International 1

Alkaissi H and McFarlane SI, 'Artificial Hallucinations in ChatGPT: Implications in Scientific Writing' (2023) 15(2) Cureus e35179

Bareis J and Katzenbach C, 'Talking AI into Being: The Narratives and Imaginaries of National AI Strategies' (2022) 47(5) Science, Technology, & Human Values 855

Belk R, 'Ethical Issues in Service Robotics and Artificial Intelligence' (2021) 41(13—14) The Service Industries Journal 860

Bengio Y and others, 'Managing Extreme AI Risks amid Rapid Progress' (2024) 384(6698) Science 842

Bildirici F, 'Open-Source AI: An Approach to Responsible Artificial Intelligence Development' (2024) 5(1) Reflektif Sosyal Bilimler Dergisi 74

Birhane A and others, 'Science in the Age of Large Language Models' (2023) 5(5) Nature Reviews Physics 277

Bi W , 'The Dilemma in the Risk Regulation of Generative Artificial Intelligence and Its Resolution: A Perspective on the Regulation of ChatGPT' (生成式人工智能

的风险规制困境及其化解：以 ChatGPT 的规制为视角) (2023) 3 Journal of Comparative Law 155

Bradford A, ‘The Brussels Effect’ (2015) 107 Northwestern University Law Review 1

Breggin PR, ‘Understanding and Helping People with Hallucinations’ (2015) 43(1) The Humanistic Psychologist 70

Buocz T, Pfotenhauer S and Eisenberger I, ‘Regulatory Sandboxes in the AI Act’ (2023) 15(2) Law, Innovation and Technology 357

Calo R, ‘Artificial Intelligence Policy: A Primer and Roadmap’ (2017) 51 UC Davis Law Review 399

Calzada I, ‘Citizens’ Data Privacy in China: The State of the Art of the PIPL’ (2022) 5(3) Smart Cities 1129

Cao B and others, ‘The Life Cycle of Knowledge in Big Language Models: A Survey’ (2024) 21(2) Machine Intelligence Research 217

Cath C and others, ‘Artificial Intelligence and the “Good Society”: the US, EU, and UK Approach’ (2018) 24(2) Science and Engineering Ethics 505

Crain M, ‘The Limits of Transparency: Data Brokers and Commodification’ (2018) 20(1) New Media & Society 88

Demszky D and others, ‘Using Large Language Models in Psychology’ (2023) 2(11) Nature Reviews Psychology 688

Dong H and Chen J, ‘Meta-Regulation: An Ideal Alternative to the Primary Responsibility as the Regulatory Model of Generative AI in China’ (2024) 54 Computer Law & Security Review 106016

Eberle EJ, ‘The Methodology of Comparative Law’ (2011) 16(1) Roger Williams University Law Review 51

Edelman LB and Suchman MC, ‘The Legal Environments of Organizations’ (1997) 23 Annual Review of Sociology 479

Essen L von and Ossewaarde M, ‘Artificial Intelligence and European Identity’ (2024) 25(2) European Politics and Society 375

Farquhar S and others, ‘Detecting Hallucinations in Large Language Models Using Semantic Entropy’ (2024) 630 Nature 625

Fernandes P and others, ‘Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation’ (2023) 11 Transactions of the Association for Computational Linguistics 1643

Ferrara E, ‘Should ChatGPT Be Biased? Challenges and Risks of Bias in LLMs’ (2023) 28(11) First Monday

Filgueiras F, ‘The Politics of AI: Democracy and Authoritarianism in Developing Countries’ (2022) 19(4) Journal of Information Technology & Politics 449

Floridi L and others, ‘AI4People—An Ethical Framework for a Good AI Society’ (2018) 28(4) Minds and Machines 689

Fraser HL and Suzor NP, ‘Locating Fault for AI Harms’ (2025) 17(1) Law, Innovation and Technology 103

Fung P and Etienne H, ‘Confucius, Cyberpunk and Mr. Science: Comparing AI Ethics Principles between China and the EU’ (2023) 3(2) AI and Ethics 505

Gao Z, ‘Response and Transcendence: Legal Regulation of Generative Artificial Intelligence’ (回应与超越：生成式人工智能法律规制——以《生成式人工智能服务管理暂行办法》为视角) (2024) 5 Social Sciences Journal 121

Greenleaf G, ‘China’s Completed Personal Information Protection Law’ (2021) 172 Privacy Laws & Business International Report 20

Guerreiro NM and others, ‘Hallucinations in Large Multilingual Translation Models’ (2023) 11 Transactions of the Association for Computational Linguistics 1500

Guo D, ‘Technology Governance with Chinese Characteristics’ (2023) 47(3) Computer Law Review International

Hacker P, Cordes J and Rochon J, ‘Regulating Gatekeeper Artificial Intelligence and Data’ (2024) 15(1) European Journal of Risk Regulation 49

Hadid A, Chakraborty T and Busby D, ‘When Geoscience Meets Generative AI and LLMs’ (2024) 41(10) Expert Systems e13654

Hasan M, ‘Regulating Artificial Intelligence: A Study in the Comparison between South Asia and Other Countries’ (2024) 5(1) Legal Issues in the Digital Age 122

Haupt CE and Marks M, ‘AI-Generated Medical Advice—GPT and Beyond’ (2023) 329(16) JAMA 1349

He X and Fang L, ‘Regulatory Challenges in Synthetic Media Governance’ (2024) 9(12) Journal of Robotic Process Automation 36

Henderson P, Hashimoto T and Lemley MA, ‘Where’s the Liability in Harmful AI Speech?’ (2023) 3 Journal of Free Speech Law 589

Hicks MT, Humphries J and Slater J, ‘ChatGPT is bullshit’ (2024) 26 Ethics and Information Technology

Hine E and Floridi L, ‘Artificial Intelligence with American Values and Chinese Characteristics’ (2024) 39(1) AI & Society 257

Hofstede G , ‘Dimensionalizing Cultures: The Hofstede Model in Context’ (2011) 2(1) Online Readings in Psychology and Culture

Huang L and others, ‘A Survey on Hallucination in Large Language Models’ (2025) 43(2) ACM Transactions on Information Systems

Huang R and Yao H , ‘“Reshaping” and ““High Risk”: The Impact of Generative Artificial Intelligence on Public Opinion Security’ (“再塑造”与“高风险”: 生成式人工智能对舆论安全的影响) (2024) 43(4) Journal of Intelligence 121

Hutchinson T and Duncan N, ‘Defining and Describing What We Do: Doctrinal Legal Research’ (2012) 17(1) Deakin Law Review 83

Hutchinson T, ‘The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law’ (2015) 8(3) Erasmus Law Review 130

Ji Z and others, ‘Survey of Hallucination in Natural Language Generation’ (2023) 55(12) ACM Computing Surveys art 248

Justo-Hanani R, ‘The Politics of Artificial Intelligence Regulation and Governance Reform in the European Union’ (2022) 55(1) Policy Sciences 137

Keith AJ, ‘Governance of Artificial Intelligence in Southeast Asia’ (2024) 15(5) Global Policy 937

Krarup T and Horst M, ‘European Artificial Intelligence Policy as Digital Single Market Making’ (2023) 10(1) Big Data & Society 205395172311538

Krkić M, ‘Cultural Perspectives on AI Usage and Regulation in Deepfake Creation’ (2025) 12 International Communication of Chinese Culture 225

Kuner C, ‘The European Union and the Search for an International Data Protection Framework’ (2014) 2(2) Groningen Journal of International Law 55

Laux J, Wachter S and Mittelstadt B, ‘Trustworthy Artificial Intelligence and the European Union AI Act’ (2024) 18(1) Regulation & Governance 3

Lee M, ‘A Mathematical Investigation of Hallucination and Creativity in GPT Models’ (2023) 11(10) Mathematics 2320

Li W and Chen J, ‘From Brussels Effect to Gravity Assists: Understanding the Evolution of the PIPL in China’ (2024) 54 Computer Law & Security Review 105994

Long Z, Parnell JA and Dent EB, ‘Individualism, Collectivism and Management in China’ (2019) 20(3) Journal of Asia-Pacific Business 166

McIntosh TR and others, ‘A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination’ (2024) 5(6) IEEE Transactions on Artificial Intelligence 2739

Mertha A, ‘“Fragmented Authoritarianism 2.0”: Political Pluralization in the Chinese Policy Process’ (2009) 200 The China Quarterly 995

Munn L, Magee L and Arora V, ‘Truth Machines: Synthesizing Veracity in AI Language Models’ (2024) 39 AI & Society 2759

Novelli C and others, ‘Generative AI in EU Law’ (2024) 55 Computer Law & Security Review 106066

Novelli C and others, ‘Taking AI Risks Seriously: A New Assessment Model for the AI Act’ (2024) 39(5) AI & Society 2493

Pan S and others, ‘Unifying Large Language Models and Knowledge Graphs: A Roadmap’ (2024) 36(7) IEEE Transactions on Knowledge and Data Engineering 3580

Park M, Wu S and Funk RJ, ‘Regulation and Innovation Revisited’ (2025) 36(1) Organization Science 240

Polyportis A and Pahos N, ‘Navigating the Perils of Artificial Intelligence: A Focused Review on ChatGPT and RRI’ (2024) 11 Humanities and Social Sciences Communications 107

Ray PP, ‘Web3: A Comprehensive Review on Background, Technologies, Applications, Zero-Trust Architectures, Challenges and Future Directions’ (2023) 3 Internet of Things and Cyber-Physical Systems 213

Roberts H and others, ‘Achieving a “Good AI Society”: Comparing the Aims and Progress of the EU and the US’ (2021) 27(6) Science and Engineering Ethics

Roberts H and others, ‘Governing Artificial Intelligence in China and the European Union’ (2023) 39(2) The Information Society 79

Robinson SC, ‘Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for AI’ (2020) 63 Technology in Society 101421

Ryan M and Stahl BC, ‘Artificial Intelligence Ethics Guidelines for Developers and Users’ (2021) 19(1) Journal of Information, Communication and Ethics in Society 61

Sallam M, ‘ChatGPT Utility in Healthcare Education, Research, and Practice’ (2023) 11(6) Healthcare 887

Schuett J, ‘Risk Management in the Artificial Intelligence Act’ (2024) 15(2) European Journal of Risk Regulation 367

Selaković M, ‘Fake News and Foreign Direct Investment Inflows’ (2022) 14(2) European Journal of Interdisciplinary Studies 24

Silva D De and Alahakoon D, ‘An Artificial Intelligence Life Cycle: From Conception to Production’ (2022) 3(6) Patterns 100489

Simões JM and Caldeira W, ‘Ethics concerns in the use of computer-generated images for human communication’ (2024) 4 Journal of Ethics in Higher Education 169

Smuha NA, ‘From a “Race to AI” to a “Race to AI Regulation”’ (2021) 13 Law, Innovation and Technology 57

Stuurman C and Lachaud E, ‘Regulating AI: A Label to Complete the Newly Proposed Act on Artificial Intelligence’ (2022) 44 Computer Law & Security Review 105657

Tandoc Jr EC, Lim ZW and Ling R, ‘Defining “Fake News”: A Typology of Scholarly Definitions’ (2018) 6(2) Digital Journalism 137

Terzi S and Stamelos I, ‘Architectural Solutions for Improving Transparency, Data Quality, and Security in eHealth Systems’ (2024) 14 Health and Technology 451

Tuzov V and Lin F, ‘Two Paths of Balancing Technology and Ethics: A Comparative Study on AI Governance in China and Germany’ (2024) 48(10) Telecommunications Policy 102850

Vaishya R, Misra A and Vaish A, ‘ChatGPT: Is This Version Good for Healthcare and Research?’ (2023) 17 Diabetes & Metabolic Syndrome 102744

Veale M and Borgesius FZ, ‘Demystifying the Draft EU Artificial Intelligence Act’ (2021) 22(4) Computer Law Review International 97

Velázquez J De Miguel and others, ‘Decoding Real-World Artificial Intelligence Incidents’ (2024) 57(11) Computer 71

Wachter S and Mittelstadt B, ‘A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI’ (2019) 2019(2) Columbia Business Law Review 494

Wachter S, Mittelstadt B and Russell C, ‘Do large language models have a legal duty to tell the truth?’ (2024) 11(8) Royal Society Open Science 240197

Wachter S, ‘The Social Contract for AI: Protecting the Individual in the Algorithmic Age’ (2021) 20(2) Human Rights Law Review 53

Wachter S, ‘The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law’ (2022) 97(2) Tulane Law Review 149

Wachter S, Mittelstadt B and Russell C, ‘Why Fairness Cannot Be Automated’ (2021) 41 Computer Law & Security Review 105567

Wang S and others, ‘Artificial Intelligence Policy Frameworks in China, the European Union and the United States’ (2025) 212 Technological Forecasting and Social Change 123971

Wu J, ‘The Choice of Legislative Regulatory Model of Generative Artificial Intelligence in China’ (中国生成式人工智能立法监管模式选择) (2024) 4 China Legal Science 133

Yang Q and others, ‘Dual Retrieving and Ranking Medical Large Language Model with Retrieval Augmented Generation’ (2025) 15(1) Scientific Reports 18062

Zhang AH, ‘The Promise and Perils of China’s Regulation of Artificial Intelligence’ (2025) 63 Columbia Journal of Transnational Law 1

Zhang X, ‘Legislative Innovation in the Era of Algorithmic Governance’ (2023) 41(2) Law and Social Development 33

Zhang Y and others, ‘Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models’ (2025) 51(4) Computational Linguistics 1373

Zhao H and others, ‘Explainability for Large Language Models: A Survey’ (2024) 15(2) ACM Transactions on Intelligent Systems and Technology

Zhu D and Amiya B, ‘The Impact of Artificial Intelligence on Business Strategy: A Review of Theoretical and Empirical Studies in China’ (2025) 2(3) International Journal of Advances in Business and Management Research 9

Zhu X and Zhao H, ‘Experimentalist Governance with Interactive Central–Local Relations’ (2021) 49(1) Policy Studies Journal 13

Conference Papers

AboulEla S and others, ‘Exploring RAG Solutions to Reduce Hallucinations in LLMs’ in 2025 IEEE International Systems Conference (SysCon) (IEEE 2025) 1

Akyürek E and others, ‘Towards Tracing Knowledge in Language Models Back to the Training Data’ in Findings of the Association for Computational Linguistics: EMNLP 2022 (ACL 2022)

Alfiani FRN and Santiago F, ‘A Comparative Analysis of Artificial Intelligence Regulatory Law in Asia, Europe, and America’ in SHS Web of Conferences vol 204 (EDP Sciences 2024) 07006

Bang Y and others, ‘A Multitask, Multilingual, Multimodal Evaluation of ChatGPT’ in Proceedings of IJCNLP-AACL 2023 (Volume 1: Long Papers) (ACL 2023) 675

Bender EM and others, ‘On the Dangers of Stochastic Parrots’ in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM 2021)

Binns R, ‘Fairness in Machine Learning: Lessons from Political Philosophy’ in Proceedings of the 1st Conference on Fairness, Accountability and Transparency (2018)

Bird C and others, ‘Typology of Risks of Generative Text-to-Image Models’ in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (ACM 2023)

Carlini N and others, ‘Extracting Training Data from Large Language Models’ in 30th USENIX Security Symposium (USENIX Association 2021) 2633

Chen J and others, 'Benchmarking Large Language Models in Retrieval-Augmented Generation' (2024) 38(16) Proceedings of the AAAI Conference on Artificial Intelligence 17754

Cheng X and others, ‘Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking’ in Findings of the Association for Computational Linguistics: ACL 2025 (ACL 2025)

Christiano PF and others, ‘Deep Reinforcement Learning from Human Preferences’ in Advances in Neural Information Processing Systems 30 (2017)

Dhamala J and others, ‘BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation’ in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM 2021) 862

Dziri N and others, ‘Neural Path Hunter’ in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (ACL 2021) 2197

Filippova K, ‘Controlled Hallucinations: Learning to generate faithfully from noisy data’ in Findings of the Association for Computational Linguistics: EMNLP 2020 (ACL 2020) 864

Gekhman Z and others, ‘Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?’ in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (ACL 2024)

Goanta C and others, ‘Regulation and NLP (RegNLP): Taming Large Language Models’ in Houda Bouamor and others (eds), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (ACL 2023) 8715

Goodrich B and others, ‘Assessing the Factual Accuracy of Generated Text’ in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (ACM 2019)

Hajian S, Bonchi F and Castillo C, ‘Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining’ in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM 2016)

Hutiri W, Papakyriakopoulos O and Xiang A, ‘Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators’ in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (ACM 2024)

Jin S and others, ‘Reasoning Grasping via Multimodal Large Language Model’ in Proceedings of the 8th Conference on Robot Learning (PMLR 2025)

Lewis P and others, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks' in Advances in Neural Information Processing Systems 33 (2020) 9459

Li M and others, ‘A Closer Look at the Existing Risks of Generative AI’ (2025) 8(2) Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 1561

Luna J and others, ‘Navigating Governance Paradigms’ (2024) 7(1) in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 917

Manakul P, Liusie A and Gales MJF, ‘SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models’ in Houda Bouamor and others (eds), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (ACL 2023)

Mascarell L, Chalumattu R and Rios A, ‘German Also Hallucinates! Inconsistency Detection in News Summaries with the Absinth Dataset’ in Proceedings of LREC-COLING 2024 (ELRA and ICCL 2024)

Mitchell M and others, ‘Model Cards for Model Reporting’ in Proceedings of the Conference on Fairness, Accountability, and Transparency (ACM 2019)

Ouyang L and others, ‘Training Language Models to Follow Instructions with Human Feedback’ in Advances in Neural Information Processing Systems 35 (2022) 27730

Qiu Y and others, ‘Detecting and Mitigating Hallucinations in Multilingual Summarisation’ in Houda Bouamor and others (eds), Proceedings of the 2023

Conference on Empirical Methods in Natural Language Processing (ACL 2023) 8940

Qiu Y and others, ‘Think While You Write: Hypothesis Verification Promotes Faithful Knowledge-to-Text Generation’ in Findings of the Association for Computational Linguistics: NAACL 2024 (ACL 2024)

Rawte V and others, ‘The Troubling Emergence of Hallucination in Large Language Models’ in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (ACL 2023)

Shen Y and others, ‘Alleviating LLM-Based Generative Retrieval Hallucination in Alipay Search’ in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM 2025)

Su WH and others, ‘Mitigating Entity-Level Hallucination in Large Language Models’ in Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (ACM 2024) 23.

Wan Z and others, ‘Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain’ in Findings of the Association for Computational Linguistics: ACL 2024 (ACL 2024)

Zhang S and others, ‘The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models’ in Findings of the Association for Computational Linguistics: ACL 2024 (ACL 2024)

Zhang Y and others, ‘Alleviating Hallucinations of Large Language Models through Induced Hallucinations’ in Findings of the Association for Computational Linguistics: NAACL 2025 (ACL 2025)

Zhao Z and others, ‘Calibrate before Use: Improving Few-Shot Performance of Language Models’ in Proceedings of the 38th International Conference on Machine Learning (PMLR 2021) 12697

Zhou C and others, ‘Detecting Hallucinated Content in Conditional Neural Sequence Generation’ in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (ACL 2021) 1393

Online Journals

Ainsworth E, Wycliffe J and Winslow F, ‘Reducing Contextual Hallucinations in Large Language Models through Attention Map Optimization’ (TechRxiv, 23 July 2024) <https://www.techrxiv.org/doi/full/10.36227/techrxiv.172166617.52554248/v1> accessed 20 July 2025

Berberette E, Hutchins J and Sadovnik A, ‘Redefining “Hallucination” in LLMs: Towards a Psychology-Informed Framework for Mitigating Misinformation’ (arXiv, 1 February 2024) <https://arxiv.org/abs/2402.01015> accessed 25 July 2025

Capellini R, Atienza F and Sconfield M, ‘Knowledge Accuracy and Reducing Hallucinations in LLMs via Dynamic Domain Knowledge Injection’ (Research Square, 2024) <https://doi.org/10.21203/rs.3.rs-4456950/v1> accessed 25 July 2025

Chen J, Huang X and Li Y, ‘Dynamic Supplementation of Federated Search Results for Reducing Hallucinations in LLMs’ (OSF Preprints, 2024) <https://osf.io/preprints/osf/s3z7h> accessed 25 July 2025

Cheng Q and others, ‘Evaluating Hallucinations in Chinese Large Language Models’ (arXiv, 25 October 2023) <https://arxiv.org/abs/2310.03368> accessed 25 July 2025

Chun J, de Witt CS and Elkins K, ‘Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US’ (arXiv, 5 October 2024) <https://arxiv.org/abs/2410.03350> accessed 26 July 2025

Cossio M, ‘A Comprehensive Taxonomy of Hallucinations in Large Language Models’ (arXiv, 3 August 2025) <https://arxiv.org/abs/2508.01234> accessed 15 August 2025

DeepSeek-AI and others, ‘DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model’ (arXiv, 7 May 2024) <https://arxiv.org/abs/2405.04434> accessed 20 July 2025

Desai DR and Riedl MO, ‘Responsible AI Agents’ (arXiv, 25 February 2025) <https://arxiv.org/abs/2502.18524> accessed 26 July 2025

Floridi L and others, ‘capAI: A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU AI Act’ (SSRN, 23 March 2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064959> accessed 17 June 2025

Guldimann P and others, ‘COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU AI Act’ (arXiv, 10 October 2024) <https://arxiv.org/abs/2410.07959> accessed 25 July 2025

Hacker P, ‘AI Regulation in Europe: From the AI Act to Future Regulatory Challenges’ (arXiv, 6 October 2023) <https://arxiv.org/abs/2310.04417> accessed 17 June 2025

Hadan H and others, ‘Who Is Responsible When AI Fails? Exploring the Legal and Ethical Perspectives on AI Liability’ (arXiv, 28 March 2025) <https://arxiv.org/abs/2503.20050> accessed 26 July 2025

Ji Z and others, ‘Towards Mitigating Hallucination in Large Language Models via Self-Reflection’ (arXiv, 10 October 2023) <https://arxiv.org/abs/2310.06271> accessed 25 July 2025

Kolt N, ‘Governing AI Agents’ (arXiv, 14 January 2025) <https://arxiv.org/abs/2501.07762> accessed 26 July 2025

Mialon G and others, ‘Augmented Language Models: A Survey’ (arXiv, 15 February 2023) <https://arxiv.org/abs/2302.07842> accessed 25 July 2025

Penedo G and others, ‘The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Removing Errors’ (arXiv, 1 June 2023) <https://arxiv.org/abs/2306.01116> accessed 20 July 2025

Sajduk B and Dziwisz D, ‘Comparative Analysis of AI Development Strategies: A Study of China’s Ambitions and the EU’s Regulatory Framework’ (EuroHub4Sino Policy Paper 2024/12, 20 September 2024) <https://storage.eh4s.eu/vitrin/files%2FComparative-Analysis-of-AI-Development-Strategies.pdf> accessed 26 July 2025

Schuett J, ‘A Blueprint for Managing AI Risks under the EU AI Act’ (arXiv, 27 March 2024) <https://arxiv.org/abs/2403.18525> accessed 25 July 2025

Shi W and others, ‘Trustworthy LLMs: A Survey and Guideline for Evaluation’ (arXiv, 30 January 2024) <https://arxiv.org/abs/2401.15391> accessed 25 July 2025

Tonge A and others, ‘A Cross-Cultural Perspective on LLM Hallucinations’ (arXiv, 23 April 2025) <https://arxiv.org/abs/2504.14321> accessed 26 July 2025

Zhang Y and others, ‘Trustworthy Large Language Models: A Critical Analysis of Hallucination, Privacy and Bias’ (arXiv, 18 December 2024) <https://arxiv.org/abs/2412.14020> accessed 26 July 2025

Websites, Blogs & News

AIbase, ‘Taobao Launches Platform-Wide AI-Generated Fake Image Governance’ (*AIbase*, 27 March 2025) <https://www.aibase.com/news/16663> accessed 1 July 2025

Akewushola N, ‘TikTok Begins Auto-Labelling of AI-Generated Content’ (*The FactCheckHub*, 15 May 2024) <https://factcheckhub.com/tiktok-begins-auto-labelling-of-ai-generated-content> accessed 10 June 2025

Alibaba Cloud, ‘AIGC and Forgery Detection Service of Image Moderation 2.0’ (*Alibaba Cloud*, 24 November 2025) <https://www.alibabacloud.com/help/en/content-moderation/latest/image-audit-enhanced-edition-detects-aigc-infringement> accessed 1 December 2025

Angwin J, Nelson A and Palta R, ‘Seeking Reliable Election Information? Don’t Trust AI’ (*Proof News*, 27 February 2024) <https://www.proofnews.org/seeking-election-information-dont-trust-ai/> accessed 15 January 2025.

Anthropic, ‘Claude is providing incorrect or misleading responses — what’s going on?’ (*Anthropic*, 2024) <https://support.anthropic.com/en/articles/8525154-claude-is-providing-incorrect-or-misleading-responses-what-s-going-on> accessed 10 June 2025

Anthropic, ‘The Claude 3 Model Family’ (*Anthropic*, 4 March 2024) <https://www.anthropic.com/news/claude-3-family> accessed 10 June 2025

Axios, ‘Chinese AI Censorship Targets “Incorrect” Speech’ (*Axios*, 10 August 2023) <https://www.axios.com/2023/08/10/china-ai-censorship-socialist-values> accessed 17 June 2025

Baidu Inc, ‘Baidu Launches ERNIE 4.5 Turbo, ERNIE X1 Turbo and New Suite of AI Tools to Empower Developers and Supercharge AI Innovation’ (*PR Newswire*, 25 April 2025) <https://www.prnewswire.com/news-releases/baidu-launches-ernie-

4-5-turbo-ernie-x1-turbo-and-new-suite-of-ai-tools-to-empower-developers-and-supercharge-ai-innovation-302438584.html> accessed 14 July 2025

Bartz D and Hu K, 'OpenAI, Google, others pledge to watermark AI content for safety, White House says' (*Reuters*, 22 July 2023) <https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/> accessed 10 June 2025

Bytefeed, 'China Ramps Up Plans for Mandatory AI Watermarks Amid Rising Concerns' (*Bytefeed*, 27 September 2024) <https://bytefeed.ai/technology/china-ramps-up-plans-for-mandatory-ai-watermarks-amid-rising-concerns/> accessed 1 July 2025

Castro D, 'China's Annual Parliamentary Meeting Shows National Commitment to Advancing AI' (*Center for Data Innovation*, 18 March 2024) <https://datainnovation.org/2024/03/chinas-annual-parliamentary-meeting-shows-national-commitment-to-advancing-ai/> accessed 23 September 2024

Central Cyberspace Affairs Commission, 'Initiative for Global AI Governance' ('全球人工智能倡议') (*Cyberspace Administration of China*, 18 October 2023) <https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm> accessed 7 February 2024

China Money Network, 'Chinese Tech Giants Dominate AI Algorithms with a Focus on Industry-Specific Applications' (*China Money Network*, March 2024) <https://www.chinamoneynetwork.com/2024/03/07/chinese-tech-giants-dominate-ai-algorithms-with-a-focus-on-industry-specific-applications> accessed 2 May 2024

ChinaTalk, ‘Hugging Face Blocked! “Self-Castrating” China’s ML Development + Jordan at APEC’ (*ChinaTalk*, October 2023) <https://www.chinatalk.media/p/hugging-face-blocked-self-castrating> accessed 2 May 2024

Deepseek-ai, ‘[Hallucination Report] Model hallucinates biological validity in psychiatric analogies’ (*Hugging Face discussion*, 11 January 2026) <https://huggingface.co/deepseek-ai/DeepSeek-R1/discussions/236> accessed 28 March 2026

Edwards B, ‘New Meta AI Demo Writes Racist and Inaccurate Scientific Literature, Gets Pulled’ (*Ars Technica*, 18 November 2022) <https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/> accessed 20 November 2024

Edwards L, ‘Regulating AI in Europe: Four Problems and Four Solutions’ (*Ada Lovelace Institute*, 31 March 2022) <https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/> accessed 26 March 2026

Edwards L, ‘The EU AI Act: A Summary of Its Significance and Scope’ (*Ada Lovelace Institute*, 11 April 2022) <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf> accessed 26 March 2026

European Commission, ‘AI Pact’ (*European Commission*, updated 19 June 2025) <https://digital-strategy.ec.europa.eu/en/policies/ai-pact> accessed 20 June 2025

European Commission, ‘Over a Hundred Companies Sign EU AI Pact Pledges’ (*European Commission*, 25 September 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4864> accessed 20 June 2025

European DIGITAL SME Alliance, ‘New EU Initiatives to Ease the Transition of AI Act Regulatory Compliance for SMEs’ (*European DIGITAL SME Alliance*, 15 July 2025) <https://www.digitalsme.eu/new-eu-initiatives-to-ease-the-transition-of-ai-act-regulatory-compliance-for-smes/> accessed 15 July 2025

Fried I, ‘Exclusive: Trust in AI Is Much Higher in China than in the U.S.’ (*Axios*, 13 February 2025) <https://www.axios.com/2025/02/13/trust-ai-china-us> accessed 10 October 2025

Giovine C and others, ‘Building AI Trust: The Key Role of Explainability’ (*McKinsey*, 26 November 2024) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability> accessed 17 July 2025

Hancock E, ‘Ireland’s Privacy Watchdog Probes Musk’s Grok AI Model’ (*The Wall Street Journal*, 11 April 2025) <https://www.wsj.com/tech/irelands-privacy-watchdog-probes-musks-grok-ai-model-4779ba4e> accessed 14 July 2025

Information Commissioner’s Office, ‘Guidance on AI and Data Protection’ (*Information Commissioner’s Office*, 15 March 2023) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/> accessed 26 March 2026

Ji, XT, ‘Prof. Tiejun Huang from the School of Computer Science at Peking University: We Should Not Simply “Kill” AI Hallucinations with One Stick’ (‘北京大学计算机学院教授黄铁军：不能简单地将 AI 幻觉“一棒子打死”’) (China Electronics News, 7 May 2024)

Jiang B, ‘Chinese Online Search Giant Baidu’s Ernie Bot Joins iFlytek’s Spark in Apple’s Mainland App Store for Local AI’ (South China Morning Post, 5 July 2023) <https://www.scmp.com/tech/big-tech/article/3226659/chinese-online-search-giant-baidus-ernie-bot-joins-iflyteks-spark-apples-mainland-app-store-local-ai> accessed 14 July 2025

Kroet C, ‘Meta’s AI Labelling “Inconsistent”, Internal Oversight Board Finds’ (*Euronews Next*, 25 June 2025) <https://www.euronews.com/next/2025/06/25/metas-ai-labelling-inconsistent-internal-oversight-board-finds> accessed 30 June 2025

Kubacka T, ‘Today I Asked ChatGPT about the Topic I Wrote My PhD About’ (*Lookalikes and Meanders*, 6 December 2022) <https://lookalikes.substack.com/p/today-i-asked-chatgpt-about-the-topic> accessed 13 March 2025

Lan X, ‘DeepSeek’s Fabrication Is Flooding the Chinese Internet’ (‘DeepSeek 的胡编乱造，正在淹没中文互联网’) (*Sina Technology*, 6 March 2025) <https://finance.sina.com.cn/tech/roll/2025-03-06/doc-inensrzp6931316.shtml> accessed 10 March 2025

Maruf R, ‘Lawyer Apologizes for Fake Court Citations from ChatGPT’ (*CNN Business*, 28 May 2023) <https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers> accessed 20 November 2024

National Public Service Platform for Standards Information, ‘TC28/SC42 National Information Technology Standardization Technical Committee Artificial Intelligence Subcommittee’ (‘TC28/SC42 全国信息技术标准化技术委员会人工智能分技术委员会’) (*National Public Service Platform for Standards Information, 2026*) <https://std.samr.gov.cn/search/orgDetailView?tcCode=TC28SC42> accessed 28 March 2026

NOYB, ‘ChatGPT Provides False Information about People, and OpenAI Can’t Correct It’ (*NOYB*, 29 April 2024) <https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it> accessed 27 June 2025

OpenAI, ‘Understanding the Source of What We See and Hear Online’ (*OpenAI*, 7 May 2024) <https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online> accessed 10 June 2025

Qwen, ‘Qwen is loosing broad knowledge since Qwen2’ (*Hugging Face discussion*, 29 April 2025) <https://huggingface.co/Qwen/Qwen3-235B-A22B/discussions/16> accessed 28 March 2026

Pelkmans J and Andrea R, ‘Does EU Regulation Hinder or Stimulate Innovation?’ (*CEPS Special Report*, November 2014) <https://cdn.ceps.eu/wp-content/uploads/2015/01/No%2096%20EU%20Legislation%20and%20Innovation.pdf> accessed 17 October 2024

PYMNTS, ‘AI Regulations: OpenAI Calls on EU to Review, Simplify AI Rules’ (*PYMNTS*, 17 April 2025) <https://www.pymnts.com/artificial-intelligence-2/2025/ai-regulations-openai-calls-on-eu-to-review-simplify-ai-rules/> accessed 20 June 2025

Reuters, ‘Spain to Impose Massive Fines for Not Labelling AI-Generated Content’ (*Reuters*, 11 March 2025) <https://www.reuters.com/technology/artificial-intelligence/spain-impose-massive-fines-not-labelling-ai-generated-content-2025-03-11> accessed 10 June 2025

Schuerger C, Vikram V and Katherine Q, ‘China and Medical AI: Implications of Big Biodata for the Bioeconomy’ (*Center for Security and Emerging Technology*, May 2024) <https://cset.georgetown.edu/publication/china-and-medical-ai/> accessed 26 July 2025

Tazbaz T and John N, ‘Blog: A Lifecycle Management Approach toward Delivering Safe, Effective AI-Enabled Health Care’ (*U.S. Food and Drug Administration*, 25 July 2024) <https://www.fda.gov/medical-devices/digital-health-center-excellence/blog-lifecycle-management-approach-toward-delivering-safe-effective-ai-enabled-health-care> accessed 31 March 2025

The General Office of the State Council of the People’s Republic of China, ‘The Inaugural Meeting of the Artificial Intelligence Subcommittee of the National Information Security Standardization Technical Committee Held in Beijing’ (全国信息安全标准化技术委员会人工智能分技术委员会成立大会在京召开) (*The Central People’s Government of the People’s Republic of China*, 20 July 2017) <https://www.gov.cn/xinwen/2017-07/20/content_5212064.htm> accessed 27 July 2025

Yerramsetti R, ‘Detecting Model Hallucinations in RAG’ (*LinkedIn*, 22 November 2022) <https://www.linkedin.com/pulse/detecting-model-hallucinations-retrieval-augmented-rag-yerramsetti-f0gec> accessed 17 June 2025