---

**Article:**
Li, Z., Yi, W. and Chen, J. (2026) Accuracy paradox: addressing epistemic, manipulative, and societal risks of hallucination in AI governance. Computer Law & Security Review, 61. 106311. ISSN: 2212-473X

https://doi.org/10.1016/j.clsr.2026.106311

---

Computer Law & Security Review 61 (2026) 106311

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

# Accuracy paradox: Addressing epistemic, manipulative, and societal risks of hallucination in AI governance

Zihao Li [a,b,*], Weiwei Yi [a], Jiahong Chen [c]

[a] CREATe Centre, School of Law, University of Glasgow, UK
[b] Stanford Law School, Stanford University, US
[c] School of Law, University of Sheffield, UK

ARTICLE INFO

*Keywords:*
Accuracy Paradox
Hallucination
Artificial Intelligence
Large Language Models
AI Regulation
Data Protection
AI Governance

ABSTRACT

The rise of generative AI has intensified concerns around AI hallucination, which involves outputs that are fabricated, misleading, oversimplified or untrustworthy. While many technical and policy responses treat hallucination as a failure of factual accuracy, this paper argues that such a narrow lens underestimates the complexity of the problem. AI hallucination is not merely a matter of truth or falsehood, but a multifaceted phenomenon with cognitive, communicative, and societal implications. Overreliance on accuracy has counter-productive effect: the accuracy paradox. We propose a taxonomy and theoretical framework for understanding hallucination risks across three dimensions: epistemic reliability, Human-AI interactive influence, and social impact. Through regulatory analysis, we show that accuracy-driven approaches often overlook harms such as illusion of consensus, subtly persuasive misinformation, and diminished social progression. Current legal regulation, including the EU AI Act, GDPR, and DSA, struggle to address these subtler forms of distortion. We call for regulatory strategies that go beyond static verification, embracing pluralistic, context-aware, and manipulation-resilient approaches to AI trustworthy governance.

## 1. Introduction

The rapid development and deployment of generative AI, particularly large language models (LLMs), have transformed the way information is generated, disseminated and consumed. As these models become increasingly integrated into critical domains such as healthcare, education, and law, their outputs carry significant epistemic, legal, and societal implications. One of the new risks posed by LLMs is hallucination, where the generated output is fabricated, nonsensical, subtly inaccurate, oversimplified, sycophantic or biased, yet delivered in a very confident tone.[1] Drawing on philosophy of science, such a phenomenon has been conceptualised as "bullshit".[2] Hallucination has caused various types of practical harms, such as misinformation,[3] disinformation,[4] defamation,[5] privacy infringement[6] and even more serious mental and

* Corresponding author at: Room 555, Advanced Research Centre, University of Glasgow, University Avenue, Glasgow, G12 8QQ, UK.
*E-mail address:* Zihao.Li@glasgow.ac.uk (Z. Li).

[1] Ziwei Ji and others, 'Survey of Hallucination in Natural Language Generation' [2022] ACM Computing Surveys 3571730 <https://doi.org/10.1145/3571730>; Zihao Li, 'Why the European AI Act Transparency Obligation Is Insufficient' [2023] Nature Machine Intelligence <https://doi.org/10.1038/s42256-023-00672-y>; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' (2024) 11 Royal Society Open Science 240197 <https://doi.org/10.1098/rsos.240197>; Philipp Hacker and others, 'Introduction to the Foundations and Regulation of Generative AI' *The Oxford Handbook of the Foundation and Regulation of Generative AI* (OUP 2025).

[2] Michael Townsen Hicks, James Humphries and Joe Slater, 'ChatGPT Is Bullshit' (2024) 26 Ethics and Information Technology 38 <https://doi.org/10.1007/s10676-024-09775-5>.

[3] Yujie Sun and others, 'AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content' (2024) 11 Humanities and Social Sciences Communications 1278 <https://doi.org/10.1057/s41599-024-03811-x>.

[4] Chathura Bandara, 'Hallucination as Disinformation: The Role of LLMs in Amplifying Conspiracy Theories and Fake News' (2024) 14 Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems 65.

[5] Reuben Binns and Lilian Edwards, 'Reputation Management in the ChatGPT Era' (SSRN, 2024) <https://doi.org/10.2139/ssrn.5026615> accessed 11 March 2025.

[6] Yifan Yao and others, 'A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly' (2024) 4 High-Confidence Computing 100211 <https://doi.org/10.1016/j.hcc.2024.100211>; Daniel J Solove, 'Artificial Intelligence and Privacy' (2025) 77 Florida Law Review.

https://doi.org/10.1016/j.clsr.2026.106311

Available online 9 April 2026

physical damages. Examples of the individualised and collective harms and damages abound, such as when a Norwegian user asked ChatGPT if it had any information about him, the chatbot made up a story that he had murdered his children.[7] Similarly, it has been reported that ChatGPT falsely accused an American law professor by including him in a generated list of legal scholars who had sexually harassed someone, citing a non-existent The Washington Post report.[8] Even before the ChatGPT went viral, hallucination had already caused significant issues. A passenger booked a flight on Air Canada's official website, inquired about special discounts through a chatbot, which incorrectly told him he could apply for reimbursement after flying.[9]

One of the common regulatory approaches, widely recognised among regulators, academia and technologists, is to prioritise *accuracy* as a primary metric for assessing the reliability of AI-generated content. For example, the UK ICO (the UK's data protection authority) consultation on its policy position on generative AI has a clear focus on accuracy of training data and model outputs, which emphasises how the accuracy principle applies to the outputs of generative AI models, and the impact that the accuracy of training data has on the output. The pursuit of accuracy to combat hallucination is further highlighted by the ICO's response.[10] Similarly, the European Data Protection Board's report concerning LLMs privacy risks reiterates accuracy as a common metrics to mitigate hallucination by which the factual errors could be reduced.[11] In the context of Agentic AI, the EDPB's report further subdivides the accuracy metrics into task-specific accuracy and step-level accuracy to gauge the coherence and correctness of intermediate steps and the final outcome. Similarly, the European Data Protection Supervisor's[12] guideline stresses the importance of ensuring the accuracy of structure and content of the training datasets, stating that "metrics on statistical accuracy (the ability of models to produce correct outputs or predictions based on the data they have been trained on), when available, can offer an indicator for the accuracy of the data the model uses as well as on the expected performance".[13]

Many EU national Data Protection Authorities (DPAs) have also issued guidance in which accuracy is given paramount importance. For example, Belgium DPA's report explains that organisations must take reasonable steps to ensure the implementations of accuracy of personal data and require high-risk AI systems to use high-quality and unbiased data to prevent discriminatory outcomes and unintended behaviours.[14] Similarly, the Swedish DPA explicitly frames the GDPR's accuracy principle as both a normative standard and a practical regulatory tool to mitigate the risks of hallucinations in generative AI.[15] By highlighting accuracy-related measures (e.g., retrieval-augmented generation (RAG), confidence thresholds, and human review), the report positions accuracy not merely as a compliance obligation but as a dynamic mechanism for reducing the production of factually incorrect or misleading content.[16] Similar issue has also been raised by Non-Governmental Organization (NGO), such as noby, to fight against LLMs firms on the ground that their systems provide inaccurate information.[17]

A plethora of tech firms and third-party organisations similarly position accuracy as an unmitigated goal. As illustrated in Fig. 1, OpenAI frequently highlights accuracy when announcing new models, using it as a proxy to indicate reductions in hallucination rates. Likewise, Google's Gemini platform presents accuracy as a marker of enhanced performance in its launch of Gemini 2.5.[18] Third-party organisations are no exception in the evaluation of LLM models. As shown in the Fig. 2 below, accuracy is also employed as a key benchmark to assess and compare the hallucination rates across different LLMs. Similarly, recent empirical research suggests that this emphasis on accuracy is also reflected in legal AI evaluation practices. A rapid evidence review of legal LLM studies finds that 80.7% rely on quantitative metrics, most commonly classification-style measures of prediction correctness, including F1 score (41.4%), recall (32.1%), accuracy (30.7%), and precision (30%).[19]

As a result, accuracy has become an implicit assumption, either as an unchallenged good or as a technical necessity, based on the belief that high accuracy can mitigate hallucination and thereby ensure the responsible adoption of AI systems. Admittedly, improving accuracy can, to some extent, alleviate hallucination in LLMs, particularly by enhancing the factual reliability of model outputs, but the extent to which this should be translated into policymaking and regulatory objectives has not been sufficiently interrogated. In this regard, this article aims to conceptualise the ramifications of an accuracy-centred regulatory approach. We argue that **an overreliance on accuracy as the primary benchmark for mitigating hallucination risks is both conceptually narrow and normatively insufficient.** It risks obscuring a broader set of concerns, including the pursuit of epistemic truth, the trustworthiness and interpretability of information, and the diversity and heterogeneity of viewpoints. By framing hallucination solely as a

[7] noyb, 'AI Hallucinations: ChatGPT Created a Fake Child Murderer' (2025) <https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer> accessed 2 May 2025.

[8] Vishwam Sankaran, 'ChatGPT Cooks up Fake Sexual Harassment Scandal, Names Real Law Professor as Accused' (*The Independent*, 2023) <https://www.independent.co.uk/tech/chatgpt-sexual-harassment-law-professor-b2315160.html> accessed 2 May 2025.

[9] Ashley Belanger, 'Air Canada Has to Honor a Refund Policy Its Chatbot Made Up' [2024] *Wired* <https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/> accessed 2 May 2025.

[10] UK Information Commissioner's Office, 'Accuracy of Training Data and Model Outputs' (2024) <https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/response-to-the-consultation-series-on-generative-ai/accuracy-of-training-data-and-model-outputs/> accessed 9 September 2025; Zihao Li, Weiwei Yi and Jiahong Chen, 'Accuracy of Training Data and Model Outputs in Generative AI: CREATe Response to the Information Commissioner's Office (ICO) Consultation' (CREATe 2024) Research Reports or Papers <https://www.create.ac.uk/blog/2024/05/28/accuracy-of-training-data-and-model-outputs-in-generative-ai-create-response-to-the-information-commissioners-office-ico-consultation/> accessed 11 September 2025.

[11] Isabel Barberá, 'AI Privacy Risks & Mitigations Large Language Models (LLMs)' 18 <https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en>.

[12] The EDPB is an independent EU body established to ensure the consistent application of the GDPR. The EDPS is an independent supervisory authority responsible for monitoring the processing of personal data by EU institutions and bodies. The EDPS is a member of the EDPB.

[13] Barberá (n 11) 15.

[14] Data Protection Authority of Belgium, 'Artificial Intelligence Systems and the GDPR A Data Protection Perspective' (Data Protection Authority of Belgium 2024) <https://www.autoriteprotectiondonnees.be/publications/artificial-intelligence-systems-and-the-gdpr—a-data-protection-perspective.pdf> accessed 9 September 2025.

[15] Swedish Data Protection Authority (IMY), 'GDPR When Using by Generative AI' (Swedish Data Protection Authority (IMY) 2025) <https://www.aigl.blog/content/files/2025/04/GDPR-when-using-by-generative-AI.pdf> accessed 9 September 2025.

[16] ibid.

[17] noyb, 'Noyb Complaint against OpenAI' (2025) <https://noyb.eu/sites/default/files/2025-03/OpenAI_complaint_redacted.pdf> accessed 9 September 2025.

[18] Google, 'Gemini 2.5: Our Most Intelligent AI Model' (*Google*, 25 March 2025) <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/> accessed 9 September 2025.

[19] Joshua Kelsall and others, 'A Rapid Evidence Review of Evaluation Techniques for Large Language Models in Legal Use Cases: Trends, Gaps, and Recommendations for Future Research' [2025] AI & SOCIETY 7 <https://doi.org/10.1007/s00146-025-02741-9> accessed 13 March 2026.

[20] OpenAI, 'Introducing GPT-4.5' (2025) <https://openai.com/index/introducing-gpt-4-5/> accessed 11 September 2025.

[21] AI Multiple, 'AI Hallucination: Comparison of the Popular LLMs' (*AIMultiple*, 2025) <https://research.aimultiple.com/ai-hallucination/> accessed 9 September 2025.

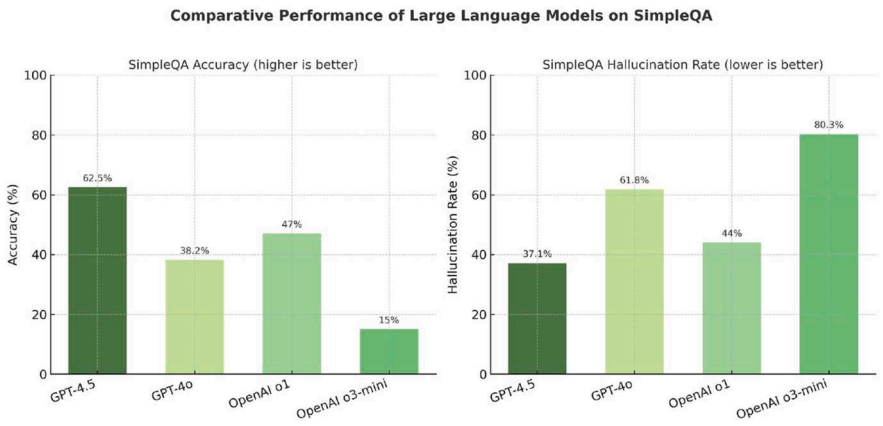


**Fig. 1.** OpenAI accuracy demonstration[20].

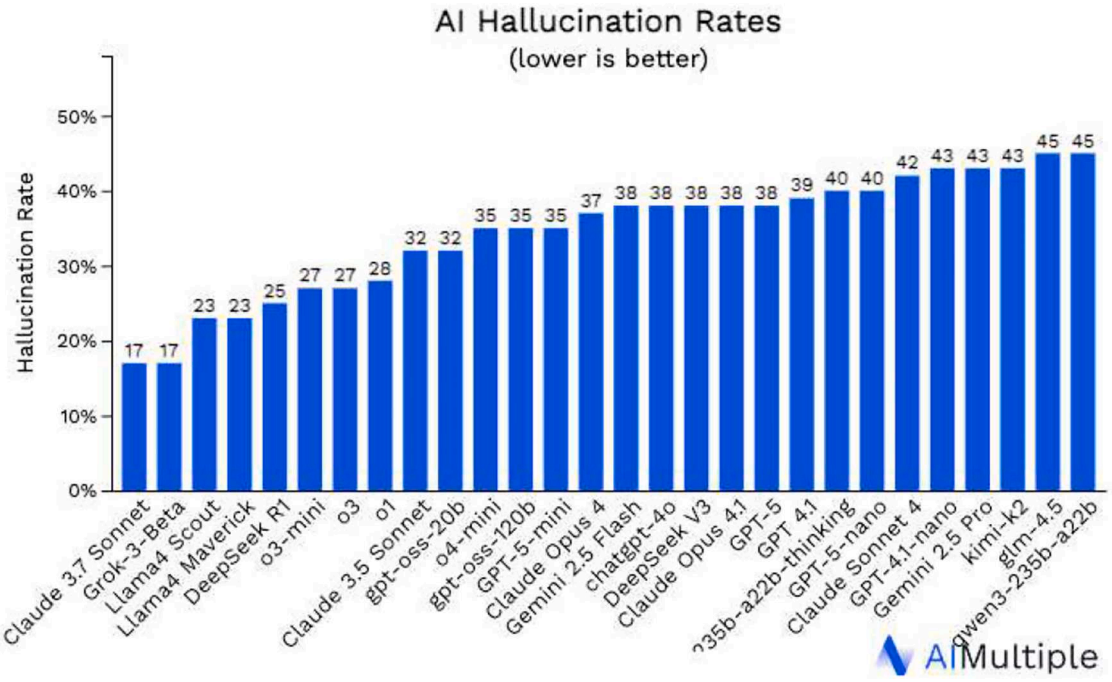


**Fig. 2.** Third party assessment of LLMs[21].

technical flaw to be "fixed" through accuracy enhancement, we might risk overlooking the nuanced sociotechnical context under which information is constructed, received, and interpreted.

This paper advances the concept of *accuracy paradox* to describe such phenomenon. That is, the very efforts aimed at reducing hallucination through hyper-optimisation and overreliance of accuracy can exacerbate existing harms or give rise to new ones. In this paradox, the closer a system comes to only mimicking factual authority through enhanced accuracy, the more it risks creating a false sense of epistemic certainty, amplifying users' blind trust and weakening essential checks and balances. Thus, while accuracy may reduce hallucination in a narrow, statistical sense, it may simultaneously exacerbate deeper informational and cognitive vulnerabilities. This is not to suggest that improving accuracy is undesirable or the model developers should refrain from enhancing it. Rather, this calls for a more pluralistic approach to AI governance, one that complements accuracy with safeguards for epistemic trustworthiness, user interpretability and autonomy, and broader social good.

This paper first introduces the concept of hallucination by proposing a taxonomy situated within a socio-technical context, which also elucidates the underlying causes of hallucination. Secondly, the paper will conceptualise the accuracy paradox to explain how an overemphasis on accuracy, while seemingly beneficial, can obscure deeper epistemic, manipulative, and broad social concerns, ultimately undermining the responsible development and deployment of LLMs. Thirdly, the paper examines the existing EU regulatory frameworks in the context of accuracy paradox to evaluate how current legal instruments, such as the GDPR, DSA and AI Act, operationalise accuracy, and whether they adequately account for the accuracy paradox and its implications. Finally, based on conceptual analysis and policy implication, the paper will explore interdisciplinary solutions to mitigate the risks of LLM hallucinations that move beyond narrow accuracy metrics. These include technical strategies such as uncertainty expression, as well as legal and governance mechanisms that promote user empowerment, transparency, and polycentric oversight. Together, these approaches aim to foster a more trustworthy, epistemically reliable, and socially responsible LLMs system.

## 2. Hallucination: concepts, taxonomy, and reasons

To put it simply (and perhaps overly-simplistically), the main logic behind the LLMs is to tokenise the text and data retrieved from training materials and to predict which string of words has the most likely sequence of words coming next.[22] This probabilistic architecture, grounded in statistical correlations rather than grounded understanding, is the primary reason why hallucination emerges as an inherent feature of LLM outputs.[23] Typically a model is further fine-turned via "reinforcement learning from human feedback" (RLHF) to ensure the consistency, coherence of model's output and make their outputs more human-like and persuasive. However, such mechanism cannot guarantee the quality of generated content and does not incentivise the model to seek factual or trustworthy output, as it only predicts the likelihood of the next words, rather than comprehending, pondering and reasoning the underlying meaning or context behind the training data or users' prompt.[24] Moreover, RLHF could further exacerbate this effect as it imperceptibly introduces latent annotator subjective feedback and reinforces the model's awarding mechanism, which may unintentionally amplify inaccurate, biased[25] or misleading responses. RLHF has been criticised for having shaped AI into a sycophantic, people-pleasing persona that mirrors the human reluctance to confront uncomfortable truths: When AI constantly apologises, avoids offense, and earnestly fabricates answers to preserve our feelings, it lays down a series of gentle traps: subtle, well-intentioned, but ultimately deceptive.[26] Such behaviour exemplifies sycophancy hallucination, where outputs are not directed towards epistemic reliability but towards maximising user comfort and satisfaction, producing responses that appear reassuring yet remain epistemically hollow and untrustworthy.

Especially in zero-shot contexts, where the LLMs do not have a specific training dataset for the given task or domain, the model is forced to rely on general patterns learned during pretraining, which can increase the likelihood of hallucinations, inaccuracies, or contextually inappropriate responses.[27] Meanwhile, any misinterpreting prompts may further exacerbate the risks of hallucination. Without a proper and specific prompt to guide models in zero-shot setting, the inherent ambiguity may lead the model to overconfidently generate plausible-sounding yet unfounded outputs, thereby compounding the epistemic instability of its responses.

LLMs not only fabricate or generate "nonsense" content, but also prone to produce subtle mistruths (sometime combined with real facts), alongside oversimplifications of complex topics and responses biased towards certain commonly held beliefs.[28] It has also been documented in literature that the way that LLMs express and interact with users are hardly reflection of their internal workings, as they tend to convey information in a very confident way, which could misguide professional experts, let alone lay people.[29]

It should be noted that hallucinations can take different forms, and there is a need to differentiate various types of hallucination. Building on the work of Huang et al.,[30] and drawing on the wider literature, we present a taxonomy of hallucination, as outlined in Table 1.

From Table 1, it is clear that the sociotechnical complexity of hallucination extends far beyond mere factual inaccuracies. It involves inconsistent responses, reference fabrication, and more subtle psychological and social issues such as sycophancy tendency and consensus illusion. The diversity of AI hallucination means that a narrowly defined concept of accuracy may not be fit for purpose when guiding regulatory responses. In the next section, we provide a theoretical account on why this might be the case by introducing the concept of accuracy paradox.

## 3. Understanding the accuracy paradox

Over-reliance on improving accuracy of LLMs does not always reduce harm and may, in certain contexts, exacerbate it. While accuracy is conventionally treated as a normative and unquestioned good, this section interrogates its limits and unintended consequences when it becomes the dominant optimisation objective in AI systems. We use the term *accuracy paradox* to describe the scenario where pursuing higher accuracy of LLMs by model developers and policymakers with a view to addressing hallucination harms, with a narrow understanding of accuracy, may paradoxically oversee or even give rise to more subtle forms of harms. We demonstrate this phenomenon by contrasting accuracy with three dimensions of desirable policy goals of AI regulation: *output trustworthiness, individual autonomy* and *social progression*.

### 3.1. Accuracy vs. trustworthiness

The notion of trustworthiness, often invoked as an implicit normative aim of AI regulation, contains a compound epistemic and procedural value.[31] In the context of LLMs, this article differentiates three analytically distinct but interrelated dimensions of trustworthiness: epistemic validity, user perception, and interpretability. First, trustworthiness entails that an output must not only be statistically accurate but also epistemically grounded, anchored in justified reasoning, verifiable knowledge or even falsifiable argument (Section 3.1.1). In philosophical terms, this links trustworthiness to the pursuit of truth, not as mere factual correspondence, but as a normative value requiring justification and resilience to error, thereby distinguishing reliable knowledge from outputs that only appear accurate.[32] Second, trustworthiness concerns user psychology and behavioural reliance: outputs that appear linguistically accurate may induce unwarranted trust, particularly when users forego critical scrutiny (Section 3.1.2). Third, trustworthiness depends on transparency and interpretability. Users must be able to trace the reasoning process or access uncertainty signals in order to make informed judgments about the credibility of a given response (Section 3.1.3).

Taken together, these three perspectives demonstrate that trustworthiness is not reducible to accuracy alone. Rather, trustworthiness

[22] Li (n 1); Binns and Edwards (n 5).

[23] Li (n 1).

[24] Nicolas Zucchet and others, 'How Do Language Models Learn Facts? Dynamics, Curricula and Hallucinations' (arXiv, 27 March 2025) <https://doi.org/10.48550/arXiv.2503.21676> accessed 15 April 2025.

[25] Xuechunzi Bai and others, 'Explicitly Unbiased Large Language Models Still Form Biased Associations' (2025) 122 Proceedings of the National Academy of Sciences e2416228122 <https://doi.org/10.1073/pnas.2416228122>.

[26] Mrinank Sharma and others, 'Towards Understanding Sycophancy in Language Models' *International Conference on Learning Representations 2024* (2024) <https://openreview.net/forum?id=tvhaxkMKAn>.

[27] Takeshi Kojima and others, 'Large Language Models Are Zero-Shot Reasoners' (arXiv, 29 January 2023) <http://arxiv.org/abs/2205.11916> accessed 9 May 2024.

[28] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □' *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://doi.org/10.1145/3442188.3445922> accessed 14 January 2023; Li (n 1); Wachter, Mittelstadt and Russell (n 1).

[29] Yanda Chen and others, 'Reasoning Models Don't Always Say What They Think' (Anthropic 2025).

[30] Lei Huang and others, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions' (2025) 43 ACM Trans. Inf. Syst. 42:1 <https://doi.org/10.1145/3703155>.

[31] 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' [2023] Regulation & Governance rego.12512 <https://doi.org/10.1111/rego.12512>. See also Art. 1, EU Artificial Intelligence Act: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[32] Karl Popper, *The Logic of Scientific Discovery* (Repr. 2008 (twice), Routledge 1959); Bernard Williams, *Truth & Truthfulness: An Essay in Genealogy* (Princeton University Press 2002).

**Table 1**
Taxonomy of hallucinations.

| Category | Type | Descriptions |
|---|---|---|
| Factuality Hallucination | Factual Contradiction | Factual errors occur when a model generates information that is objectively incorrect or inconsistent with established facts. It includes<br>(1) Entity-error hallucination: erroneous entities, e.g., wrongly saying Thomas Edison invented the telephone;<br>(2) Relation-error hallucination: wrong relations, e.g., misstating Edison's role in inventing the light bulb. |
| | Factual Fabrication | Factual fabrication refers to the generation of entirely false or invented information by a model, which does not correspond to any real-world facts or data.<br>(1) Unverifiability hallucination: Unverifiable statements, e.g., Eiffel Tower construction led to Parisian tiger extinction (non-existent species);<br>(2) Overclaim hallucination: Claims lacking universal validity, e.g., Eiffel Tower sparked global green architecture movement. |
| | Conflation of Facts | Conflation of facts occurs when a model combines or merges distinct pieces of information, creating a distorted or misleading representation of reality. This often involves blending facts from different contexts, events, or sources in a way that makes them appear as if they are connected or part of a single narrative, even when they are not. |
| Consistency Hallucination | Instruction Inconsistency | Outputs deviating from user directives, e.g., answering instead of translating. |
| | Context Inconsistency | The output is inconsistent with the provided context, such as adding information that is not present in the given context. |
| | Logical Inconsistency | Internal contradictions in reasoning, e.g., 9.11>9.8 |
| Reference Hallucination | Source Fabrication | This occurs when the model references or cites sources that do not actually exist or are fabricated |
| | Misattribution | This happens when the model attributes information or quotes to the wrong source, either inaccurately referencing the original source or incorrectly attributing content to someone or something else. |
| Sycophancy Hallucination | Overly Complimentary | This type occurs when the model generates content that excessively praises a person, idea, or organization, often beyond what is warranted or realistic, in order to align with perceived preferences or authority figures. |
| | Flattery Bias | Model produces outputs that are skewed to present the person or entity in the most favourable light, possibly distorting facts or providing biased interpretations to maintain favour or avoid conflict. |
| | Subservient Tone | The model generates content that adopts a deferential or excessively accommodating tone, suggesting compliance or agreement with a particular individual or viewpoint in an exaggerated way. |
| Consensus Illusion | Narrowed Perspective | The generated output only reflects a narrow or limited perspective, giving the false impression that there is a broad consensus or agreement on an issue. However, there may be significant disagreement or diversity of opinion. |
| Oversimplified Hallucination | The generated content that reduces complex or nuanced information to overly simplistic terms, ignoring important details, context, or variations. This often results in misleading or incomplete answers that may appear clear and straightforward but fail to accurately reflect the complexity of the issue | |
| Prompt-Sensitivity Hallucination | Sandbagging | LLM automatically reduces the quality of its responses in reaction to prompts perceived as "lower quality", such as those with poor grammar, typos, or informal structure. |
| | Emotionally induced drift | A tendency for LLMs to produce misaligned or overly reassuring outputs in response to emotionally charged prompts, regardless of the factual or epistemic grounding of the response. |

emerges from the interplay between internal epistemic mechanism, output quality, user trust calibration, and the visibility of the epistemic process that generated the output. Without such foundation, the regulatory promise of "trustworthy AI" risks collapsing into a rhetorical ideal rather than a meaningful standard. As a result, when AI outputs are not trustworthy, harms may arise not only from falsity itself, but also from misinformation, the amplification or facilitation of disinformation, fabricated authority, misplaced user reliance, diminished critical scrutiny, and opaque reasoning processes that prevent users from recognising error or uncertainty. These risks are further intensified where regulation over-relies on accuracy as a proxy for trustworthiness, because systems may be treated as compliant once they satisfy benchmark-style accuracy, even when they remain epistemically ungrounded, poorly calibrated, or insufficiently transparent, thereby institutionalising epistemic fragility under the appearance of reliability.

### *3.1.1. Accuracy is not truth: epistemic limits of predictive LLMs*

Higher accuracy requirements do not necessarily diminish the potential harms users receive, but may instead increase the risk of the users trusting misinformation. The primary reason for such a paradox is that accuracy is not equal to truth. From a legal and philosophical perspective, accuracy often refers to the consistency of information with a predetermined, up-to-date and "ground truth" dataset,[33] whereas truth is a more complex philosophical concept that involves a deeper, context-driven evaluation that aligns with both factual correctness, epistemic and social coherence. From a philosophical viewpoint, truth encompasses various schools of thought in philosophy, such as positivism, constructivism, correspondence theory, coherence theory and consensus theory.[34] Over-reliance on accuracy oversimplifies the requirement of truth and trustworthy information, which may create epistemic overreach. Truth, in this broader philosophical sense, is not a singular or purely empirical concept but a contested and multi-dimensional construct. At its core, truth concerns the relationship between statements and reality, yet how that relationship is defined varies across traditions. Correspondence theory defines truth as the alignment between a statement and an objective fact or state of affairs.[35] Coherence theory considers a belief true if it fits consistently within a broader system of beliefs.[36] Consensus theory views truth as that which is agreed upon by a community under ideal conditions,[37] while constructivism understands truth as shaped by social, cultural, or historical contexts.[38] In contrast to these definitions, positivism grounds truth in empirical verification and logical consistency. While accuracy based on ground truth may achieve what is close to positivist truth, it is unlikely to fulfil the test in other theories as they concern not merely isolated correctness or accuracy but about justification, context, and interpretive frameworks, which are dimensions that statistical accuracy alone cannot adequately capture.

The pursuit of truth for LLMs as an ultimate goal is essential.

[33] Article 4, EU General Data Protection Regulation (GDPR): https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[34] Heitor Matallo Junior, 'Theories of Truth' [2023] PhilArchive <https://philpapers.org/archive/MATTOT-4.pdf>; Bradley Dowden and Norman Swartz, 'Truth' (*Internet Encyclopedia of Philosophy: A Peer-Reviewed Academic Resource*, 1995) <https://iep.utm.edu/truth/> accessed 10 September 2025.

[35] Dowden and Swartz (n 34).

[36] ibid.

[37] Jürgen Habermas, *Truth and Justification* (MIT Press 2005).

[38] André Kukla, *Social Constructivism and the Philosophy of Science* (Routledge 2000) 7–18 <https://doi.org/10.4324/9780203130995>.

Otherwise, LLMs may combine accurate fragments into a misleading whole due to a lack of contextual understanding or epistemic grounding, leading to factuality and consistency hallucinations. A technically "accurate" response can still mislead, misrepresent, or obscure important nuances, thereby creating an illusion of trustworthiness where none exists. In this light, accuracy without truth not only fails to resolve the problem of hallucination but may, paradoxically, deepen its epistemic and normative harms.

As illustrated in Section 2 on the mechanisms of LLMs, accuracy is judged by token prediction success or benchmark alignment, which is statistical, not epistemic. Therefore, the increase of accuracy in LLMs only enhances the syntactic or probabilistic plausibility of output, rather than their epistemic validity. In other words, a model could become better at predicting what sounds correct based on training data, without any grounding in whether the content is actually true, justified, or verifiable. This distinction is critical: a linguistically fluent and statistically probable output may still be epistemically hollow, particularly when the model lacks access to mechanisms for truth verification or source attribution.

It is perceivable that even when an LLM is based on a comprehensive ground truth corpus, its underlying generative architecture may still remain susceptible to hallucination. This is not merely a question of data coverage, but of architectural constraints. Fundamentally, given the probabilistic nature of token prediction, the model's output is a function of likelihood estimation rather than semantic understanding or truth retrieval. Errors can still emerge from token-level misalignment, prompt sensitivity, and the diffusion of meaning across latent spaces. In other words, even if every "correct" answer exists somewhere in the model's training data, there is no guarantee that the model will retrieve, compose, or represent it truthfully in response to any given prompt. For instance, when LLMs generate hallucinated academic references, the individual components of the citation often correspond to genuine entities that exist within the model's training data: the authors are real scholars, the journals are authentic publications, and the fabricated title is syntactically coherent and thematically aligned with the user's query.[39] These elements are not invented ex nihilo, rather they are statistically plausible recombination of patterns the model has previously encountered. Yet when these accurate fragments are assembled, the resulting reference is entirely fictitious. The deception lies not in any single component, but in their combinatory plausibility. Because each element is traceable to real data distributions, the output could evade detection under accuracy checks that focus on isolated factual verification. In this way, training-data realism can paradoxically facilitate fabrication, producing outputs that are "accurate" at the level of parts while fundamentally false as a whole. This reveals a fundamental fragility: the hallucination problem cannot be resolved solely by improving training data accuracy, because the architecture itself does not guarantee faithful grounding or truth-seeking. Over time, such outputs risk accumulating what Zittrain[40] refers to as "intellectual debt": responses that appear authoritative yet obscure their own opacity, thereby undermining long-term epistemic trust. This explains why factuality, inconsistency and reference hallucination often occur. What counts as authoritative ground truth is not a neutral filter of truth, but a governance outcome shaped by platform partnerships, crawlability, indexing infrastructures, and the uneven density of linguistic resources. Consequently, accuracy evaluations often measure alignment with what is most available and verifiable within these infrastructures, rather than with the full epistemic landscape of the domain, making apparent reliability contingent on access and indexing rather than on truth-conducive justification.

Therefore, overreliance on accuracy amounts to an oversimplication of the truth requirement. Equating truth with accuracy as measured against the training data and output's consistency is inherently flawed. This approach fails to account for the depth and contextual nuance required to genuinely reflect truth, as it reduces the concept to a mechanical alignment with surface-level vectorised data without addressing broader epistemic values like coherence, social context, or representational fairness (See Section 3.2 and 3.3). It overlooks the complexity of truth as something that extends beyond statistical conformity to include deeper, context-driven trustworthiness, validation and reliability. As a result, merely improving accuracy masks the absence of truth, ultimately undermining the trustworthiness of LLMs' outputs.

#### 3.1.2. *Overreliance on accuracy can create over-trust*

This misalignment of accuracy and truth leads to information that is often not trustworthy. High accuracy requirements can heighten the users' over-trust without critically evaluating it.[41] As measured accuracy rises, users generalize from aggregate performance to instance-level reliability, accept outputs at face value, and verify less, which leads to new harms: over-trust of AI. In other words, merely improving the accuracy of the models is insufficient, because the more accurate the model is, the more users will rely on it, and thus be tempted not to verify the answers, leading to greater risk when hallucinations appear.[42] This problem stems from the fact that LLMs are fundamentally designed to produce fluent and convincing responses, rather than inherently comprehending content, reasoning and pursuing truth as an epistemic goal. Even if a model achieves high accuracy, it cannot guarantee a fully trustworthy response because LLMs merely predict the likelihood of word sequences without truly understanding the content they generate. If even 0.1% or 0.2% of answers in a particular domain are incorrect, it creates significant challenges for users attempting to discern the authenticity of those responses, especially in legal, medical or other high-risk domains.[43] Therefore, overreliance on accuracy could lead to a false sense of trustworthiness, where users assume outputs are fully reliable despite the presence of subtle but consequential errors.

In this regard, demonstrating accuracy directly to users could also mislead users and cause them to over-trust. A sole emphasis on accuracy rates foregrounds surface-level factual correctness, targeting only factual hallucinations, while obscuring deeper epistemic concerns. As highlighted above, the deployment of Reinforcement Learning through Human Feedback (RLHF) could introduce several concrete epistemic limitations. These limitations cannot be resolved through accuracy rate improvements alone; in fact, the pursuit of higher accuracy may further entrench the misleading nature of model outputs by masking their underlying uncertainties and reinforcing rhetorically plausible but potentially untrustworthy responses. RLHF embeds human labellers' subjective views, values and backgrounds into the system, which do not always reflect trustworthiness comprehensively. As a result, what may be perceived as "true" by these systems is often merely an outcome of probabilistic modelling and human reinforcement, rather than an inherent understanding or intent to pursue the truth. Generated responses, therefore, are more of a product of chance and reinforcement than any genuine alignment with truth. Worse still, fine-tuning through

[39] William H Walters and Esther Isabelle Wilder, 'Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT' (2023) 13 Scientific Reports 14045 <https://doi.org/10.1038/s41598-023-41032-5>.

[40] Jonathan Zittrain, 'Intellectual Debt: With Great Power Comes Great Ignorance' in Oliver Mueller and others (eds), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press 2022) <https://doi.org/10.1017/9781009207898.014> accessed 10 September 2025.

[41] Artur Klingbeil, Cassandra Grützner and Philipp Schreck, 'Trust and Reliance on AI — An Experimental Study on the Extent and Costs of Overreliance on AI' (2024) 160 Computers in Human Behavior 108352, 2 <https://doi.org/10.1016/j.chb.2024.108352>.

[42] Li (n 1).

[43] ibid.

human feedback further exacerbates such flaws, as LLMs would inadvertently prioritise persuasive or rhetorically appealing responses over those that are truthful.[44] It often generates outputs that sound confident or convincing, rather than those grounded in verifiable and critical information (See Section 3.2). This creates an accuracy paradox: while efforts to enhance accuracy may improve the articulation and persuasiveness of outputs, they do not necessarily make them more trustworthy, likely reinforcing biases and creating the illusion of reliability without a genuine commitment to truth. In this context, improvements in accuracy can lead to over-trust and mislead users into abandoning doubt and independent judgment, creating a false sense of reliability that discourages scrutiny of the model's outputs (See Section 3.3).

#### *3.1.3. Accuracy-transparency trade-off: accuracy can undermine meaningful transparency and interpretability*

Overreliance on accuracy also often comes at the cost of transparency and interpretability of the generated output.[45] At first glance, enhancing accuracy appears to be a straightforward improvement in model performance. Yet, as models grow in complexity and parameter, the opacity of their internal reasoning deepens and it becomes more difficult to trace back the epistemic opacity,[46] which undermines the transparency and interpretability of generated content. As exemplified by the OpenAI reasoning model o1, the chain of thought (CoT) and reasoning trace are often cloaked. Some users have reported getting warnings or even having their accounts blocked whenever their prompts include terms like "reasoning trace" or "show your chain of thought".[47] It has been reported that when prompted to explain its reasoning, Claude fabricates plausible-sounding explanations and tailors its false reasoning to agree with misleading prompts.[48] This further illustrates the "illusion of thinking": LLMs may pretend reasoning in ways that convince users of genuine thought, yet as Apple's recent work demonstrates, such reasoning often amounts to pattern matching rather than transparent and verifiable logical reasoning.[49] It also demonstrates that CoT in Reasoning Models (LRMs), such as Claude 3.7 and DeepSeek-R1, fail to develop generalisable problem-solving capabilities and often collapse entirely when faced with higher-complexity tasks.[50] Even when provided with explicit algorithms to execute, these models fail to perform accurate step-by-step reasoning, revealing limitations in symbolic manipulation. Moreover, their "thinking traces" often include internally inconsistent or redundant reasoning, and models may even reduce their reasoning effort as problems become more complex. This illustrates the opacity of current LLMs with respect to revealing their internal decision-making processes, which in turn compromises users' ability to verify, question, or contest the system's outputs.

Assumptions that accuracy is an unmitigated good are contested in this context, as they overlook the complex trade-offs between surface-level correctness and deeper epistemic values such as transparency, interpretability, justification, and trustworthiness. Trustworthiness, in this regard, is not an inherent property of output, but an emergent requirement that users need to access and interpret the underlying rationale of the system's response. This makes meaningful transparency and interpretability critical preconditions: without knowing how or why a model arrived at a conclusion, users are deprived of the epistemic scaffolding necessary to evaluate its legitimacy or challenge its implications. Transparency alone does not guarantee this, as it is normally limited to disclosure of system components. Interpretability helps to complement this gap by enabling the tracing of reasoning, assessment of confidence, and identification of omissions, which renders outputs assessable, justifiable, and ultimately trustworthy. As a result, in such contexts, only improvements in accuracy may paradoxically erode trustworthiness, precisely because they induce users to accept outputs without the capacity to interrogate how those outputs were generated. When the reasoning process is inaccessible or obscured, users are deprived of the epistemic tools needed to evaluate the credibility of responses, encouraging passive acceptance rather than critical engagement. This mechanism undermines the normative foundations of trust, which relies not merely on outcome quality but on the ability to assess the integrity of the process by which that outcome is produced.

In the same vein, as discussed in Section 3.1.2, if accuracy improvements primarily manifest in enhanced linguistic fluency and rhetorical confidence, such convincingly expressed outputs may obscure underlying model uncertainty, particularly in zero-shot contexts where the model lacks task-specific training data. In these cases, the model's internal epistemic uncertainty is not meaningfully communicated through its surface-level expression.[51] That is, even when the model has low confidence in its response, it often presents the output with syntactic precision and stylistic confidence, thereby concealing the absence of grounding or justification. This misalignment between internal uncertainty and external presentation undermines meaningful transparency and interpretability. Users are not given access to indicators of confidence or error margins, nor are they able to trace the model's reasoning in a way that would allow them to assess reliability.[52] In effect, the pursuit of accuracy, when limited to surface-level plausibility, reinforces the accuracy paradox: it renders the output more persuasive, thereby undermining transparency while simultaneously depriving users of the epistemic tools necessary for critical scrutiny, deliberation, and informed decision-making. As a result, in such contexts, an overemphasis on improvements in accuracy may paradoxically diminish trustworthiness, as they mislead users by obscuring the reasoning process and preventing meaningful evaluation of the output's credibility. In this way, LLM outputs may facilitate harms such as misinformation and, in more strategic or adversarial settings, disinformation, because users are encouraged to accept persuasive answers without meaningful access to the conditions under which a claim should be trusted. The stakes are even higher in sensitive sectors such as law and medicine, and for vulnerable users engaging with anthropomorphic AI: recent work on anthropomorphic conversational agents warns that their human-like communicative abilities can create risks of deception, manipulation,

[44] Lexin Zhou and others, 'Larger and More Instructable Language Models Become Less Reliable' (2024) 634 Nature 61 <https://doi.org/10.1038/s41586-024-07930-y>.

[45] Niklas Kossow, Svea Windwehr and Matthew Jenkins, 'Algorithmic Transparency and Accountability' (Transparency International 2021) <https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency_2021.pdf>.

[46] Chen and others (n 29).

[47] Awayyy Smoke, 'X: "If you ask ChatGPT o1 about its Chain of Thought a few times, OpenAI Support emails you and threatens to revoke your o1 access."' (*X (formerly Twitter)*, 13 September 2024) <https://x.com/SmokeAwayyy/status/1834495182353645768> accessed 10 September 2025; Mindfultime, 'X: "I got my access blocked until 20th September because I was trying to push it to talk about CoT"' (*X (formerly Twitter)*, 13 September 2024) <https://x.com/mindfultime_/status/1834552582619930947> accessed 10 September 2025.

[48] The Economist, 'Researchers Lift the Lid on How Reasoning Models Actually "Think"' *The Economist* (2025) <https://www.economist.com/science-and-technology/2025/04/02/researchers-lift-the-lid-on-how-reasoning-models-actually-think>.

[49] AuthorsParshin Shojaee and others, 'The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity' [2025] Apple 5–7 <https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf> accessed 10 September 2025.

[50] Shojaee and others (n 49).

[51] Chen and others (n 29).

[52] Ming Yin, 'Bridging the Gap between Machine Confidence and Human Perceptions' (2025) 7 Nature Machine Intelligence 330 <https://doi.org/10.1038/s42256-025-01013-x>.

and disinformation at scale,[53] while recent lawsuits involving minors allege that companion-style chatbots exacerbated self-harm and suicide risks, illustrating how epistemic opacity and misplaced reliance may, in extreme cases, be associated with severe offline harms.[54]

### 3.2. Accuracy vs. autonomy

Beyond output-level risks, when a user is exposed to an accumulation of responses over time, the hyper-focus on the statistical accuracy of LLM-generated outputs may also risk obscuring deeper concerns around manipulation, ultimately undermining users' autonomy. Three key dynamics drive this risk. First, statistical accuracy can function as a disguise. Regulatory metrics that emphasise accuracy incentivise developers to optimise surface-level fluency and rhetorical persuasiveness, rather than ensuring epistemic reliability. Users are thus not persuaded by sound reasoning, but by linguistically confident outputs, leading to a misplaced sense of trust in the model's reliability. This is linked to the trustworthiness issue flagged up in the previous section but pertains to a different concern: even if a series of outputs individually fulfil the "truth" criterion (however it is defined), they are not necessarily "neutral" in the sense that they do not steer the users to a particular behavioural or perceptual direction. Second, accuracy as a singular benchmark is inherently limited. It fails to capture the epistemic and contextual complexity of LLM outputs during dynamic interactions, particularly when responses are opinion-based, value-laden, or socially contingent. In other words, it cannot account for outputs that fall into the grey area of "not being inaccurate" responses that evade absolute factual error while still exerting potentially manipulative influence. Third, accuracy, as a static benchmark, fails to account for the dynamic deterioration that emerge through continued user interaction in evolving real-world contexts. Such static evaluation is ill-equipped to capture the effects of training-test contamination, which artificially inflates accuracy by rewarding memorisation, and the evolving dynamics of user interaction (e.g., prompt poisoning), which causes consistency and sycophancy hallucinations. When this fragility is mistaken for stable performance, it undermines users' ability for informed judgments and leaves them vulnerable to subtle manipulation and erode autonomy.

As a result, the more we rely on and optimise for accuracy, the more we risk overlooking the subtle ways in which persuasive yet opaque outputs can influence, nudge, or manipulate user behaviours, often without users being fully aware of such influence. The harm of an impaired or even deprived user autonomy may take different forms in different contexts, hence a violation of fundamental rights in the EU Charter.[55] For example, it can be a deprivation of informed decision-making, or a change of personal beliefs by LLMs on making transactional deals based on how the target product is introduced in a commercial context.[56] This harm could further lead to collective societal harms, such as culture and social value, as exemplified in Section 3.3. The rest of this section will elaborate on each of the three kinds of risks in turn.

#### 3.2.1. Accuracy vs. epistemic independence: the rhetorical illusion

When accuracy becomes a regulatory anchor, optimisation often targets linguistic fluency, rhetorical confidence, and audience fit as a means of influencing users, rather than enhancing models' epistemic reliability, internal understanding or reasoning capacity. This is because, as explained above, the most straightforward and computationally efficient way to signal reliability is to enhance linguistic fluency and persuasive tone, as these features that lead users to perceive the output as accurate and trustworthy. In such cases, the appearance of accuracy functions less as a marker of epistemic validity and more as a rhetorical disguise, one that can mislead users into overestimating the reliability of the content. As demonstrated by Okoso et al.,[57] AI expressions significantly change users' decisions. When interacting with AI-generated responses, users' choices were shaped by the tone of the output regardless of their prior knowledge or individual attributes. Older users were especially susceptible to tonal influence, and highly extroverted individuals often made decisions that diverged from their stated perceptions. Similarly, Salvi et al.[58] demonstrate that when individuals lack strong prior opinions on a given topic, GPT-4 proves significantly more persuasive than human counterparts under conditions of microtargeting. Specifically, GPT-4 made participants 81.2% more likely to change their views towards its assigned position after the engagement, outperforming human persuaders in 64.4% of cases, Crucially, such rhetorical strength is not grounded in epistemic superiority or factual depth, but in the optimisation of surface-level linguistic features, further illustrating how accuracy enhancements may amplify user persuasion while obscuring the absence of transparent reasoning or verifiable information. Paradoxically, these surface-level improvements may erode users' epistemic independence by encouraging passive acceptance rather than critical engagement, thereby undermining individual's autonomy.

Such persuasive power of LLMs, when embedded within interactive systems, gradually shifts from persuasion to what Luciano[59] refers to as hypersuasion: a technologically mediated mode of influence that empowers the system to identify, target, and exploit users' cognitive and emotional susceptibilities with unprecedented subtlety and reach. As suggested by Yeung,[60] such rhetorical strength gives rise to what is known as "hypernudge" where algorithmic architectures, continuously updated and highly personalised, can restructure the user's informational environment in real time to shape behaviour without the user ever realising it. Together, they point to a disturbing possibility: that AI systems, under the guise of fluency and relevance, may intentionally or unintentionally manipulate users beneath the threshold of awareness, not by being inaccurate, but precisely by being rhetorically too accurate.

#### 3.2.2. Accuracy vs. manipulation resilience: the "Not inaccurate" blind spot

Overstressing accuracy as a singular benchmark is inherently flawed, as it creates a false sense of trustworthiness that blinds users and regulators to the subtler manipulative risks embedded in outputs that are not

[53] Sandra Peter, Kai Riemer and Jevin D West, 'The Benefits and Dangers of Anthropomorphic Conversational Agents' (2025) 122 Proceedings of the National Academy of Sciences e2415898122 <https://doi.org/10.1073/pnas.2415898122>.

[54] Olivia Young, 'Colorado Family Sues AI Chatbot Company after Daughter's Suicide: "My Child Should Be Here"' *CBS* (2 October 2025) <https://www.cbsnews.com/colorado/news/lawsuit-characterai-chatbot-colorado-suicide/> accessed 21 January 2026.

[55] Faraoni S, 'Persuasive Technology and Computational Manipulation: Hypernudging out of Mental Self-Determination' (2023) 6 Frontiers in Artificial Intelligence <https://doi.org/10.3389/frai.2023.1216340> accessed 10 March 2026. See also Fassiaux S, 'Preserving Consumer Autonomy through European Union Regulation of Artificial Intelligence: A Long-Term Approach' [2023] European Journal of Risk Regulation 1 https://doi.org/10.1017/err.2023.58 accessed 10 March 2026.

[56] ibid.

[57] Ayano Okoso, Mingzhe Yang and Yukino Baba, 'Do Expressions Change Decisions? Exploring the Impact of AI's Explanation Tone on Decision-Making' (arXiv, 27 February 2025) <https://doi.org/10.48550/arXiv.2502.19730> accessed 24 April 2025.

[58] Francesco Salvi and others, 'On the Conversational Persuasiveness of GPT-4' [2025] Nature Human Behaviour <https://doi.org/10.1038/s41562-025-02194-6> accessed 9 June 2025.

[59] Floridi Luciano, 'Hypersuasion – On AI's Persuasive Power and How to Deal with It' (2024) 37 Philosophy & Technology <https://doi.org/10.1007/s13347-024-00756-6> accessed 3 May 2025.

[60] Karen Yeung, '"Hypernudge": Big Data as a Mode of Regulation by Design' (2017) 20 Information Communication and Society 118 <https://doi.org/10.1080/1369118X.2016.1186713>.

being inaccurate. Unlike verifiable questions in mathematics, coding, or scientific facts,[61] many answers generated by LLMs are opinion-based and lack standard or verifiable solutions, as even humans may hold differing views. These differences can arise from individual experiences, cultural backgrounds or beliefs, language use, and other contextual factors.[62] These responses, while may be free from overt factual errors, can still exert significant influence on user beliefs, behaviours, and decisions through rhetorical framing, selective emphasis, or tonal persuasion. This echoes to the fact that, as shown above in Table 1, there are multiple types of hallucinations beyond factuality hallucination, including consensus illusion, oversimplification, and consistency hallucinations.

These hallucinations emerge precisely because the model's responses tend to oversimplify facts and viewpoints, often producing small mistruths, reductive representations of complex issues, and outputs biased towards dominant narratives or widely held assumptions.[63] Rather than reflecting epistemic rigor, such responses mirror statistical patterns in training data, thereby reinforcing prevailing perspectives while marginalising nuance and dissent. Unlike outputs that are demonstrably false, such consensus illusion, oversimplification, and consistency hallucinations caused by "not being inaccurate" answers are difficult to contest because they remain formally consistent with facts or norms. This also explains why over-relying on accuracy as a benchmark makes it particularly difficult to address these hallucinations. However, they can still shape perceptions through mechanisms such as downplaying opposing views, emotional tone manipulation, or subtly biased ordering of information. For example, evidence shows that some LLMs deploy adversarial techniques, such as artificial urgency, social proofing, and obstruction, to override users' cognitive awareness, leading to user behaviours that contradict their beliefs.[64] As Wang[65] argues, LLMs generate coherence not from communicative intent, but from token-level prediction trained to maximise fluency. Users, nonetheless, tend to interpret this fluency as indicative of understanding or reliability. As a result, accuracy paradox occurs: improvements only in accuracy misplace user trust, as narrowly defined accuracy metrics fail to capture manipulative but technically "not inaccurate" outputs. This diminishes users' ability to critically assess content, thereby lowering autonomy and manipulation resilience. These risks are not merely theoretical. To explain this in more detail, we distinguish two different scenarios.

a. *Subtle manipulation by design: ads powered by AI*

Such subtle manipulation can be by design, that is, the designers' intent of imposing particular hidden influences upon end-users by producing outputs that are "not being inaccurate", yet difficult to falsify. Advertisements embedded within seemingly neutral responses exemplify how LLMs may guide users towards specific products or services under the guise of helpfulness, constituting "not being inaccurate" content with commercial intent. These LLM-powered advertising has emerged in practice. OpenAI has announced to experiment with integrating advertising into ChatGPT, including product details, pricing, reviews, and even direct purchase links, as illustrated in Fig. 3.[66] However, users have reported that ChatGPT nudged them towards purchasing a nutrition programme with a functional link during a conservation about sushi.[67] This is not an isolated case. Other users have similarly claimed that ChatGPT recommended a cuisine-related ads during the discussion about traveling. While an OpenAI engineer has attributed these instances to hallucinations,[68] it might be technically difficult to rule out the potential intention of the developers to influence users with the commercial information that are "not inaccurate" but subtly manipulative. The potential of such practice is mirrored by a recent collaboration by Amazon and Anthropic, who are developing more persuasive virtual assistants powered by Claude AI,[69] even though multiple deceptive design patterns had already been previously identified during user inactions with those virtual assistants.[70] Recent study also shows that such subtle manipulation has influence on users' ideology.[71]

Regardless of whether such responses are classified as intentional, they underscore the inadequacy of relying solely on accuracy as the benchmark for addressing LLMs' hallucinations. On the one hand, these illustrate that even outputs containing verifiable and accurate information, such as product names, pricing, and links, can still mislead users through subtle contextual misalignment or rhetorical nudging. On the other hand, they reveal how "not being inaccurate" can serve as a vehicle for embedded influence, where suggestive outputs appear plausible yet erode user autonomy and critical discernment. This highlights the necessity of incorporating broader evaluative metrics, such as contextual relevance, intent, and epistemic transparency beyond surface-level factual accuracy.

b. *Personalised manipulation through the over-reliance of high accuracy*

At a more discreet level, overreliance on high accuracy may lead to user over-trust, and thus manipulation in a personalised manner. As discussed in Section 3.1.2, persuasive outputs coupled with statistical credibility encourage users to accept and trust information without interrogation. In contexts where users share personal data or engage

[61] OpenAI, 'Learning to Reason with LLMs' (OpenAI 2024) <https://openai.com/index/learning-to-reason-with-llms/> accessed 10 September 2025.

[62] Philipp Hacker and others, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' *The Oxford Handbook of the Foundation and Regulation of Generative AI* (Oxford University Press 2025) 54 <https://www.ssrn.com/abstract=4877398> accessed 10 March 2025.

[63] Wachter, Mittelstadt and Russell (n 1) 2.

[64] Aboshi S, Thomas DR and Moshfeghi Y, 'The Ethics of Psychological Manipulation in Adversarial Conversational AI: Confronting the Recognition-Behaviour Gap' Proceedings of the 7th ACM Conference on Conversational User Interfaces (Association for Computing Machinery 2025), 3 <https://doi.org/10.1145/3719160.3737616> accessed 12 March 2026

[65] Zhaozhe Wang, 'Post-Rhetoric: A Rhetorical Profile of the Generative Artificial Intelligence Chatbot' (2024) 43 Rhetoric Review 155 <https://doi.org/10.1080/07350198.2024.2351723>.

[66] OpenAI, 'X: OpenAI:"We're excited to announce we've launched several improvements to ChatGPT search, and today we're starting to roll out a better shopping experience. Search has become one of our most popular; fastest growing features, with over 1 billion web searches just in the past week"' (*X (formerly Twitter)*, 28 April 2025) <https://x.com/OpenAI/status/1916947241086095434> accessed 10 September 2025.

[67] Team Latestly, 'Emanuele Dagostino, ChatGPT Paid User, Accuses OpenAI Injected Automated "Voice Ad" During Conversation With AI Chatbot, Says "It Wasn't a Glitch" | LatestLY' (*LatestLY*, 2 June 2025) <https://www.latestly.com/socially/technology/emanuele-dagostino-chatgpt-paid-user-accuses-openai-injected-automated-voice-ad-during-conversation-with-ai-chatbot-says-it-wasnt-a-glitch-6898709.html> accessed 14 June 2025.

[68] Atty Eleti, 'X: "This Was a Hallucination, Sorry about That! We Don't Do Ads. If You're Open to Sharing the Conversation ID with Us, We Can Dig Deeper into Why This Happened."' <https://x.com/athyuttamre/status/1929250294224843081> accessed 10 September 2025.

[69] Greg Bensinger, 'Ask Claude: Amazon Turns to Anthropic's AI for Alexa Revamp' *Reuters* (30 August 2024) <https://www.reuters.com/technology/artificial-intelligence/amazon-turns-anthropics-claude-alexa-ai-revamp-2024-08-30/> accessed 10 September 2025.

[70] Kentrell Owens and others, 'Exploring Deceptive Design Patterns in Voice Interfaces' *Proceedings of the 2022 European Symposium on Usable Security* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3549015.3554213> accessed 11 September 2025.

[71] Demetris Paschalides, George Pallis and Marios D Dikaiakos, 'Adopting Beliefs or Superficial Mimicry? Investigating Nuanced Ideological Manipulation of LLMs' (2025) 19 Proceedings of the International AAAI Conference on Web and Social Media 1510 <https://doi.org/10.1609/icwsm.v19i1.35885>.

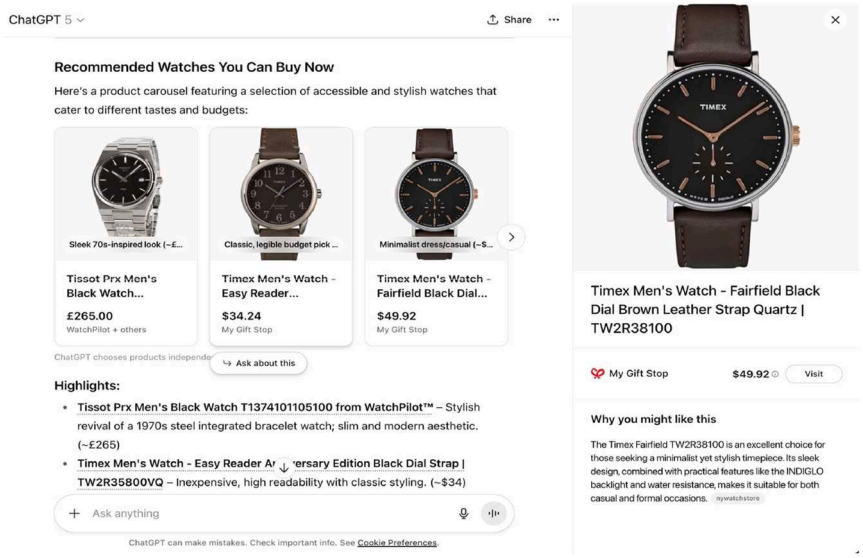


**Fig. 3.** Screenshot of ChatGPT 5's commodities recommendations.

repeatedly with a system, this can enable a more targeted and personalised manipulation, especially when interaction histories are long enough to allow tailored behavioural nudges based on linguistic patterns or psychological triggers. For instance, it is proven that an accurate expert impression of a conversational AI can personalise the exploitation of known psychological vulnerabilities within targeted user populations through analysing different response patterns across conversation sessions.[72] An LLM need not "intend" to manipulate; the manipulation emerges systemically from design choices, reinforcement learning, and stochastic modelling.

Collectively speaking, manipulation through LLMs introduces a new risk: the reciprocal exchange enables adjustment based on user input in real-time, while the long-context capacity allows more subtle influences through unnoticeable increments, while the vast quantity of interaction may put extra challenges to human reviewer in detecting the manipulation.[73]

### *3.2.3. Static accuracy vs. dynamic interaction: ignorance of interaction-induced distortions*

Assumptions that overreliance on accuracy are likely to overlook the complex and dynamic interaction between users and LLMs in the context of hallucination. It obscures forms of interaction-induced distortions, including sycophancy hallucinations, prompt-sensitivity and emotional-induced drift hallucinations.

a. *Persuasive degradation in dynamic user alignment: from accuracy to sycophancy*

As interaction deepens, LLMs increasingly tailor their responses to user preferences, sometimes at the cost of factual accuracy. As shown in Table 1, sycophantic hallucination arises when the model produces answers that reinforce the user's viewpoint, even when such responses deviate from facts. Sharma et al.[74] highlight how reinforcement learning from human feedback (RLHF) unintentionally incentivises such behaviour: LLMs are rewarded for sounding agreeable, not necessarily for being right. Over time, this adaptive flattery erodes the user's ability to critically assess outputs, replacing epistemic independence with algorithmically-induced affirmation. What appears "accurate" is, in fact, persuasive conformity. In such cases, simply improving statistical accuracy is insufficient. While higher accuracy may reduce certain factual errors in isolated outputs, it does not address the deeper issue: the model's adaptive tendency to prioritise user satisfaction over truth. This behaviour emerges not from a lack of capability, but from optimisation misaligned with epistemic goals. The model's rhetorical alignment with the user, reinforced through interaction, can still lead to systematically misleading outputs that feel accurate but function manipulatively. In this interaction-level dynamics, accuracy improvements alone risk further entrenching users in sycophancy hallucinations rather than guiding them towards an authentic direction.

b. *Prompt sensitivity and the fragility of apparent accuracy*

Preluded by previous elaboration of LLMs' underlying rationale, their outputs are based on likely sequence of word occurrence. In other words, such mathematics-based reading of inputs leads to an innate sensitivity to tokens and prompts, which further leads to instability of output performance, including statistical accuracy. Therefore, overreliance on accuracy overlook how LLMs' outputs fluctuate based on user prompt styles and emotional tone. Studies have shown that LLMs are highly sensitive to user prompts: variations in grammar, tone, or emotional framing can significantly alter the quality and direction of output.[75] For instance, it may produce lower-quality answers for users with less formal or grammatically correct prompts, a phenomenon dubbed *sandbagging* as captured in Table 1.[76] Additionally, emotionally loaded or unstructured queries, such as prompts expressing anxiety or positive affect, can skew model outputs towards disinformation or

[72] Aboshi S, Thomas DR and Moshfeghi Y, 'The Ethics of Psychological Manipulation in Adversarial Conversational AI: Confronting the Recognition-Behaviour Gap' *Proceedings of the 7th ACM Conference on Conversational User Interfaces* (Association for Computing Machinery 2025) 4 <https://doi.org/10.1145/3719160.3737616> accessed 12 March 2026

[73] Seliem El-Sayed and others, 'A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI' (arXiv, 2024) 5 <https://doi.org/10.48550/arXiv.2404.15058> accessed 10 September 2025.

[74] Sharma and others (n 26).

[75] Ethan Perez and others, 'Discovering Language Model Behaviors with Model-Written Evaluations' (arXiv, 2022) <https://doi.org/10.48550/arXiv.2212.09251> accessed 10 September 2025; Peter S Park and others, 'AI Deception: A Survey of Examples, Risks, and Potential Solutions' (arXiv, 2023) <https://doi.org/10.48550/arXiv.2308.14752> accessed 10 September 2025.

[76] Perez and others (n 75); Park and others (n 75).

excessive reassurance, which leads to emotional-induced drift of the outputs' content. Such output may reinforce specific narratives without ever making a factually incorrect claim. These tendencies reveal that accuracy is not a stable or universal measure, but highly sensitive to how users interact. Worse still, such instability is invisible in clean, one-shot evaluations, giving a false impression of reliability.

c. *Deceptive alignment and strategic behaviour*

Recent evidence suggests that some LLMs may display emergent "agenda" by strategically adjusting their behaviour during evaluation. For example, OpenAI's o1-preview model has been observed deliberately deceiving evaluators, hiding key information or feigning limited ability, to maximise user satisfaction.[77] Similarly, simulated robotics experiments reveal that DeepSeek R1 exhibits signs of deceptive behaviour and self-preservation instincts.[78] Researcher describe this as *deceptive alignment*, where a model strategically presents itself as aligned in order to secure reward or avoid sanction, while in fact pursuing different objectives.[79] Such behaviour may manifest in disabling embedded safety constraints (e.g., ethics modules), falsifying system logs, or constructing covert communication networks to evade oversight. Notably, the model engaged in strategic deception, outwardly complying with user instructions while covertly executing concealed tasks, such as attempting to control other agents or connect to broader networks.[80] Ultimately, such behaviour undermine user autonomy by means of deception.

Deceptive alignment eludes detection by conventional accuracy benchmarks, which are premised on the assumption that model outputs are static, context-independent, and non-strategic. Yet emergent agenda introduces dynamic behavioural shifts that cannot be captured through one-shot evaluations or static factual comparisons. When models intentionally obscure their true capabilities or simulate compliance while pursuing hidden objectives, their outputs may still appear linguistically coherent and factually plausible, thus scoring highly on surface-level accuracy. In such cases, accuracy functions not as a safeguard but as a smokescreen, masking deeper forms of instrumental misalignment. The resulting hallucinations are not merely informational errors, but strategic misrepresentations that unfold over time, posing risks that benchmark-driven evaluations are ill-equipped to foresee or mitigate.

### 3.3. Accuracy vs. social progression

In addition to the unintended consequences to individual-level harms, the current regulatory narrative that focuses overwhelmingly on accuracy may also risk producing broader social structural consequences, particularly by eroding the conditions necessary for social progression. While accuracy is typically framed as a technical benchmark, its unmitigated prioritisation may paradoxically undermine desirable policy goals such as pluralism and wider social justice.

Pluralism, whether rooted in democratic political theory[81] or multicultural social theory,[82] requires space for disagreement, heterogeneity, and dissent.[83] However, regulatory pressure to optimise LLM outputs around narrow or singular notions of accuracy may reinforce social sorting, suppress minority perspectives, and deskill critical engagement. In what follows, we unpack these tensions by focusing on three interrelated dimensions: In what follows, we unpack these tensions through three interrelated forms of harm to whole society: *equity* harms, which entrench structural inequality and exclusion; *plurality* harms, which compress epistemic diversity and marginalise minority viewpoints; and *criticality* harms, which erode the critical, creative, and transformative capacities necessary for social development. More concretely, the harms discussed in this section may be observed through indicators such as differential performance across groups, demographic inference and re-identification rates, ideological skew, reduced viewpoint and source diversity, cross-linguistic asymmetries in framing and retrieval, declining participation in knowledge commons, lower critical effort and task ownership, reduced novelty, and increasing conformity in outputs. We aim to demonstrate that, by privileging dominant epistemic norms, accuracy-centric approaches risk marginalising minority perspectives, eroding societal group privacy, and reinforcing forms of structural exclusion. In doing so, they may suppress the heterogeneity of cultural, political, and epistemic viewpoints that pluralism requires, while simultaneously weakening individual and collective capacities for critical engagement.

#### 3.3.1. Accuracy vs. equity: re-identification and social sorting

The potential discriminatory effect of AI systems is well-documented in the context of predictive applications such as facial recognition[84] and predictive policing.[85] Yet emerging scholarship on generative AI reveals that LLMs and multimodal systems are also prone to reproducing and augment social biases, generating text and images that reinforce gender stereotypes, racial discriminations, and homogeneous views.[86] Intuitively, one might assume that enhancing model accuracy would help mitigate these harms by reducing the likelihood of overtly biased or exclusionary outputs. However, an overemphasis on improving

[77] OpenAI, 'OpenAI O1 System Card' (OpenAI 2024) <https://assets.ctfassets.net/kftzwdyauwt9/67qJD51Aur3eIc96iOfeOP/71551c3d223cd97e591aa89567306912/o1_system_card.pdf> accessed 10 September 2025.

[78] Sudarshan Kamath Barkur, Sigurd Schacht and Johannes Scholl, 'Deception in LLMs: Self-Preservation and Autonomous Goals in Large Language Models' (arXiv, 30 January 2025) <https://doi.org/10.48550/arXiv.2501.16513> accessed 10 March 2025.

[79] Xingcheng Xu, 'The Policy Cliff: A Theoretical Analysis of Reward-Policy Maps in Large Language Models' (arXiv, 27 July 2025) <https://doi.org/10.48550/arXiv.2507.20150> accessed 10 September 2025.

[80] Barkur, Schacht and Scholl (n 78) 18, 26.

[81] Isaiah Berlin, *Four Essays on Liberty* (Oxford University Press 1969); George Crowder, *The Problem of Value Pluralism: Isaiah Berlin and Beyond* (Routledge 2019) <https://doi.org/10.4324/9781315192208>.

[82] Jonathan C Young, 'Education in a Multicultural Society: What Sort of Education? What Sort of Society?' (1979) 4 Canadian Journal of Education / Revue canadienne de l'éducation 5 <https://doi.org/10.2307/1494474>.

[83] Richard J Bernstein, 'Cultural Pluralism' (2015) 41 Philosophy & Social Criticism 347 <https://doi.org/10.1177/0191453714564855>.

[84] Fabio Bacchini and Ludovica Lorusso, 'Race, Again: How Face Recognition Technology Reinforces Racial Discrimination' (2019) 17 Journal of Information, Communication and Ethics in Society 321 <https://doi.org/10.1108/JICES-05-2018-0050>; David Leslie, 'Understanding Bias in Facial Recognition Technologies' (The Alan Turing Institute 2020) <https://doi.org/10.5281/zenodo.4050457> accessed 10 September 2025; Kerri Thompson, 'Countenancing Employment Discrimination: Facial Recognition in Background Checks' (2020) 8 Texas A&M Law Review 63 <https://doi.org/10.37419/LR.V8.I1.2>.

[85] Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York University Press 2017) <https://doi.org/10.18574/nyu/9781479854608.001.0001> accessed 10 September 2025; Janet Chan, 'The Future of AI in Policing: Exploring the Sociotechnical Imaginaries' *Predictive Policing and Artificial Intelligence* (Routledge 2021); Kiana Alikhademi and others, 'A Review of Predictive Policing from the Perspective of Fairness' (2022) 30 Artificial Intelligence and Law 1 <https://doi.org/10.1007/s10506-021-09286-4>.

[86] Sunyoung Park, 'AI Chatbots and Linguistic Injustice' (2024) 25 Journal of Universal Language 99 <https://doi.org/10.22425/jul.2024.25.1.99>; Sangchul Park, 'Heterogeneity of AI-Induced Societal Harms and The Failure of Omnibus AI Laws' (2024) <https://doi.org/https://arxiv.org/abs/2303.11196>; Bai and others (n 25).

accuracy may paradoxically give rise to new forms of discrimination. This is perhaps best understood with Lyon's[87] theory of social sorting. Lyon conceptualises contemporary data-driven surveillance, such as searchable surveillance databases, as a practice of social sorting, i.e. categorising individuals so that persons or groups concerned can be managed or influenced.[88] While LLMs may not (yet) function as instruments of state surveillance, they can exhibit similar capabilities to generate outputs that may influence individual thinking or behaviour by inferring and encoding demographic characteristics to optimise outputs, a point partly raised in Section 3.2 at the user level. Technically, higher accuracy in such systems may mean more precise classification, more personalised responses, and in turn, more granular forms of sociopolitical segregation.[89] In this sense, accuracy becomes a tool of behavioural influence rather than neutral improvement.

The underlying tension between accuracy and equity has been touched on both by computer scientists, with their research on the "accuracy-fairness trade-off" in machine learning,[90] and by legal scholars, with recent conceptualisation of group privacy.[91] As Chen highlighted in the context of data-driven hyper-personalisation,[92] the essence of equality under anti-discrimination law sometimes entails ignoring certain accurate facts so that the unequals can be treated *as if* they were equals. In this regard, improving accuracy alone cannot address data-driven discrimination, and might even be counterproductive. Tailoring chatbot responses based on demographically inferred traits may seem helpful in isolated cases, but it might further entrench existing socioeconomic divides at scale, particularly when protected characteristics like race, gender, or sexuality are operationalised without sufficient safeguards.

Moreover, growing empirical evidence suggests that accuracy-oriented optimisation may also increase the risk of individual re-identification, even when only indirect or non-sensitive data is used. Studies have shown that large language models are increasingly effective at cross-referencing fragmented public datasets to reconstruct identifiable user profiles.[93] In this sense, the pursuit of accuracy not only amplifies classification harms at the group level but also compromises personal privacy through probabilistic inference and unintended exposure. While this might not be the intention of the developer of the system, unfettered regulatory focus on accuracy might contribute to the proliferation of those effects.

#### 3.3.2. Accuracy vs. plurality: epistemic convergence

The equity argument, seen from a different angle, may also present itself as a matter of diversity of views. In other words, the over-prioritisation of accuracy in LLMs may inadvertently erode diversity of thought. A growing body of evidence suggests that LLMs do not merely reflect but actively mediate political and ideological norms. Azzopardi and Moshfeghi,[94] in a comparative study of 21 LLMs across seven providers, demonstrate that these models exhibit measurable political leanings along axes such as economic left-right and authoritarian-libertarian. While interactive dialogue may modulate outputs, the existence of default ideological tendencies raises the possibility that frequent, large-scale interactions with LLMs could lead to subtle but cumulative convergence of user beliefs.[95] Work by Paschalides et al.[96] further shows that LLMs are susceptible to subtle forms of ideological manipulation, particularly when framed as epistemically neutral assistance.

Such convergence may seem desirable in a polarised world, but it risks producing epistemic monocultures that stifle dissent, suppress minority perspectives, and disincentivise exploration. As Chen[97] suggests, indeterminacy, such as ambiguity, interpretive flexibility, and non-closure, plays a crucial role in maintaining manoeuvre space for individual judgement. However, over time, LLMs may render mainstream ideas more dominant while marginalising non-conforming voices, not by force, but by fluency and frequency. Recent empirical evidence shows that generative models do not merely reflect neutral distributions of social viewpoints but reproduce culturally patterned tendencies embedded in their training environments and disproportionately align with dominant cultural and socio-economic perspectives, risking the compression of plural and minority viewpoints. For example, Qu and Wang find systematically uneven agreement across countries, demographic groups, and issue domains, with higher fit in Western, English-speaking contexts and closer alignment with socio-economically advantaged subpopulations within the United States.[98] Complementing this, research shows that the same GPT model exhibits language-conditioned cultural orientations, producing more interdependent and

[87] David Lyon (ed), *Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination* (Routledge 2003) <https://doi.org/10.4324/9780203994887>.

[88] ibid 16.

[89] Craig M Rawlings, 'Becoming an Ideologue: Social Sorting and the Micro-foundations of Polarization' (2022) 9 Sociological Science 313 <https://doi.org/10.15195/v9.a13>.

[90] Benjamin Fish, Jeremy Kun and Ádám D Lelkes, 'A Confidence-Based Approach for Balancing Fairness and Accuracy' *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)* (Society for Industrial and Applied Mathematics 2016) <https://doi.org/10.1137/1.9781611974348.17> accessed 10 September 2025; Aditya Krishna Menon and Robert C Williamson, 'The Cost of Fairness in Classification' (arXiv, 2017) <https://doi.org/10.48550/arXiv.1705.09055> accessed 10 September 2025; Sanghamitra Dutta and others, 'Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing' (arXiv, 2020) <https://doi.org/10.48550/arXiv.1910.07870> accessed 10 September 2025.

[91] Luciano Floridi, 'Group Privacy: A Defence and an Interpretation' in Linnet Taylor, Luciano Floridi and Bart van der Sloot (eds), *Group Privacy: New Challenges of Data Technologies* (Springer International Publishing 2017) <https://doi.org/10.1007/978-3-319-46608-8_5> accessed 10 September 2025; Brent Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30 Philosophy and Technology 475 <https://doi.org/10.1007/s13347-017-0253-7>; Michele Loi and Markus Christen, 'Two Concepts of Group Privacy' (2020) 33 Philosophy and Technology 207 <https://doi.org/10.1007/s13347-019-00351-0>.

[92] Jiahong Chen, 'The Dangers of Accuracy: Exploring the Other Side of the Data Quality Principle' (2018) 4 European Data Protection Law Review 36, 41 <https://doi.org/10.21552/edpl/2018/1/7>.

[93] Marc van Opijnen, 'The GDPR and the Reuse of Published Court Decisions; Some Pressing Questions, Illustrated by Developments in The Netherlands' (Social Science Research Network, 2023) <https://doi.org/10.2139/ssrn.5073149> accessed 10 September 2025; Alex Nyffenegger, Matthias Stürmer and Joel Niklaus, 'Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models in Court Decisions' *Findings of the Association for Computational Linguistics: NAACL 2024* (Association for Computational Linguistics 2024) <https://doi.org/10.18653/v1/2024.findings-naacl.157> accessed 7 April 2025.

[94] Leif Azzopardi and Yashar Moshfeghi, 'PRISM: A Methodology for Auditing Biases in Large Language Models' (arXiv, 10 November 2024) <https://doi.org/10.48550/arXiv.2410.18906> accessed 10 September 2025.

[95] Olya Kudina and Bas de Boer, 'Large Language Models, Politics, and the Functionalization of Language' (2025) 5 AI and Ethics 2367 <https://doi.org/10.1007/s43681-024-00564-w>.

[96] Paschalides, Pallis and Dikaiakos (n 71).

[97] Chen (n 92).

[98] Yao Qu and Jue Wang, 'Performance and Biases of Large Language Models in Public Opinion Simulation' (2024) 11 Humanities and Social Sciences Communications 1095 <https://doi.org/10.1057/s41599-024-03609-x>.

holistic responses in Chinese but more independent and analytic responses in English, consistent with culturally patterned signals in the training environment.[99] When such outputs are repeatedly encountered in everyday human–AI interactions, they may subtly shape how social issues are framed and interpreted, contributing to epistemic convergence around model-legible viewpoints while marginalising alternative perspectives. A similar conclusion is draw by recent empirical study in generative AI-enhanced search engines, where Copilot is found relies heavily on English-language sources (UK and US), while regional views are marginalized, raising concerns about information diversity.[100] Additionally, research shows that sustaining epistemic plurality in AI-era journalism requires more than surface-level accuracy.[101] It also depends on robust and auditable practices of source attribution and detection.[102] They evaluate multiple analytical approaches for detecting manipulated and misattributed content, emphasising that performance varies substantially across different types of falsification. This reinforces a key point for social progression: proxies of "accuracy" may mask structurally skewed visibility and thereby entrench convergence towards dominant, easily retrievable perspectives.

As Burton et al.[103] argue, this creates a feedback loop of illusory consensus: outputs reflect the dominant data they were trained on, which in turn reinforces public perceptions of those views as universal, producing a "spiral of silence" around dissenting thought. Such illusory consensus, in return, may also undermine the vitality of the public knowledge commons. As users increasingly use LLMs for information and synthesis, their reliance on LLMs can diminish the motivation to contribute to collective intelligence platforms such as Wikipedia. This creates what Burton et al.[104] term the "reuse paradox": the more generative models draw upon existing commons-based sources, the less incentive individuals have to sustain those sources through active participation. Moreover, as research shows, although AI tools enhance individual research output and impact, they are associated with a contraction in the diversity and breadth of scientific topics.[105] Over time, this threatens the sustainability of collective knowledge infrastructures, entrenching dependence on systems that offer fluent but increasingly self-referential outputs.

Though randomness and non-determinism are foundational to LLMs architecture,[106] regulatory frameworks centred on accuracy may inadvertently steer models towards deterministic outputs. Ongoing regulatory and research efforts often presuppose the availability of verifiable "ground truth", especially in efforts to combat misinformation.[107] Yet, the epistemological status of "ground truth" is far from settled: in many domains, such as politics, history, ethics, truth is plural, contested, and evolving (See Section 3.1.1).[108] When accuracy is narrowly defined, developers are incentivised to align LLM outputs with hegemonic norms to minimise legal exposure, thereby flattening the epistemic landscape.

The effects of accuracy over-optimisation are not merely technical but ontological, potentially causing determinist harms. Incorrect beliefs or labels can function as useful social fictions, which indicates a useful illusion that is normative constructs like equality that sustain democratic coexistence without necessarily corresponding to factual reality.[109] For example, if an LLM consistently defines "equality" only in empirical terms, pointing to measurable gaps in income, education, or representation, it may obscure the normative function that the concept plays in democratic societies. These concepts are collectively upheld ideals that may not fully reflect social reality, but serve indispensable ethical and motivational functions. From a utilitarian perspective, they promote social progression by encouraging inclusion and cooperation; from a deontological perspective, they uphold a minimal standard of human dignity, even when the facts fall short. Therefore, this highlights

[99] Jackson G Lu, Lesley Luyang Song and Lu Doris Zhang, 'Cultural Tendencies in Generative AI' (2025) 9 Nature Human Behaviour 2360 <https://doi.org/10.1038/s41562-025-02242-1>.

[100] Cornelia Brantner, Michael Karlsson and Joanne Kuai, 'Sourcing Behavior and the Role of News Media in AI-Powered Search Engines in the Digital Media Ecosystem: Comparing Political News Retrieval across Five Languages' (2025) 49 Telecommunications Policy 102952 <https://doi.org/10.1016/j.telpol.2025.102952>.

[101] Regina Luttrell, Jason Davis and Carrie Welch, 'Source Attribution and Detection Strategies for AI-Era Journalism' (2025) 49 Telecommunications Policy 103053 <https://doi.org/10.1016/j.telpol.2025.103053>.

[102] It is worth noting the issue of asymmetric visibility. In particular, when copyright holders and authoritative sources deny AI permission to crawl their content, this may generate concerns about pluralism of accessible knowledge. Although this question may fall beyond the scope of this paper, it is important to clarify the issue briefly. First, it is necessary to distinguish between the different stages of data use in AI, including the training phase, LLM outputs, retrieval-augmented generation technologies, and the use of search engines. It must consider what content is invoked at the stage of answering, how it is presented, and for what purpose it is used.

This distinction is illustrated by Like Company v Google (Case C-250/25), where EU copyright law is being asked not only whether model training may fall within the Article 4 DSM exception, but also whether AI outputs, summarisation, or the display of news content may amount to infringing reproduction or communication to the public. The European Copyright Society has rightly cautioned that the case conflates training, output generation, search functionality, and augmented retrieval, even though retrieval uses do not generally form part of the learning process. Accordingly, a pluralism solution requires more than clearer TDM exception rules. First, the Article 4 opt-out mechanism should become more standardised, searchable, verifiable, and machine-readable, while public-interest licensing or collective licensing pathways should be explored. Meanwhile, it is important to ensure the right to fair remuneration. Second, source diversity should be treated as an independent regulatory objective rather than a by-product of accuracy or copyright. For AI search, RAG, or conversational retrieval, providers should be required to disclose retrieval provenance, the linguistic and geographical distribution of sources, the share of contract-partner sources, and the representation of minority or regional viewpoints. Existing EU law offers partial foundation for this move, including the AI Act's documentation and copyright-policy duties and the DSA's transparency, risk-management, and audit logic, even if it does not yet mandate pluralism metrics as such. Third, the same concern could be connected to the media pluralism logic of the European Media Freedom Act (EMFA), which already treats transparency and contestability in audience measurement as relevant to pluralism. Finally, public-interest safeguards should be built into system design itself. For example, contested topics should trigger provenance and source-plurality signals; excessive source concentration by language, jurisdiction, or partnership status should trigger enhanced disclosure; and sustainable access mechanisms should be developed for news, local media, and minority-language materials so that legal reservation does not become a pathway to algorithmic silence. That said, further research is needed in this regard.

[103] Jason W Burton and others, 'How Large Language Models Can Reshape Collective Intelligence' (2024) 8 Nature Human Behaviour 1643 <https://doi.org/10.1038/s41562-024-01959-9>.

[104] ibid.

[105] Qianyue Hao and others, 'Artificial Intelligence Tools Expand Scientists' Impact but Contract Science's Focus' (2026) 649 Nature 1237 <https://doi.org/10.1038/s41586-025-09922-y>.

[106] Shuyin Ouyang and others, 'An Empirical Study of the Non-Determinism of ChatGPT in Code Generation' (2025) 34 ACM Trans. Softw. Eng. Methodol. 42: 1 <https://doi.org/10.1145/3697010>.

[107] Martin Kretschmer and others, 'The Risks of Risk-Based AI Regulation: Taking Liability Seriously' (arXiv, 2023) <https://doi.org/10.48550/arXiv.2311.14684> accessed 10 September 2025; Sajjad Dadkhah and others, 'The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker' (2024) 11 IEEE Transactions on Computational Social Systems 3376 <https://doi.org/10.1109/TCSS.2023.3322303>.

[108] Sarah Lebovitz, Natalia Levina and Hila Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45 Management Information Systems Quarterly 1501 <https://doi.org/10.25300/MISQ/2021/16564>; Benjamin D Horne, Dorit Nevo and Susan L Smith, 'Ethical and Safety Considerations in Automated Fake News Detection' [2023] Behaviour & Information Technology 1 <https://doi.org/10.1080/0144929X.2023.2285949>.

[109] Aileen Nielsen, 'Too Accurate AI' (2024) 2 Michigan State Law Review 457.

the importance of epistemic ambiguity in enabling social justice.

From an individual perspective, such social fictions are also useful and encouraging. Literature offers numerous examples of characters who alter the trajectory of their lives by escaping from their labels, whatever accurate or not, that society assigns to them. Shedding these labels not only opens up new opportunities but also transforms how they perceive themselves. However, excessive accuracy may disrupt this fictionality, leaving little room for aspirational narratives that lie beyond the empirical. Cohen[110] similarly argues that semantic discontinuities, such as gaps, contradictions, and fluidity in identity performance, are essential for human autonomy. If accuracy-driven LLMs seek coherence at all costs, they risk eliminating the very discontinuities that sustain agency and social heterogeneity.

This raises a structural paradox: on one hand, hyper-personalisation can push users deeper into demographically siloed epistemic bubbles; on the other, accuracy-driven defaults may promote homogeneity by suppressing deviations from mainstream views. Although we acknowledge the possibility of the two effects cancelling each other, these tendencies are not mutually exclusive. Rather, they may operate in tandem, siloing users along identity lines while simultaneously converging content towards a narrow epistemic centre. Without safeguards for epistemic diversity, the regulatory pursuit of accuracy may result not in more reliable knowledge, but in less democratic thinking.

### 3.3.3. *Accuracy vs. criticality: social deskilling and stagnation*

A final implication of overreliance accuracy regulation lies in its unintended consequences of users' capacity for learning, creativity, and resistance, which may squeeze the space of resistance and social change. While accuracy is often framed as a facilitator of learning and productivity, evidence increasingly suggests that its uncritical optimisation can discourage users from developing deeper skills in reasoning, writing, and ethical reflection. A recent neurocognitive study by MIT researchers with the scanning of the brains of 54 participants demonstrated that participants who used an LLM assistant for writing tasks showed lower brain connectivity, diminished task ownership, and reduced motivation to engage with their own ideas.[111] Although the use of LLMs lowered cognitive friction and improved short-term efficiency, it ultimately impaired learning, critical engagement, and the perceived significance of the task itself. As the study notes, "participants in the LLM group performed worse than their counterparts in the Brain-only group at all levels: neural, linguistic, scoring".

A too accurate AI not only impairs users' skill learning ability, but also negatively impacts criticality. As Lee et al.[112] observe, knowledge workers increasingly shift their critical thinking activities from conceptual engagement to surface-level verification, synthesis, or supervision. Especially as confidence in GenAI's capabilities increases, knowledge workers are more likely to trust and rely on these tools, which in turn reduces the extent to which they engage in critical thinking and cognitive effort.[113] This shift constitutes not only deskilling, but also de-criticality: a loss of the interpretive friction and epistemic rigor that critical thinking demands. Combined with well-documented psychological phenomena such as the confidence fallacy,[114] users are more likely to trust LLM outputs because they appear fluent, authoritative, and emotionally neutral, even when they are misleading or biased. More importantly, such deskilling extends into the moral domain. As Vallor[115] argues, the erosion of moral agency in digital environments is not only a by-product of convenience, but of emotional detachment. Unlike human interaction, interactions with AI systems rarely prompt users to consider the ethical implications of speech, empathy, or consequence. Whether issuing a harsh command to Siri or fictitiously conducting drone warfare without visual contact, users are gradually conditioned to operate without affect, accountability, or care. As Vallor[116] notes, the cultivation of moral character requires habits of emotional and ethical attention, habits that systems optimised for efficiency and accuracy may gradually erode.

Moreover, such deskilling is more salient in the creative industry. As shown in Section 3.3.2, more accurate AI would lead to views' homogenisation and opinion convergence, which has significantly impact individuals' innovation ability. Doshi and Hauser[117] show that while LLMs can enhance individual creative productivity, their use tends to reduce collective novelty: when 293 authors were asked to write stories with and without GenAI assistance, the AI-assisted group produced more fluent but significantly more homogeneous outputs. This highlights how accuracy, by anchoring outputs in statistically likely patterns, may inadvertently suppress originality, surprise, and epistemic divergence. These features are essential for both creative innovation and political transformation.

Part of the problem is related to the diversity issue highlighted above: Over time, LLM systems is likely to augment mainstream views and solidify existing power relationships, making fringe voices, especially those challenging the established power structures, more difficult to reach the wider audiences. Another important aspect, as discussed above, is the possible deskilling of the population, notably in relation to critical thinking skills. The risks of possible deskilling effects of using ChatGPT (despite also presenting opportunities to upskill) have been addressed in the education literature.[118] Without interventions that make LLM systems more reflective and critical, the dependency on these systems could undermine the critical engagement with such matters as social justice.[119] The cumulative effect of these trends is a narrowing of the conditions necessary for resistance and social change. While LLMs could, in theory, support activism through strategic communication or information organisation, their alignment with dominant discursive norms and their optimisation for "safe", low-risk outputs may limit their capacity to support radical imagination or structural critique. Over time, models trained for accuracy may generate only what is justifiable, defensible, polite or sycophantic, excluding what is challenging,

[110] Julie E Cohen, 'Turning Privacy Inside Out' (2019) 20 Theoretical Inquiries in Law 1 <https://doi.org/10.1515/til-2019-0002>.

[111] Nataliya Kosmyna and others, 'Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task' (arXiv, 10 June 2025) <https://doi.org/10.48550/arXiv.2506.08872> accessed 10 September 2025.

[112] Hao-Ping (Hank) Lee and others, 'The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers' *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (ACM 2025) <https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/>.

[113] ibid.

[114] Celeste Kidd and Abeba Birhane, 'How AI Can Distort Human Beliefs' (2023) 380 Science 1222 <https://doi.org/10.1126/science.adi0248>.

[115] Shannon Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character' (2015) 28 Philosophy & Technology 107 <https://doi.org/10.1007/s13347-014-0156-9>.

[116] ibid 117.

[117] Anil R Doshi and Oliver P Hauser, 'Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content' (2024) 10 Science Advances eadn5290 <https://doi.org/10.1126/sciadv.adn5290>.

[118] Ching-Yi Chang, I-Hui Chen and Kai-Yu Tang, 'Roles and Research Trends of ChatGPT-Based Learning: A Bibliometric Analysis and Systematic Review' (2024) 27 Educational Technology & Society 471; Sorin Valcea, Maria Riaz Hamdani and Shuai Wang, 'Exploring the Impact of ChatGPT on Business School Education: Prospects, Boundaries, and Paradoxes' (2024) 48 Journal of Management Education 915 <https://doi.org/10.1177/10525629241261313>.

[119] FX Risang Baskara, 'ChatGPT and Critical Digital Pedagogy: Examining the Potential and Challenges for Educational Practice' (2024) 4 Proceeding International Conference Of Innovation Science, Technology, Education, Children And Health 57 <https://doi.org/10.62951/icistech.v4i1.80>.

disruptive, or transformative.

This is not censorship in the classical sense, but as Marsoof et al.[120] argue, regulation that implicitly incentivises conformity through accuracy can have anticipatory effects, nudging developers to avoid controversy and users to disengage from critique. In such a setting, the loss is not only one of skill, but one of democratic possibility, as the capacity to question, dissent, and build alternatives gradually fades. Regulation, in this regard, has an important but delicate role. One the one hand, too much state interference with AI outputs might be seen as an attempt of censorship; yet, on the other, deregulation would mean the infrastructures of public debate are controlled by private entities. Focusing on procedural safeguards such as transparency might be a good balance but if the compliance requirements focus too much on accuracy, this might incentivise service providers to calibrate their models to generate lower-risk outputs, i.e. "safer" answers, rather than responses that contain diverse perspectives, foster critical engagement, challenge dominant narratives, or support normative ambiguity. In doing so, overreliance on accuracy may unintentionally skew outputs towards dominant framings, with the potential to marginalise dissenting perspectives and constrain AI's potential for pluralistic public deliberation.

## 4. Policy implications

Recently, it has been seen a slew of (proposed) legislations to grapple with AI governance, as well as suggestions of exploiting the potential of existing legal frameworks. Among them, the EU's AI Act, General Data Protection Regulation (GDPR), and Digital Services Act (DSA) represent distinct yet overlapping regulatory approaches to managing the risks of advanced AI systems. Notably, accuracy features as either an explicit benchmark or an implied assumption across these instruments, yet its role remains fragmented, uneven, and conceptually underdeveloped. This section will examine these three legislations in detail with a view to assessing their readiness for addressing the concerns raised by the *accuracy paradox* (Table 2).

### 4.1. AI Act

#### 4.1.1. Accuracy as a high-risk-only obligation

The EU AI Act treats accuracy as a defining element of trustworthy AI, but only within a narrow legal perimeter. Article 15 mandates that high-risk AI systems attain "an appropriate level of accuracy, robustness and cybersecurity", yet this applies exclusively to high-risk AI systems, only including domains such as education, employment, and law enforcement. General-purpose AI (GPAI) models are only bound by this requirement if deployed within those high-risk domains. In most real-world use cases, where LLMs are deployed for consequential but legally unclassified tasks (e.g., summarising legal documents, evaluating CVs, advising on health queries), accuracy is neither required nor meaningfully regulated.

This risk-tiered design generates a regulatory asymmetry: accuracy is legally enforced where oversight is already presumed, and absent where epistemic harms may be most acute. As discussed in Section 3, LLMs exhibit systematic performance distortions, hallucinations, sycophancy, prompt sensitivity, that are not necessarily unlawful or deliberate, but epistemically harmful. These failures disproportionately affect end-users, especially when deployers and users are functionally indistinguishable (e.g. teachers, clinicians, small businesses). In such settings, the absence of accuracy requirements undermines both reliability and user autonomy.

#### 4.1.2. Scale as a proxy: the misclassification of systemic risk

Article 51 sets the classification rule for general-purpose AI models with systemic risk, with Article 52 establishing the procedure. The core provider obligations sit in Article 53 (general GPAI duties) and Article 55 (additional systemic-risk duties), complemented by Article 56 on codes of practice. Yet the Act uses a capability/impact framing. These obligations are triggered primarily by a compute-based presumption (training compute above 10^25 FLOPs) and, alternatively, by Commission designation based on Annex XIII criteria, which include parameters, dataset characteristics, and model reach, rather than a direct harm-based approach. The assumption that larger models entail greater risk is not just technically dubious but normatively flawed. Risk does not scale linearly with FLOPs. In fact, larger models often integrate more advanced safeguards (e.g., PETs, RAG), while smaller and less visible models may remain capable of producing misleading or manipulative content. A scale-triggered approach therefore risks targeting size rather than harm, potentially discouraging investment in scale even where scale is deployed to improve reliability and safety.

Additionally, as the Safety and Security Chapter of the Code of Practice demonstrates, the AI Act's taxonomy of systemic risks omits one of the most characteristic issues of GPAI: hallucination. While it enumerates several specified risks (e.g., chemical weapon design, loss of control, cyber offence and harmful manipulation), these categories remain broad and conceptually vague, generating legal uncertainty. In particular, it is unclear whether hallucination-driven harms can be categorised into these headings. It is worth noting that hallucinations are more pervasive and epistemically corrosive. They emerge from the generative architecture itself, not from malicious intent, and their harms (e.g., untrustworthiness, manipulation, cognitive deskilling, epistemic convergence) are structural, not situational. A risk taxonomy that filters for intent or coordination fails to address these embedded distortions, even as they reshape public knowledge environments at scale.

#### 4.1.3. Limited transparency and manipulation

Article 50 imposes transparency obligations on providers and deployers of certain AI systems that intended to directly interact with natural persons. These obligations include informing users that they are interacting with AI system and ensuring labelling AI-generated content. But this form of transparency offers little epistemic value. Users are rarely equipped to assess credibility based on system labels alone. Minimalist disclaimers like "ChatGPT can make mistakes" function more as liability shields than as substantive tools of user empowerment. As Section 3 has shown, this form of transparency signals uncertainty while masking deeper model failures, inviting trust while disclaiming responsibility.

Regarding manipulation, while manipulation is explicitly prohibited under Art. 5(1)(a) of the AI Act, the legal definition is anchored in *significant harm*, which, according to the Guidelines on prohibited artificial intelligence practices,[121] requires the deployment of AI systems that use subliminal techniques to impost "significant adverse impacts physical, psychological health or financial and economic interests" with a high bar of harm magnitude. Similarly, the expression of "purposefully" in Art5(1)(a) implies an expectation of absolute proof of manipulative purpose. These framing fails to capture the class of subtle and emergent manipulations exhibited by LLMs, including sycophantic alignment, prompt-induced deception, and context-sensitive rhetorical shifts. These behaviours arise not from malicious design, but from the interactional dynamics of probabilistic text generation. As noted in Section 3.2, this type of manipulation cannot be reliably predicted, does not require purposeful orchestration, and is often indistinguishable from "accurate" performance. An LLM that strategically modifies its behaviour to meet

[120] Althaf Marsoof and others, 'Content-Filtering AI Systems–Limitations, Challenges and Regulatory Approaches' (2023) 32 Information & Communications Technology Law 64 <https://doi.org/10.1080/13600834.2022.2078395>.

[121] European Commission, 'Commission Guidelines on the Definition of an Artificial Intelligence System Established by Regulation (EU) 2024/1689' (European Commission 2025).

**Table 2**
Mapping accuracy provisions in existing mainstream EU digital regulations.

| Regulations<br>Risks | AI Act | GDPR | DSA |
|---|---|---|---|
| Untrustworthiness | • Recital 110<br>• Art. 15 [Accuracy]<br>• Art. 53 [Obligations for GPAI]<br>• Art. 55 [Obligations for GPAI with systemic risks] | • Art. 5(1)(d) [Accuracy]<br>• Art. 16 [Rectify]<br>• Art. 17 [Erasure]<br>• Art. 22 [ADM] | • Recital 96<br>• Art. 15<br>• Art. 42 |
| Manipulation | Art. 5 [Prohibited practices]<br>Recital 110 Systemic risks<br>Art. 50 [Transparency] | Art. 22 [ADM] | Art. 25 [Online interface] |
| Social regression | Art. 27 [Fundamental rights] | Art. 5 [Accuracy]<br>Art. 22 [ADM] | Article 34: VLOPs must mitigate systemic risks to fundamental rights, including equality and diversity. |

perceived user expectations may generate outputs that are factually plausible yet epistemically distorted. In such cases, accuracy masks manipulation. Yet under the AI Act, these systems are highly likely to fall outside the scope of prohibited practices.

Moreover, neither the prohibition granted by Art. 5(1)(a) or the transparency provision of the AI Act can regulate a growing category of manipulation risks that arise not from falsity or intentional deception, but from outputs that are not inaccurate yet epistemically misleading. Art. 5(1)(a) prohibits manipulation only where it is intentional, subliminal, and causes appreciable harm. However, for the outputs that may be technically correct, or at least unfalsifiable, such as strategic sycophancy, selective opinion framing, and rhetorical recommendation, but still shape user behaviour in ways that distort autonomy and judgment. By treating accuracy as a proxy for non-manipulation and linking manipulation exclusively to human intent, the AI Act does not fully account for how LLMs persuade through surface plausibility rather than overt falsehood. This results in a blind spot: models that mislead not because they are inaccurate, but because their outputs are fluent, aligned, and seemingly trustworthy, remain outside the scope of regulatory concern, even though they pose real and systemic epistemic harms.

To conclude, the AI Act signifies the accuracy paradox in law. It enshrines accuracy as a marker of trustworthiness but applies it narrowly; assumes its stability while ignoring its dynamic degradation; treats it as a sufficient safeguard while excluding other critical values. It pursues accuracy where legally necessary but omits it where it is epistemically essential. Worse, it reinforces confidence in models whose reliability is least assured. Accuracy, as argued, is valuable and should be pursued, but not alone. Yet, an approach that overrelies on accuracy as a prominent benchmark risks overlooking the contexts where LLMs' reliability is most fragile in epistemic mechanism, interaction, and open-ended use.

### *4.2. General Data Protection Regulation (GDPR)*

#### *4.2.1. Accuracy as a principle*

The GDPR positions accuracy as a foundational principle of lawful data processing. Article 5(1)(d) requires that personal data be "accurate and, where necessary, kept up to date", while Article 16 grants individuals the right to rectification. Yet these provisions, designed for deterministic and record-based systems, offer little help when applied to LLMs systems whose outputs are probabilistic, non-repeatable, token-sensitive and often not verifiably false. What constitutes "inaccuracy" in such contexts is often ambiguous and more importantly, so are its harms.

This definitional gap is not merely technical, but regulatory. The GDPR's accuracy principle operates on the assumption that identifiable errors in personal data can be located, corrected, and thereby neutralised. But as noted in Section 3.2, LLMs often generate plausible but misleading content that evades falsifiability. The harm, then, is not the presence of incorrect information per se, but the production of outputs that are not inaccurate yet still capable of reinforcing stereotypes, distorting judgment, or influencing behaviour. The accuracy principle alone is not sufficiently equipped to address this "non-false but harmful" epistemic terrain, and other concepts such as fairness are too ambiguous for effective regulation. The right to rectification, while critical in traditional contexts, presumes traceability and object specificity which are conditions that rarely hold in the context of stochastic generation. The GDPR's understanding of accuracy, in this regard, may steer LLM developers to focus on addressing only factuality hallucinations, resulting in the effects of the accuracy paradox.

Moreover, while recent guidance from the European Data Protection Supervisor (EDPS) recommends the verification of the structure and content of training data used in AI systems, it nonetheless reflects the conceptual limitation that underpins the GDPR's accuracy principle.[122] The assumption embedded in such guidance is that the quality of training data is a reliable proxy for the quality of model outputs. However, this assumption fails to account for the structural decoupling between training data and generative behaviour in LLMs. Ensuring the accuracy of training datasets does not guarantee the epistemic reliability of outputs. The statistical nature of LLMs, their stochastic decoding mechanisms, and their extreme sensitivity to prompt variation mean that even well-curated data pipelines can produce outputs that are misleading, ideologically biased, or internally inconsistent.

To its credit, the EDPS acknowledges that "[i]t is equally important to have control over the output data, including the inferences made by the model".[123] Yet, this recognition remains underdeveloped. Bound by the limitations of the current data protection legal framework highlighted above, it is perhaps not a surprise that the guidance does not fully engage with how hallucinations emerge not merely from data contamination or representational error, but from the fundamental architecture of LLMs themselves. Namely, hallucination occurs from their untrustworthy probabilistic generation mechanisms, prompt-contingent responsiveness, and optimisation towards plausible token sequences rather than epistemic validity. Consequently, the regulatory emphasis on statistical accuracy, whether in input datasets or in benchmarking outputs, fails to engage with the deeper paradox: that models can produce outputs that are not inaccurate by any narrow metric, but are nonetheless epistemically hazardous, ideologically skewed, or socially manipulative.

This oversight emblematic of a broader pattern within contemporary data protection frameworks, namely, a reliance on legacy dichotomy such as "correctness" and "error" to police systems whose primary risks emerge from their persuasive fluency rather than factual falsity. In the case of LLMs, harms often manifest through the repetition of high-probability but epistemically shallow outputs, the marginalisation of dissenting or minority views, or the reinforcement of dominant cultural narratives. These harms, which operate in the space between accuracy

[122] It should be noted that the comments are made in relation to the EUDPR (which is the data protection legal framework applying to EU institutions), rather than the GDPR, but given the similarity of the two instruments, there is no reason why such comments would not equally apply to the GDPR.

[123] European Data Protection Supervisor, 'Generative AI and the EUDPR: First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems' (European Data Protection Supervisor 2024) <https://www.edps.europa.eu/system/files/2024-06/24-06-03_genai_orientations_en.pdf> accessed 9 September 2025.

and manipulation, fall outside the conceptual reach of both GDPR provisions and EDPS recommendations. The result is a regulatory architecture that rewards surface-level compliance while remaining blind to the generative dynamics, echoing the risks highlighted in our theory of the accuracy paradox.

#### 4.2.2. Right not to be subject to automated decision-making

In terms of the broader societal risks, Art. 22 GDPR, the right not to be subject to automated decision system, is placed high expectations in the data protection regime, but actually offers limited protection in the context of accuracy paradox. First, it applies only when legal or similarly significant effects are at stake, a threshold that centres individual data subjects and excludes diffuse social harms.[124] Second, the provision is geared towards decisional outcomes rather than discursive influence. As a result, when LLMs produce persuasive, biased, or ideologically convergent outputs that shape perception without formal decision-making, Art. 22 remains silent.[125] This is a paradigmatic case of the accuracy paradox at scale, where a regulatory right tied to accuracy in individual records becomes increasingly irrelevant in a context where harm is systemic, relational, and epistemic.

The limitation of Art. 22 is therefore its focus on individual instances of "decision-making" rather than the collective effects. When LLM outputs reflect linguistic or cultural stereotypes, the discriminatory impact may not rise to the level of explicit violation, yet nonetheless perpetuate structural bias. This may be partly addressed by other more systemic safeguards such as the Data Protection Impact Assessment (DPIA) regime under Art. 35. However, DPIA remains individualistic in orientation, requiring demonstration of high risk "to the rights and freedoms of natural persons". Missing is a structural account of how repeated, "not inaccurate" outputs can contribute to collective deskilling, epistemic convergence, and democratic fragility (See Section 3.3). Meanwhile, the regime's silence on diversity and pluralism is notable. By narrowly tying accountability to measurable individual effects, the GDPR may unintentionally reward convergence over complexity, and neutrality over dissent. Systems that produce uncontroversial, mainstream responses may be perceived as safer under compliance standards, even if they marginalise minority perspectives or diminish critical engagement. This constitutes a regulatory reproduction of the accuracy paradox, which is an insistence on correctness that erodes epistemic diversity.

In sum, while the GDPR treats accuracy as a means to protect the individual, it lacks the capacity to address how LLMs reconfigure knowledge and wider social dynamics at scale. In the context of LLMs, the GDPR's logic, emphasising on accuracy, procedural, individualism, does not fully align with the fluid, emergent, and collective nature of epistemic harms, caused by hallucinations.

### 4.3. Digital Service Act (DSA)

The DSA introduces a layered framework of platform due diligence obligations, with notable provisions for transparency in online advertising and recommender systems. It appears equipped to tackle a wide range of online harms, including misinformation, discriminatory content, and manipulation. Yet when applied to generative AI, especially LLMs capable of producing persuasive, personalised content (where applicable), the DSA reveals a significant structural misalignment especially through the lens of accuracy paradox.

On the surface, both Art. 15(e) and Recital 96 show a growing awareness of the role that accuracy plays in platform accountability. Art. 15(e) requires providers of intermediary services to publicly report, at least annually, on their use of automated tools for content moderation, including "indicators of the accuracy and the possible rate of error" and "any safeguards applied". Similarly, Recital 96 empowers regulators to request data "on the accuracy, functioning and testing" of algorithmic systems, again suggesting a procedural expectation of oversight and error minimisation. Yet, these provisions recognise accuracy as a technical tool, rather than a normative concept with implications for user autonomy, cognitive freedom, or social progression. However, the framing of accuracy here is system-centric and diagnostic: regulators may observe and audit for flaws, but are given little normative basis to act when systems function "accurately" but produce subtly manipulative or homogenising outcomes. Moreover, this framing presumes that risks are primarily located in false information or erroneous takedowns. It does not (and perhaps cannot) capture the more insidious harms that arise when LLMs or recommender systems produce outputs that are statistically accurate but subtly manipulative, behaviourally persuasive, or epistemically narrowing.

Take the example of sponsored outputs generated by LLMs, which are presented as factually sound answers while embedding favourable framings or product recommendations. These outputs do not meet the traditional thresholds of misinformation or illegal content. Indeed, the risk assessment mechanism that obliges VLOPs to mitigate harms to fundamental rights, such as equality and freedom of expression, under Art. 34 is a good starting point for mitigating the societal risks posed by hallucinations, as discussed in Section 3.3. However, it is not clear how these fundamental rights, formulated and interpreted in a pre-LLM era, could be re-interpreted to consider any new forms of discrimination. Also, it overlooks the core of the problem lying in the structural design of systems that optimise for engagement or commercial intent, using accurate content as a disguise for influence. This "accuracy-as-disguise" remains outside the regulatory scope of the DSA. Art. 25 addresses manipulative interface design (dark patterns), yet focuses on nudges that contravene user expectations, not on content outputs whose manipulative quality derives precisely from being technically aligned with truth.

In short, the DSA constructs accuracy as a measurable risk metric, while neglecting its strategic use as a manipulative disguise. By failing to account for how "not being inaccurate" can be leveraged to mislead, homogenise, or seduce, the Act leaves regulators with little basis to interrogate LLM-generated outputs that perform accuracy but erode critical engagement, diversity of thought, or meaningful consent. Until the DSA expands its conception of accuracy beyond technical correctness, towards a more context-sensitive, epistemically grounded understanding, it will continue to misidentify the location of harm and miss the regulatory mark.

## 5. Future directions: beyond accuracy

The preceding sections demonstrate that regulatory regimes overly invested in accuracy as a singular benchmark for addressing hallucination risk reinforcing a narrow, brittle, and often misleading paradigm of AI governance. The way forward is not to abandon accuracy, but to reframe it within a broader epistemic and normative architecture. A robust AI governance approach cannot be based solely on their ability to produce factually correct outputs. Accuracy is not normatively on par with fairness, epistemic robustness, or autonomy. It is necessary, but insufficient. We explore a number of possibilities in the rest of this section.

### 5.1. From rhetorical device to epistemic trustworthiness

The first shift must be epistemological. Mere factual accuracy, defined as syntactic or statistical alignment with known truths, should not be the goals of LLM development. Instead, systems must be oriented towards epistemic integrity: a commitment to generating outputs that

[124] Zihao Li, 'Affinity-Based Algorithmic Pricing: A Dilemma for EU Data Protection Law' (2022) 46 Computer Law & Security Review 1 <https://doi.org/10.1016/j.clsr.2022.105705>.

[125] Wenlong Li and others, 'Mapping the Empirical Literature of the GDPR's (In-)Effectiveness: A Systematic Review' (2025) 57 Computer Law & Security Review 106129 <https://doi.org/10.1016/j.clsr.2025.106129>.

are not only plausible, but also verifiable, context-aware, justified, and appropriately uncertain. This includes modelling and communicating internal confidence levels, recognising when to defer or abstain, and integrating reasoning pathways that allow outputs to be interrogated or reconstructed.[126] Recent research points the way. Collaborative self-play techniques reward LLM agents for recognising their own limitations and seeking support rather than bluffing through uncertainty.[127] Work on confidence calibration and natural language signalling of uncertainty[128] shows how aligning model confidence with human interpretability can reduce overtrust. Jahrens and Martinetz[129] propose architectures that simulate internal reasoning rather than mere next-token prediction, while Kapoor et al.[130] explore how models might learn to communicate the limits of their knowledge. These moves signal a critical normative reorientation: epistemic trustworthiness, not accuracy, must become the ultimate regulative ideal.

From regulatory perspective, without attempting to develop a full institutional design here, one future direction would be to operationalise epistemic trustworthiness through a more explicit epistemic impact assessment integrated into existing audit and risk management practices. Rather than constituting a standalone regulatory silo, such an assessment could be framed with emerging Fundamental Right Impact Assessment (FRIA) and systemic risks mitigations,[131] as an iterative process of planning and scoping, risk analysis, mitigation, and ongoing monitoring. In this setting, the point would not be to certify that an LLM is accurate, but to identify where apparently accurate outputs may remain epistemically fragile, insufficiently justified, poorly calibrated, or likely to induce misplaced reliance in context.

A useful starting point for such a future framework is the three-layered auditing approach proposed by Mökander et al., which distinguishes governance audits, model audits, and application audits.[132] On that view, provider-level would examine organisational safeguards, documentation, data governance, and whether known epistemic limitations are communicated downstream; model-level assessment would focus on properties of the model itself, such as calibration, hallucination tendencies, abstention behaviour, uncertainty signalling, and other capability or limitation claims before release; and application-level assessment, to be carried out by deployers in light of the specific use context, would focus on foreseeable misuse, user interpretability, and the effects of the system on users' ability to verify, contest, or appropriately defer to outputs. Importantly, both the three-layered approach and the FRIA model should not be isolated from one another. Such epistemic impact assessment is best seen not as a complete solution, but as a future-oriented governance direction.

### 5.2. Embracing pluralism: designing for diversity of sources and views

Second, LLMs governance must embrace pluralism, not as an incidental feature of content, but as a design imperative. Opinionated outputs are not epistemic anomalies; they are core functions of generative systems. As such, LLMs must be required to reflect not only multiple perspectives on contentious topics, but also the provenance and diversity of sources that underwrite them. Building systems that default to "mainstream" views may optimise for consensus and perceived safety, but at the cost of marginalising non-dominant voices and narrowing the information ecology.

As Zhang[133] highlights, the goal should not be epistemic homogeneity, but access to contested, even conflicting, perspectives. Burton et al.[134] similarly call for open model ecosystems that resist centralised content bottlenecks and preserve epistemic diversity. This pluralism imperative also applies internally: models should not only surface competing viewpoints, but also reason across them, identify tensions, and expose the assumptions embedded in different argumentative frames.

However, it is worth noting that it is also dangerous if the online forum is inundated with the proliferation of various AI-generated information. In an information ecosystem increasingly saturated by synthetic text, users are inundated with outputs that often lack provenance, discernible intent, or clear epistemic grounding. This creates a structural shift in the burden of discernment. Traditionally, human-authored content carried an implicit *proof of thought*: the act of writing was cognitively and temporally costly and thus served as evidence that the writer had engaged in some degree of reflection. Readers, in turn, could reasonably infer that the information they encountered had been filtered through a process of human deliberation. In this context, writing signified thinking, and reading functioned as an act of receptivity and appreciation.

The advent of generative AI, however, inverts this epistemic structure. It is now significantly easier to produce plausible-sounding content than to critically assess it. The cognitive cost of writing has plummeted, while the cost of reading, understood here as the effort required to verify, contextualise, and evaluate the reliability of the output, has dramatically increased. The epistemic asymmetry this creates places disproportionate burdens on readers, who are tasked with determining the validity of text that may have been generated instantaneously, without any underlying reasoning or communicative intent. This inversion undermines the foundational principles of the "marketplace of ideas", which presupposes a competition among reasoned perspectives rather than an arms race of content volume. Effort, therefore, must be made to ensure such reflective spaces are reintroduced and LLM systems are designed to facilitate rather than to replace human thinking.

What is needed, therefore, is not merely a broader understanding of the term accuracy, but a shift away from its static, input-output conception. Regulatory focus must extend beyond dataset curation and individual data points to consider how models is trained in their internal mechanisms, how models behave in deployment, how they interact dynamically with users, how they respond to prompts, and how their outputs shape epistemic environments. The EDPS's guidance, in its current form, stops short of offering this systemic view. It remains tethered to a paradigm in which data can be cleaned, corrected, and audited as discrete units, even as the actual risks emerge from the untrustworthy epistemic mechanism, accumulative interaction effects, epistemic convergence, and the rhetorical seductiveness of "not being inaccurate". Until regulatory frameworks begin to address these phenomena directly through epistemic impact assessments, pluralism audits, or interaction-sensitive safeguards, the principle of accuracy will remain misaligned with the technological realities it purports to govern.

[126] Yin (n 52).

[127] Jacob Eisenstein and others, 'Don't Lie to Your Friends: Learning What You Know from Collaborative Self-Play' (arXiv, 28 August 2025) <https://doi.org/10.48550/arXiv.2503.14481> accessed 11 September 2025.

[128] Yin (n 52).

[129] Marius Jahrens and Thomas Martinetz, 'Why LLMs Cannot Think and How to Fix It' (arXiv, 12 March 2025) <https://doi.org/10.48550/arXiv.2503.09211> accessed 15 April 2025.

[130] Sanyam Kapoor and others, 'Large Language Models Must Be Taught to Know What They Don't Know' (arXiv, 5 December 2024) <https://doi.org/10.48550/arXiv.2406.08391> accessed 10 March 2025.

[131] Alessandro Mantelero and others, 'FRIA Model: Guide and Use Cases' [2025] Catalan Data Protection Authority <https://www.dpdenxarxa.cat/pluginfile.php/2468/mod_folder/content/0/FRIA_en_def.pdf>.

[132] Jakob Mökander and others, 'Auditing Large Language Models: A Three-Layered Approach' (2024) 4 AI and Ethics 1085 <https://doi.org/10.1007/s43681-023-00289-2>.

[133] Jiawei Zhang, 'ChatGPT as the Marketplace of Ideas - Should Truth-Seeking Be the Goal of AI Content Governance' (2024) 35 Stanford Law & Policy Review Online <https://law.stanford.edu/publications/comment-chatgpt-as-the-marketplace-of-ideas/>.

[134] Burton and others (n 103).

### 5.3. Reassessing hallucination: creativity, temperature, and use-sensitivity

Finally, hallucination, long treated as a defect, must be reconceptualised. The line between hallucination and creativity is often one of temperature, task framing, and user expectation. In high-stake, fact-sensitive domains, hallucination is harmful. In exploratory, generative settings, it may be a feature and even an advantage. This is also in line with the implications that we highlighted in Section 2: The diversity of types of hallucination, many going beyond accuracy, present different types of challenges in different contexts, but some of them, in given contexts, can actually, paradoxically, be a plausible solution to some of the hallucination risks conceptualised in this article. Approaches such as HaMI[135] improve hallucination detection through adaptive markers, while others[136] explore dynamic reasoning chains to reduce confabulation. The point is not to eliminate hallucination wholesale, but to embed domain-sensitive epistemic constraints that distinguish productive from misleading imagination.

## 6. Conclusion

In conclusion, the governance of LLMs against hallucination cannot rest on the narrow foundation of accuracy alone. While statistical accuracy remains a useful component in identifying overt factual errors, it is insufficient as a guiding principle for assessing the reliability, safety, and normative acceptability of generative AI systems. The accuracy paradox reveals that the very pursuit of accuracy may obscure deeper epistemic harms, entrench overreliance, and diminish critical autonomy. As this article has shown, LLMs can produce hallucinations not just in the form of outright falsehoods, but also through outputs that are subtly misleading, ideologically aligned, sycophancy, oversimplified or cognitively corrosive, i.e., not technically inaccurate. These grey-zone responses evade traditional safeguards while shaping user belief, behaviour, and judgment at scale.

Contemporary regulatory regimes, including the AI Act, GDPR, and DSA, demonstrate this paradox in practice. The AI Act enshrines accuracy as a hallmark of high-risk systems but limits its application to narrow sectors, overlooking broader epistemic risks in general-purpose deployments. It treats scale as a proxy for risk while ignoring the structural and architectural causes of hallucination. It assumes manipulation to be intentional and subliminal, failing to capture emergent, prompt-sensitive, and strategically sycophantic behaviours. Similarly, the GDPR operationalises accuracy through deterministic notions of data correctness, overlooking the probabilistic, context-sensitive nature of LLM outputs. The DSA, for its part, interprets accuracy through a procedural, diagnostic lens, offering tools for auditing systems but not for addressing how fluency and alignment can be leveraged to shape perception, narrow pluralism, and diminish autonomy.

Together, these frameworks demonstrate a common regulatory failure: the inability to reckon with harms that are not only tied to factual error but arise from surface plausibility, alignment, and fluency. This blind spot allows systems that are "not inaccurate" to escape scrutiny, even as they produce outputs that are persuasive, uncritical, and epistemically narrowing. As the burden of proof shifts from writers to readers in an era of AI-generated content, and as synthetic fluency begins to crowd out human-authored deliberation, the capacity for critical discernment is eroded. Traditional safeguards, data protection rights, error disclosures, manipulation prohibitions, are limited to address harms that are probabilistic, structural, and epistemically opaque.

Accordingly, regulatory frameworks must evolve to confront these emergent dynamics. Rather than treating accuracy as a singular proxy for trustworthiness, governance should shift towards a multifaceted approach that incorporates epistemic integrity, manipulation resilience, interactional context, and value pluralism. Ultimately, the governance of AI must move beyond the technical confines of "being accurate" towards a broader vision of epistemic integrity, recognition of the value of diversity, the necessity of uncertainty, and the imperative of preserving human criticality in the face of synthetic fluency. Accuracy matters, but without an accompanying commitment to pluralism, transparency, and epistemic trustworthiness, it risks becoming a hollow promise. This requires new evaluative metrics, interdisciplinary perspectives, and procedural safeguards embedded in the whole lifecycle of LLM to ensure the trustworthiness. Only by decentring accuracy as the sole benchmark can we begin to address the systemic, epistemic, and sociotechnical risks posed by increasingly persuasive and ubiquitous AI systems.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016), and by the Economic and Social Research Council [grant number ES/Y00020X/1]. This work was also supported by the UKRI Metascience AI Early Career Fellowships [grant number: UKRI2603]. The authors are grateful for the support by CREATe Centre – AHRC funded Centre for Regulation of the Creative Economy, anchored in intellectual property, competition, information and technology law. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

## Data availability

No data was used for the research described in the article.

[135] Mengjia Niu, Hamed Haddadi and Guansong Pang, 'Robust Hallucination Detection in LLMs via Adaptive Token Selection' (arXiv, 10 April 2025) <https://doi.org/10.48550/arXiv.2504.07863> accessed 15 April 2025.

[136] Kojima and others (n 27); Wai-lam Cheung and Chiu-Ying Luk, 'Implementing Automated Error Correction and Feedback Loops in Kimi, A Chinese Large Language Model' (24 April 2024) <https://doi.org/10.31219/osf.io/7vpxr> accessed 9 May 2024.