Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37185

Systematically Reducing Hallucination in Generative AI from Users to Data to Models

Junhua Chen

School of Software and Internet of Things Engineering, University of Finance and Economics, Nanchang, China
chenjunhua@jxcjdxydmh.wecom.work

Abstract. AI hallucination is not just about a one-off technical glitch at the model level but is a systemic issue with the interplay between humans, the data, and the model itself. Most existing reviews focus on isolated parts of model development, like model architecture or data governance, and do not consider the whole model development chain from users to data to the model as an integrated whole and analyze it accordingly. To systematically review advances in research addressing hallucinatory outputs in generative Artificial Intelligence, this paper takes a tripartite approach covering users, data, and models. From the user perspective, from low prompting capabilities, cognitive biases to information discernment, which result in hallucinations, and then collectively devise prompt engineering, interactive clarification, and information literacy enhancement as three countermeasures to reduce the effects of hallucinations. From the data perspective, it explores how data quality, retrieval-augmented generation, and knowledge graphs can help strengthen the factual consistency of data and outline their intrinsic weaknesses. In terms of the model, it outline factuality-oriented decoding strategies, alignment training paradigms, and self-inspection with corrective refinement techniques. This paper clarifies the above three aspects and how they work together to form a closed-loop hallucination cycle, starting with user feedback as a departure point, moving to the next step of turning error-correction signals into training instances at the data tier, and lastly to internalising the ability of veracity discrimination in the model parameters via preference alignment and knowledge rectification at the model tier.

Keywords: Generative Artificial Intelligence, Hallucination, Large Language Models, Retrieval-Augmented Generation, Human-AI Interaction

1. Introduction

In more niche and specialised fields where professional standards are high, like medicine or law, where the effects of hallucinations of generative AI are more stark, it may be possible to cause actual, real-world damage. An example of this is when a patient was hospitalized for bromine intoxication due to a misdiagnosis by ChatGPT that had "hallucinated" the diagnosis [1], and a US attorney used ChatGPT to research legal precedents but submitted multiple made-up cases to the court without any verification [2]. Technical solutions to mitigate the incidence of hallucination at

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37185

the model level have spanned from noise mitigation in pre-training corpora and optimisation of decoding strategies to retrieval-augmented generation and reinforcement learning with human feedback, with the ultimate goal of reducing hallucination incidence through architectural improvements. In fact, efforts involving the quality assessment of training corpora, debiasing and cleansing, and even knowledge boundary annotation have increasingly become essential preliminary steps to improve factual consistency of the model outputs at the data level. At the user level, scholarly focus has progressively shifted towards investigations of prompt engineering, clarification mechanisms, user information literacy education, etc., relevant aspects.

These outputs are then the result of inappropriate user inputs which are catalysed by low-quality data in the model's generative processes. Such outputs, when accepted and trusted by the users, are again used in subsequent interaction cycles, thus creating a closed-loop hallucination system. Building on this awareness, the current work takes a tripartite view of users, data, and models and seeks to provide a systematic overview of existing approaches to hallucination mitigation in generative AI, and will also aim to provide an integrated reference framework conducive to the systematic governance of this area.

2. Causes and analysis of user-level hallucinations

This tendency to make up something rather than asking for more clarification whcih is a core issue in the reward system of RLHF. Human annotators are also likely to be biased toward accepting model answers that fit their own expectations, even if they are factually incorrect. According to the SycEval benchmark, the average conformity percentage of all models tested is 58.19%, and the initial-response conformity percentage for the medical domain is as high as 100%. Furthermore, if the first turn of dialogue is triggered, about 78% of the subsequent dialogue turns will deteriorate progressively. Moreover, when user instructions have fabricated facts or logical errors, LLMs absorb the incorrect contextual details and generate speculations that significantly increase the risk of hallucination.

First, the ability to effectively give instructions to an AI in human-AI interaction is universally limited. Both the validity and the reliability of the content generated by the model are directly and decisively affected by the quality of prompts. It has been demonstrated that using clear and precise prompts can significantly improve the quality of the AI-generated content, while vague prompts can lead to logical inconsistencies and, in extreme cases, factual mistakes [3]. For human-AI interaction, the model lacks grounding and tends to speculate and misinterpret. Users are advised to consider AI-generated content as preliminary and not final, and to continue cross-checking, particularly when it comes to facts, data, and professional opinions.

Secondly, people's information sense is generally poor. Users often treat the responses of large language models as authoritative responses instead of as generative artifacts, and show a significant lack of awareness and skill to critically evaluate AI-generated content. But when the model does fabricate content, users face huge hurdles to identify the fabrications; they also may go on to spread false information as if it were real, thereby creating a vicious cycle of uncritical use and re-use. However, to change users' mental models about how to interact with AI, it is essential that they understand that interacting with AI is not a series of question-and-answer exchanges, but a dialogic process that requires clarification, iterative changes, and progressive corrections. Prompt literacy is not just a skill of human-AI interaction. From a more basic perspective, it is an essential quality for students, which includes cognitive strategies, critical thinking, and ethical consciousness [4].

In addition, users have strong cognitive biases regarding the scope of what AI can and cannot do. When users have access to insufficient information, they often overestimate how accurate the LLMs

DOI: 10.54254/2755-2721/2026.37185

are factually and underestimate how much they speculate when they don't know, which is a systematic tendency in users. These cognitive misperceptions can lead to a reduced level of vigilance on the part of users about the quality of output, and at the same time, reduce the users' proactive tendency to add to or explain their intentions in the process of interaction. Thus, it is necessary to create a virtuous cycle involving user feedback and corrective recalibration. To a significant degree, the success of large language models depends on human prompting. That is, the human interlocutors should provide a paradigmatic framework of response, based on which the model can provide responses of reasonably high accuracy [5].

3. Causes & analysis of data-level hallucinations

Large language models try to consolidate a wide range of corpora from various sources during the data acquisition stage. Nevertheless, the training data are intrinsically limited in covering the entire range of factual knowledge in the real world and make the knowledge base of the model bounded and incomplete. If even the model's knowledge itself is plagued by biases, then every step from there on is built upon an inherently shaky and tenuous basis. Furthermore, the models face significant difficulties in keeping data resources up-to-date in terms of timeliness, and are unable to synchronize fully with the fast-changing real world.

Obsolescence and inaccuracies are common problems in source data, and the first priority in data preparation is to deal with these problems. In this regard, a fact-checking process should be implemented before data is ingested to allow for item-level verification of the data being ingested. Knowledge graph-based entity constraint mechanisms use structured knowledge repositories to perform consistency checking of entity relationships that are part of the data and detect and resolve possible conflicts in knowledge. The next step would be to temporality-annotate the output content of the model and to incrementally update it, while at the same time adding validity periods and timestamps to individual knowledge entries. Last but not least, there should be a hierarchical authority classification system used for the results. All data from different sources should have credibility assessment and graded annotation. During retrieval and generation, preferential credence should be given to a high-authority source, and the impact of low-authority sources like social media posts and forum discussions on the output generated should be reduced.

So far, it has talked about static data quality control during the pre-training phase. In actual deployment contexts, however, generative AI should also be dynamic and real-time to meet the needs of knowledge acquisition. The accuracy and effectiveness of retrieval-augmented generation heavily rely on how the external documents are chunked and indexed and how successfully the retrieval of relevant content to a specific query can be achieved. The analysis focuses on semantic chunking optimisation and hybrid retrieval strategies. In the context of semantic chunking, the most pressing one is the optimisation of text segmentation strategies. When it comes to the granularity of segmentation, research has gradually evolved away from the universal applicability of a given heuristic towards the more nuanced and finely calibrated options. Recursive character splitting is done by successively splitting at a higher level down to words and characters, which in turn preserves. Identify the integral boundaries of sentences and paragraphs as much as possible. Semantic similarity-based segmentation, on the other hand, measures the similarity of embedding vectors of adjacent sentences and places segmentation boundaries where there is a sharp change in semantic similarity, which results in high intra-chunk informational cohesion and low inter-chunk coupling. There is also a need for metadata augmentation of the segmentation process for the sake of preserving semantic completeness. Metadata includes provenance information for retrieval and filtering, sectional information for document structure reconstruction, and semantically annotated

DOI: 10.54254/2755-2721/2026.37185

information generated by models. These elements are an additional re-ranking dimension to vector retrieval. The basic idea behind metadata augmentation is that the superimposition of structured constraints on semantic similarity provides an effective orthogonal way of attaining retrieval precision, which ensures that what is retrieved is what is sought, while vector retrieval is good at semantic fuzzy matching and metadata filtering is good at providing precise structured constraints [6].

Most existing research efforts have focused on solving the ubiquitous issue of poor correlation between similarity in vector space and the meaning of the words. The most popular and empirically supported "decoupling" strategy for this dilemma is hybrid retrieval. The rationale behind it is the concurrent execution of both vector retrieval and keyword-based retrieval, thus realizing the synergistic complementarity in both the two directions of semantic similarity and literal matching. Chi, Zhang, and their collaborators took an empirical approach, testing their method on a dataset of 1, 000 policy subsidy programmes and 500 question-answer pairs, and they found that it outperforms conventional vector-based RAG by 6.6% in terms of context precision, 2.6% in context recall, and 2.8% in faithfulness. Compared to tag-based retrieval RAG, it achieved even more significant improvements, with 31.1% in context precision, 20.7% in context recall, and 22.6% in faithfulness [7]. Vector retrieval enables semantic-level similarities to be captured, which includes synonymous expressions, cross-lingual mappings, and ambiguous demand matching. In contrast, keyword-based retrieval guarantees literal consistency and exact matches, and is especially effective in certain domains where vector representations can have blind spots, like proper names, numerical identifiers, and code snippets. A weighted fusion or reciprocal rank fusion is performed on the retrieval results of these two paradigms to consolidate and rank the results, thus giving them semantic coverage and precision control [8]. To further alleviate cases of vector similarity without semantic relevance, it uses a coordinated optimisation framework combining multi-channel recall with fine-grained reranking mechanisms. The multi-channel recall strategy builds upon hybrid retrieval, traversing vector similarity pathways, as well as keyword matching and metadata filtering pathways, in parallel to widen the scope of the candidates, extend the Top-K coverage radius, and then rerank the expanded set for fine-grained ordering.

4. Causes and analysis of model-level hallucinations

During the pre-training phase, models are trained via causal language modelling, which is an approach that inherently engenders reasoning-induced hallucinations. However, there are still some obvious reasoning gaps in existing models in some specialised reasoning scenarios, such as multi-hop questions that require connecting multiple queries, or adversarially rephrased questions, which increase the possibility of hallucinatory generation of generative language models [9].

In the first place, optimisation for factuality is introduced at the level of decoding strategies. The conventional nucleus sampling is based only on the probabilistic output distribution of the language model itself and does not include any factually consistent adjudicator. To this paradigm, factuality-grounded nucleus sampling adds an exogenous factual veracity scoring module that considers each sampling candidate token as well as a conditional probability under the language model, but also an alignment score with authoritative knowledge sources. After sampling, the sampling space is then capped on less successful tokens, which helps limit the generation to a set of candidates that is more factually grounded, but with only a slight decrease in output fluency [10]. This method has been empirically tested on a 530-billion-parameter model and shown to achieve a 56% relative reduction in named entity error rate from 33.3% to 14.5% [10]. The second key approach is to put in place a confidence-thresholding mechanism. When the high-uncertainty decoding nodes are detected, this

DOI: 10.54254/2755-2721/2026.37185

policy is triggered, and the sampling distribution is tightened by decreasing the temperature parameter, thereby reducing the likelihood of hallucinatory generation, or the statistically more common expressions that have been extensively verified in similar contextual regimes in the training corpus are given preference [11]. Lastly, contrastive decoding procedures are used as an additional development. The family of approaches leverages the distributional mismatch between the models of various scales, or between the training checkpoints of different models, to enhance the signals that are grounded in facts. It is now possible to flag and filter these regions of tokens that are considered plausible by the weaker model, but are only statistically common artefacts, by calculating the ratio between probability distributions of the two models at a decoding step. At the same time, this process strengthens the factual signal strength of tokens in the unique competency aperture of the larger model.

Factuality-oriented refinements are applied at the model level in the context of the alignment training paradigm. Foremost, a multi-dimensional reward modelling framework is adopted. The scores of several dimensions are combined through weighted sums and/or gating to yield a composite reward, which extends the user-satisfaction-is-all-that-matters objective to one of user satisfaction conditioned on factual correctness and epistemic honesty [12]. Later, it resorts to adversarial preference data construction because the implicit preference signals in the training data directly influence the tendencies of the model's behaviors. This approach explicitly includes preference pairs in the dataset that include cases where the user query can be interpreted as containing factual inaccuracies, and compares the factually correct model answer with the answer that does not correct the factual mistake, using the former as the preferred option. Trained to prefer truth over fabrication, it develops a factuality-driven behavior. If the factuality score for any intermediate step is below a certain threshold, it is marked as a potential source of hallucination. Further, the model is now able to backtrack or correct itself at that point, which can prevent the progressive transmission of hallucinatory signals through the reasoning chain by using a process-supervised reward model. Fact-consistency evaluation follows. The basic method involves breaking down the generated content of the model into atomic factual statements, and then checking them against external knowledge bases on an item-by-item basis, labelling each atomic statement as corroborated, unverifiable, or contradictory to known facts [13]. In addition, a self-correction mechanism should be added to the model. Self-correction, on top of detection, gives the model a metacognitive ability to not only identify the error, but also to correct it on its own.

5. Conclusion

Initially, the pointwise optimisation should be overcome, and a tripartite synergistic governance framework should be achieved. First, it requires an upgrade towards full-chain coordination across the entire trajectory from users to data to models. Secondly, there is a dire need for systematic and quantitative studies on the user's side. In particular, a fine-grained taxonomy for user-input characteristics should be created. Finally, it is important to continue research into trust-aware interaction interfaces. Especially, efforts are needed in the research to represent internal signal information, like the certainty estimate of a model and the authority of retrieved sources, in a way that is easily understandable to the end user.

The hallucination issue in generative AI is not a one-off issue at any given point but a systemic issue that comes from the complex interdependencies between users, data, and models. Following this three-fold analytical point of view on users, data, and models, the present review systematically examines existing work on the mitigation of hallucination. After providing an overview of what has been accomplished in each of the dimensions separately, it explains the interdependencies of the

DOI: 10.54254/2755-2721/2026.37185

three, as well as the inherent limitations of the current situation, where each dimension is being tackled in isolation. The main argument of this paper is that technological iterations on the model side are not sufficient to achieve good hallucination mitigation. Instead, it requires breaking the barriers between user feedback, data provisioning, and model optimisation to set up a governance model based on full-chain synergistic coordination.

References

[1] Eichenberger, A., Thielke, S., & Van Buskirk, A. (2025). A case of bromism influenced by use of artificial intelligence. *Annals of Internal Medicine: Clinical Cases, 4*(8), Article e241260. https://doi.org/10.7326/aimcc.2024.1260

[2] Curlin, J. H., IV. (2025). ChatGPT didn't write this... or did it? The emergence of generative AI in the legal field and lessons from Mata v. Avianca. *Arkansas Law Review, 78*(1), 123–135. https://doi.org/10.54119/alr.mlrt4575

[3] Chen, W., & Zhang, F. (2025). Research on the impact of prompt constituent elements on the cross-modal content generation quality of large language models: An exploratory experiment based on the text-to-image function of iFlytek Spark model. *Journal of Zhengzhou University (Philosophy and Social Sciences Edition), 58*(2), 99–104.

[4] Wan, L., Wang, T., Zhang, T., & Zhao, J. (2026). Learner prompt literacy for generative artificial intelligence: Framework, indicators, and cultivation paths. *Modern Distance Education Research*. https://link.cnki.net/urlid/51.1580.G4.20260714.1701.002 [in Chinese].

[5] Zhao, L., Ju, H., Zhang, P., et al. (2024). Capability boundaries of large models and their applications in command and control. *Journal of Command and Control, 10*(6), 653–660. https://doi.org/10.3969/j.issn.2096-0204.2024.06.0653 [in Chinese].

[6] Yousuf, R. B., Xu, S., Sharma, M., Neeser, A., Latimer, C., & Ramakrishnan, N. (2026). Utilizing metadata for better retrieval-augmented generation. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval (ECIR 2026) (LNCS Vol. 16483)*, 305–319. Springer.

[7] Chi, Z., Pan, X., Huang, X., Zhou, X., & Huang, X. (2026). Optimization study of RAG system based on hybrid retrieval of vectors and tags in policy consulting. *Industrial Control Computer, 39*(7), 96–97. [in Chinese].

[8] Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. ACM. https://doi.org/10.1145/1571941.1572114

[9] Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2309.12288

[10] Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P., Shoeybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. https://proceedings.neurips.cc/paper_files/paper/2022

[11] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. https://arxiv.org/abs/2302.09664

[12] Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., & Hajishirzi, H. (2023). Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

[13] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741