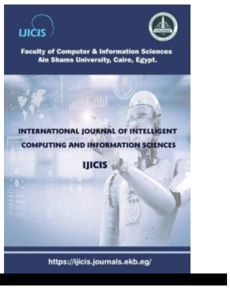
IJICIS, Vol.24, No.4, 73-81
DOI: 10.21608/ijicis.2024.336135.1365

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

HALLUCINATION MITIGATION TECHNIQUES IN LARGE LANGUAGE MODELS

Mohamed Ali*
Computer Science Department,
Computer and Information Science,
Ain Shams University,
Cairo, Egypt
mohamed.abdelghafour@cis.asu.edu.eg

Mohamed Mabrouk
Computer Science Department,
Computer and Information Science,
Ain Shams University,
Cairo, Egypt
mohamed.mabrouk@cis.asu.edu.eg

Zaki Taha
Computer Science Department,
Computer and Information Science,
Ain Shams University,
Cairo, Egypt
ztfayed@hotmail.com

Abstract: *Large language models (LLMs) have demonstrated impressive natural language understanding and generation capabilities, enabling advancements in diverse fields such as customer support, healthcare, and content creation. However, a significant challenge with LLMs is their tendency to produce factually inaccurate or nonsensical information, commonly known as hallucination. Hallucinations not only compromise the reliability of these models but can also lead to serious ethical and practical issues, particularly in high-stakes applications. This survey comprehensively reviews recent advancements in hallucination mitigation strategies for LLMs. We explore retrieval-augmented models, which enhance factual grounding by integrating external knowledge sources; human feedback mechanisms, such as reinforcement learning, which improve accuracy by aligning model responses with human evaluations; knowledge augmentation techniques that embed structured knowledge bases for enhanced consistency; and controlled generation, which restricts output to ensure alignment with factual constraints. Additionally, we examine the challenges of integrating these techniques and the limitations of current methods, including scalability, resource intensity, and dependency on quality data. Finally, we discuss future research directions to improve factual reliability in LLMs and explore hybrid solutions to create accurate and adaptable models for a wider range of real-world applications.*

Keywords: *Large Language Models, Hallucinations, Retrieval-Augmentation, Knowledge-Augmentation, Human Feedback.*

Received 2024-11-13; Revised 2024-12-22; Accepted 2024-12-23

1. Introduction

Large language models (LLMs) such as GPT-3, BERT, and OPT have revolutionized NLP, automated tasks such as summarization, translation, and conversational agents. Despite these advancements, the phenomenon of hallucinations—instances where models produce factually incorrect or irrelevant information—continues to be a significant challenge. In real-world applications such as healthcare, legal

***Corresponding Author**: Mohamed Ali

Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: mohamed.abdelghafour@cis.asu.edu.eg

advice, and finance, hallucinations can result in serious consequences, damaging the trust and reliability of AI systems.

For instance, in healthcare, incorrect medical information generated by an LLM could lead to misdiagnosis, while in the legal domain, hallucinated case precedents might misinform legal professionals. In finance, inaccurate financial summaries could lead to erroneous investment decisions. These risks highlight the critical need for effective hallucination mitigation techniques.

Recent efforts to mitigate hallucinations in LLMs focus on combining various approaches, including retrieval-augmented generation, human-in-the-loop methods, and knowledge augmentation. This paper provides an in-depth survey of these techniques and explores their strengths, limitations, and future directions for research. By addressing these challenges, the future of LLMs can be more reliable, making them applicable to high-stakes industries where factual consistency is vital.

2. Hallucination Mitigation Techniques

Mitigating hallucinations in large language models (LLMs) has become a priority, especially in contexts where factual consistency is crucial. Numerous strategies have been developed to address hallucinations by grounding models in external knowledge, refining training processes, and controlling generation outputs. These techniques generally fall into four key categories: Retrieval-Augmented Models, which integrate external data sources to provide contextually relevant information; Knowledge-Augmentation, where structured databases or knowledge graphs are embedded within the model to enhance factual reliability; Human Feedback, which involves leveraging human evaluations during training to correct and refine model outputs; and Controlled Text Generation, a technique that uses prompts or predefined conditions to constrain model responses within specific factual boundaries. Each of these approaches offers unique strengths and limitations, and together, they form a comprehensive toolkit for enhancing the factual accuracy and reliability of LLMs across various applications. Table 1 provides a summary of key studies and findings related to hallucination mitigation techniques, including Retrieval-Augmented Models, Knowledge Augmentation, Human Feedback, and Controlled Text Generation. This table highlights how each method contributes to minimizing hallucinations in specific applications.

2.1. Retrieval-Augmented Models

Retrieval Augmented Generation (RAG) is a technique that grants generative artificial intelligence models information retrieval capabilities. It modifies interactions with a large language model (LLM) so that the model responds to user queries with reference to a specified set of documents, using this information to augment information drawn from its own vast, static training data. This allows LLMs to use domain-specific and/or updated information.[1] Use cases include providing chatbot access to internal company data or giving factual information only from an authoritative source.

In practice, retrieval-augmented models like RAG have demonstrated the ability to significantly reduce hallucination. For example, in customer service applications, these models can retrieve specific product information from a database to ensure that the responses provided to customers are accurate. The REALM model, as described by Guu et al.[2], also shows potential in improving NLP tasks by grounding text generation with relevant information from external sources.

These models have been particularly effective in industries that require up-to-date and factual content. For example, in legal research, a retrieval-augmented model can fetch relevant legal precedents, improving the factuality of summaries produced by legal professionals. The ability of these models to access external databases ensures that the information they provide is both accurate and relevant to the task at hand.

2.2. Knowledge-Augmentation

A Knowledge-Augmented Language Model (KALM) is a type of machine learning model that integrates external, structured knowledge sources, such as knowledge graphs, databases, or other domain-specific resources, into its natural language processing (NLP) capabilities. By augmenting the model's understanding with this external knowledge, KALMs are able to generate more accurate and contextually informed responses compared to traditional language models.

Knowledge-augmented models play an essential role in fact-based systems, e.g. virtual assistants or question-answering systems. The integration of structured knowledge sources such as Wikipedia or domain-specific databases can help ensure that the model consistently produces factual information. This approach has been widely adopted in applications like FAQ generation for websites, where accuracy is paramount.

For instance, in medical chatbot systems, integrating medical knowledge bases ensures that the generated responses are factually grounded and reliable. This helps mitigate hallucinations by providing the model with concrete, verified knowledge during the generation process. Additionally, this approach has been utilized in education, where knowledge-augmented systems support factually accurate tutoring systems for students.

2.3. Human Feedback

A Human Feedback Model refers to an approach in machine learning and artificial intelligence (AI) where human feedback is used to guide and improve the performance of models, especially in tasks like natural language processing (NLP) and reinforcement learning. This type of model leverages human judgment, preferences, or evaluations to refine the model's behavior and outputs, ensuring that it aligns more closely with human expectations and goals.

An excellent example of human feedback in action is in content moderation systems. By incorporating human evaluations, models can be trained to better detect offensive or inappropriate content, minimizing the chances of hallucinations in these sensitive contexts. Models like the one developed by Stiennon et al. [3] exemplify how continuous human feedback can guide LLMs toward more factual outputs.

In advertising and social media management, human feedback systems are crucial for ensuring that generated promotional content adheres to brand guidelines and factuality. This process not only improves the accuracy of the generated content but also aligns it with the strategic goals of the organization, thus reducing the risk of producing off-brand or false information.

Table 1: Results from Key Studies on Hallucination Mitigation Techniques

Technique	Study	Results	Metrics	Applications
Retrieval-Augmented Models	Guu et al. (2020) - REALM Model [2]	Improved accuracy in open-domain QA through document retrieval.	40% top-1 accuracy on Natural Questions dataset	Open-domain QA for knowledge-intensive tasks
	Lewis et al. (2020) - RAG Model [5]	Combined generation with retrieval to enhance response accuracy.	44% top-1 accuracy on Natural Questions dataset	Knowledge-based dialogue systems
	Borgeaud et al. (2022) - Retrieval Techniques [6]	Achieved better factual alignment and fluency with large-scale retrieval.	Achieved 86% accuracy on Wiki-based QA tasks	Conversational AI, Knowledge-grounded summarization
Knowledge-Augmentation	Petroni et al. (2019) - Knowledge-Augmented BERT [7]	Integrated knowledge bases to improve factual responses.	Hits@1 improved by 15% in factual tasks	Virtual assistants and technical FAQs
	Yao et al. (2019) - KG-BERT [8]	Enhanced accuracy in link prediction and knowledge graph completion.	88% accuracy in link prediction	Enterprise knowledge and QA
	Bosselut et al. (2019) - COMET Model [9]	Increased commonsense reasoning through external knowledge integration.	74% accuracy on commonsense QA tasks	Educational tools, interactive tutoring
Human Feedback	Stiennon et al. (2020) - Human Feedback for Summarization [3]	Improved factual accuracy in summarization through RLHF.	Achieved 95% consistency in factual summaries	Summarization and content moderation
	Ouyang et al. (2022) - RLHF for Conversational AI [10]	Enhanced truthfulness in conversational models through RLHF.	20% improvement in truthfulness scores	Customer support chatbots, open-domain conversations
	Bai et al. (2022) - Red-Teaming and Feedback Loops [11]	Reduced hallucinations through feedback loops and red-teaming.	Increased precision (92%) and recall (88%)	Sensitive content generation
Controlled Text Generation	Keskar et al. (2019) - CTRL Model [4]	Used control codes to guide content generation, ensuring topic relevance.	Improved domain adherence by 18%	Advertising, news generation
	Holtzman et al. (2020) - Control Codes [12]	Maintained topic relevance with control codes, enhancing accuracy in specific domains.	Increased coherence in generated outputs	Scriptwriting, legal documents
	Zhang et al. (2022) - OPT Model [13]	OPT constraints improved factual accuracy in open-domain generation.	87% factual accuracy on open-domain QA tasks	Open-domain chatbots, real-time reporting

2.4. Controlled Text Generation

A Controlled Text Generation Model is a type of language model designed to generate text that adheres to specific guidelines, constraints, or desired attributes. These models allow fine-grained control over the generated output to ensure it aligns with certain goals, such as tone, style, content, or structure. This control can be achieved during the model's training or by using specific techniques during inference.

The use of controlled text generation is particularly valuable in creative industries. For example, models like CTRL [4] can generate content that adheres to specific narrative or stylistic guidelines, reducing the risk of producing hallucinated or irrelevant information. This approach is often applied in news summarization and scriptwriting, where maintaining control over the tone and factual accuracy is crucial.

In the field of journalism, for instance, using controlled text generation ensures that articles or summaries adhere to a specific factual template, thereby mitigating the risk of introducing false information. Additionally, in entertainment, controlled generation models help scriptwriters maintain creative freedom while ensuring that characters, events, and settings remain consistent and accurate within the storyline.

Table 1 provides a summary of key studies and findings related to hallucination mitigation techniques, showcasing the impact of Retrieval-Augmented Models, Knowledge Augmentation, Human Feedback, and Controlled Text Generation on reducing hallucinations in LLMs across various applications.

The comparison between the REALM and RAG models for their performance in open-domain question answering. Using datasets like Natural Questions, REALM achieves a top-1 accuracy of 40%, while RAG achieves 44% due to its enhanced integration of retrieved information into the generative process. However, RAG's higher computational requirements make it less suitable for latency-sensitive tasks.

3. Developing Models to Reduce Hallucinations

The development of models to reduce hallucinations requires improvements in architecture, training techniques, and the incorporation of external knowledge. Recent work focuses on techniques like domain-specific fine-tuning, faithfulness-aware decoding strategies, and the integration of external knowledge bases, each contributing to reducing the likelihood of hallucinated content in various applications. For instance, in the healthcare sector, retrieval-augmented models enhance chatbot capabilities by accessing real-time medical databases, enabling accurate symptom checks. In finance, controlled text generation ensures that financial reports are accurate and compliant, reducing risks associated with generating misleading investment advice.

3.1 Domain-Specific Fine-Tuning

Domain-specific fine-tuning equips language models with knowledge in specific fields, allowing them to generate more accurate and reliable outputs. For example, in the healthcare industry, a model fine-tuned on medical texts and verified clinical guidelines can reduce the chances of hallucinated medical advice, which could be dangerous. Similarly, legal models can be fine-tuned with legislative documents and case law to ensure that outputs are factually aligned with the legal framework. [14]

3.2 Faithfulness-Aware Decoding Strategies

Faithfulness-aware decoding strategies are critical for ensuring that models generate content that stays true to the source information. In tasks like abstractive summarization, these techniques help maintain factual consistency by using metrics or reward-based systems that prioritize outputs aligned with the original input. This approach has shown promise in reducing hallucinations in summarization tasks. [15]

3.3 Knowledge-Enhanced Models

Knowledge-enhanced models incorporate structured knowledge bases, such as knowledge graphs, to ground their outputs in factual data. These models can dynamically reference external knowledge during text generation, ensuring that the generated information remains accurate. [16]

3.4 Hybrid Models

Hybrid models combine natural language models with rule-based systems or retrieval mechanisms to verify the accuracy of generated content. For example, a customer service chatbot could generate responses using a neural model but check the factuality of the generated text against a structured database. By using a combination of neural generation and rule-based verification, hybrid models help prevent hallucinations, ensuring factual accuracy. [17]

3.5 Uncertainty-Aware Models

Uncertainty-aware models are designed to assess their own confidence when generating responses. When the model is uncertain about its output, it can either abstain from responding or consult an external knowledge source. This approach is particularly valuable in fields like healthcare, where the risk of hallucinated responses can be high, and model accuracy is critical. [18]

4. Conclusion

Hallucinations in large language models (LLMs) remain a significant challenge, especially in applications where factual accuracy is critical, such as healthcare, finance, legal advisory, and customer service. While LLMs have shown great promise in various natural language processing (NLP) tasks, their tendency to generate incorrect or misleading information limits their applicability in high-stakes environments.

This paper has reviewed several approaches for mitigating hallucinations, including retrieval-augmented generation, knowledge augmentation, reinforcement learning with human feedback, controlled text generation, and domain-specific fine-tuning. Each of these techniques offers unique advantages but also presents challenges. For example, retrieval-augmented models can improve factual grounding but rely heavily on the availability and quality of external knowledge bases. Similarly, human feedback has proven effective in refining models, but its scalability remains a challenge due to the labor-intensive nature of the process.

One key takeaway from the survey of these techniques is that no single method is a silver bullet for hallucination mitigation. Instead, a hybrid approach combining multiple techniques is likely to offer the most robust solution. For example, retrieval-augmented models could be paired with knowledge-enhanced systems, while reinforcement learning with human feedback could be applied to fine-tune outputs for specific domains. Such hybrid models would allow for more comprehensive control over the generation process, ensuring that outputs are both fluent and factually accurate.

Looking ahead, future research should explore more advanced techniques for improving the factual reliability of LLMs. One potential avenue is the development of self-evaluating models, where models can assess their own confidence in the accuracy of their outputs and flag potentially hallucinated content for further verification. Moreover, zero-shot learning and few-shot learning could be explored further, allowing models to perform more accurately in domains where little domain-specific data is available.

Another critical area for future work is the integration of dynamic knowledge sources that are updated in real-time, ensuring that LLMs have access to the latest information. This would be especially useful in industries such as journalism or medical research, where new facts and discoveries frequently emerge. Moreover, advances in unsupervised learning could help in reducing reliance on large, manually annotated datasets for model training.

In conclusion, while significant progress has been made in mitigating hallucinations in LLMs, the road ahead remains long. The integration of multiple techniques, combined with the development of innovative new strategies, will be crucial for ensuring that LLMs can operate reliably in environments where factual accuracy is essential. By addressing these challenges, LLMs can be transformed into more trustworthy tools capable of delivering accurate, consistent, and reliable information across a wide range of applications.

5. Discussion and Limitations

While substantial progress has been made in mitigating hallucinations in large language models (LLMs), each of the primary techniques faces inherent challenges. Retrieval-augmented models, for instance, rely heavily on the availability of accurate, up-to-date knowledge bases. Without robust and well-maintained data sources, these models may still produce hallucinated or outdated information. Furthermore, integrating external retrieval processes can introduce latency, which may be problematic for real-time applications, such as chatbots in customer support or medical consultations.

Human feedback and reinforcement learning, another promising approach, demonstrate effective results in improving factual consistency, particularly for sensitive applications. However, scaling these methods is resource-intensive, as they require continual human evaluation and feedback during training. Additionally, these models can sometimes reflect subjective biases introduced by human trainers, potentially affecting the neutrality of outputs in high-stakes domains like legal or medical services.

Knowledge-augmented models that incorporate structured knowledge bases, such as knowledge graphs, improve factual accuracy but introduce their own set of challenges. The quality of these models is intrinsically linked to the comprehensiveness and accuracy of the embedded knowledge, and maintaining updated and unbiased knowledge bases across diverse domains can be labor-intensive. Furthermore, the complexity added by integrating knowledge structures can impact the model's fluency, particularly when generating free-form text.

Controlled text generation, which restricts model outputs within predefined boundaries, is highly effective in reducing hallucinations. However, the technique can limit the model’s flexibility, sometimes resulting in outputs that are less creative or nuanced. This limitation is particularly relevant in fields like marketing or creative writing, where a balance between accuracy and expressive freedom is essential.

Future research should aim to address these limitations by exploring hybrid models that combine multiple techniques. Such models could utilize retrieval-augmented generation alongside self-evaluation mechanisms to enhance both accuracy and scalability. Additionally, developing methods to dynamically update knowledge sources and improve real-time retrieval will be essential for ensuring that LLMs remain reliable as they become more widely integrated into high-stakes applications. Ultimately, by addressing these challenges, LLMs can be made more robust, enabling them to provide factually accurate and contextually appropriate responses across diverse fields.

6. Ethical Considerations

The ethical importance of reducing hallucinations in large language models (LLMs) becomes paramount in high-stakes fields such as healthcare and law. In healthcare, AI-generated misinformation could lead to incorrect diagnoses, inappropriate treatment recommendations, or undue alarm for patients. Similarly, in legal settings, hallucinations could result in erroneous legal advice or case analyses, potentially impacting judicial outcomes. As LLMs are increasingly integrated into these critical areas, ensuring factual accuracy is not just a technical requirement but an ethical responsibility. Implementing robust hallucination mitigation techniques helps protect the welfare of users and upholds trust in AI applications. Addressing these ethical implications highlights the need for ongoing research and development to refine LLM outputs, especially in domains where inaccuracies can have profound real-world consequences.

References

[1] Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” arXiv preprint arXiv:2312.10997, 2023.

[2] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in International conference on machine learning, PMLR, 2020, pp. 3929–3938.

[3] N. Stiennon et al., “Learning to summarize with human feedback,” Adv Neural Inf Process Syst, vol. 33, pp. 3008–3021, 2020.

[4] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “Ctrl: A conditional transformer language model for controllable generation,” arXiv preprint arXiv:1909.05858, 2019.

[5] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” Adv Neural Inf Process Syst, vol. 33, pp. 9459–9474, 2020.

[6] S. Borgeaud et al., “Improving language models by retrieving from trillions of tokens,” in International conference on machine learning, PMLR, 2022, pp. 2206–2240.

[7] F. Petroni et al., “Language models as knowledge bases?,” arXiv preprint arXiv:1909.01066, 2019.

[8] L. Yao, C. Mao, and Y. Luo, “KG-BERT: BERT for knowledge graph completion,” arXiv preprint arXiv:1909.03193, 2019.

[9] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, “COMET: Commonsense transformers for automatic knowledge graph construction,” arXiv preprint arXiv:1906.05317, 2019.

[10] L. Ouyang et al., “Training language models to follow instructions with human feedback,” Adv Neural Inf Process Syst, vol. 35, pp. 27730–27744, 2022.

[11] Y. Bai et al., “Training a helpful and harmless assistant with reinforcement learning from human feedback,” arXiv preprint arXiv:2204.05862, 2022.

[12] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” arXiv preprint arXiv:1904.09751, 2019.

[13] S. Zhang et al., “Opt: Open pre-trained transformer language models,” arXiv preprint arXiv:2205.01068, 2022.

[14] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] T. Goyal and G. Durrett, “Annotating and modeling fine-grained factuality in summarization,” arXiv preprint arXiv:2104.04302, 2021.

[16] F. Petroni et al., “Language models as knowledge bases?,” arXiv preprint arXiv:1909.01066, 2019.

[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: a large-scale dataset for fact extraction and VERification,” arXiv preprint arXiv:1803.05355, 2018.

[18] Y. Ovadia et al., “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift,” Adv Neural Inf Process Syst, vol. 32, 2019.