Question:
Provide a literature survey on hallucination mitigation techniques in LLMs?

Answer:
1. Introduction

The problem of hallucinations in Large Language Models (LLMs) has become increasingly significant in recent years, particularly in sensitive applications such as summarizing medical records or financial analysis reports [5]. Hallucination mitigation techniques have emerged as a crucial area of research to address this issue. This literature survey aims to provide an overview of the current state of knowledge on hallucination mitigation techniques in LLMs.

2. Key Concepts

Hallucinations in LLMs refer to the generation of false or nonsensical information [5]. Mitigation techniques can be broadly categorized into four approaches: prompt engineering, model tuning, knowledge retrieval, and inference-time interventions [1]. These methods aim to reduce the likelihood of hallucinations by modifying the input prompts, fine-tuning the models, retrieving relevant knowledge, or applying interventions during inference. For example, Retrieval Augmented Generation (Lewis et al., 2021) and Knowledge Retrieval (Varshney et al., 2023) are notable techniques that have been shown to be effective in reducing hallucinations.

3. Evidence from Context

A comprehensive survey of over 32 techniques designed to combat hallucinations in LLMs has been presented in a recent research paper [1]. The study categorizes these methods based on their approach and highlights key challenges, limitations, and future directions for mitigating hallucinations in AI systems. Another study provides a systematic taxonomy of hallucination mitigation techniques for LLMs, encompassing Vision Language Models (VLMs) [2]. This taxonomy categorizes the methods into model development branches and prompt engineering approaches.

4. Limitations / Open Questions

Despite the progress made in understanding and mitigating hallucinations in LLMs, several challenges remain. The existing literature has adopted a task-oriented taxonomy, which may not be sufficient to capture the complexities of hallucination mitigation [3]. Moreover, no single method universally eliminates hallucinations, pointing to the need for hybrid mitigation pipelines [4]. Furthermore, the field is still lacking in terms of standardization and evaluation metrics, which hinders the comparison and advancement of different techniques.

5. Conclusion

In conclusion, this literature survey provides an overview of the current state of knowledge on hallucination mitigation techniques in LLMs. The evidence from context highlights the importance of prompt engineering, model tuning, knowledge retrieval, and inference-time interventions in reducing hallucinations. However, further research is needed to address the limitations and open questions in this field, including the development of standardized evaluation metrics and the exploration of hybrid mitigation pipelines.

References:
[1] web, https://medium.com/@arghya05/the-battle-against-ai-hallucinations-a-deep-dive-into-mitigation-strategies-for-large-language-7fe8561db5b6
[2] web, https://www.alphaxiv.org/overview/2401.01313v3
[3] web, https://www.mdpi.com/2673-2688/6/10/260
[4] web, https://pmc.ncbi.nlm.nih.gov/articles/PMC12518350/
[5] web, https://arxiv.org/abs/2401.01313