HALO: Hallucination Analysis and Learning Optimization to Empower LLMs with Retrieval-Augmented Context for Guided Clinical Decision Making

Anonymous ACL submission

Abstract

Large language models (LLMs) have significantly advanced natural language processing tasks, yet they are susceptible to generating inaccurate or unreliable responses, a phenomenon known as hallucination. In critical domains such as health and medicine, these hallucinations can pose serious risks. This paper introduces *HALO*, a novel framework designed to enhance the accuracy and reliability of medical question-answering (QA) systems by focusing on the detection and mitigation of hallucinations. Our approach generates multiple variations of a given query using LLMs and retrieves relevant information from external open knowledge bases to enrich the context. We utilize maximum marginal relevance scoring to prioritize the retrieved context, which is then provided to LLMs for answer generation, thereby reducing the risk of hallucinations. The integration of LangChain further streamlines this process, resulting in a notable and robust increase in the accuracy of both open-source and commercial LLMs, such as Llama-3.1 (from 44% to 65%) and ChatGPT (from 56% to 70%). This framework underscores the critical importance of addressing hallucinations in medical QA systems, ultimately improving clinical decision-making and patient care. The open-source *HALO* is available at: `https://anonymous.4open.science/r/HALO-636A/`.

1 Introduction

Large language models (LLMs) have emerged as powerful tools with remarkable capabilities across diverse applications, ranging from natural language understanding to text generation. However, despite their widespread application, LLMs often suffer from a critical issue known as hallucination, where they generate incorrect or unreliable answers. Several factors lead to these hallucinations. The training data for LLMs often includes misconceptions (Lin et al., 2021), social biases (Ladhak et al., 2023), duplication biases (Lee et al., 2021), knowledge cutoffs (Onoe et al., 2022), and a lack of specific domain training (Li et al., 2023b), all of which can lead to the model generating incorrect outputs. During training, the model is often pushed to provide answers beyond its knowledge, known as capability misalignment (Ranzato et al., 2015). Belief misalignment, or sycophancy (Burns et al., 2022), occurs when the model generates answers it believes the user wants to hear, regardless of their correctness, due to training on datasets that emphasize user satisfaction over factual accuracy. The prompts given to the LLM are also very crucial in generating hallucinations. Poorly written prompts (Yu et al., 2022) fail to fully grasp the context of a query, leading to insufficient context attention (Shi et al., 2023). It focuses too much on nearby words, resulting in inaccurate answers. In fields such as health and medicine, where precision and accuracy are paramount (Umapathi et al., 2023), the use of hallucinated answers can be extremely dangerous because they directly impact patient safety, treatment effectiveness, and resource utilization (Rosoł et al., 2023).

Existing solutions to mitigate hallucinations in LLMs have several limitations. For instance, Med-HALT (Umapathi et al., 2023) proposed a set of instructions to help identify hallucinations and prompt LLMs to avoid incorrect answers for reasoning- and memory-based medical questions. However, it did not explore strategies such as integrating external knowledge to enhance the model's performance when it lacks the specific information needed to answer a question. In the financial domain, Kang and Liu (2023) evaluated LLM hallucination mitigation techniques, including few-shot learning. However, these methods were not tested across a wide range of real-world scenarios and demonstrated varying levels of effectiveness. The Hypotheses-to-Theories framework (Zhu et al., 2023b) improved LLM reasoning but was constrained by the model's knowledge and input

capacity. Additionally, approaches addressing "hallucination snowballing" (Zhang et al., 2023) are limited by API constraints, hindering the large-scale exploration of mitigation. These drawbacks underscore the need for more comprehensive and effective hallucination mitigation solutions.

This paper addresses the challenge of improving the accuracy and reliability of medical question answering (QA) systems by proposing *HALO*, a novel framework that enhances LLMs' contextual reasoning abilities. Building upon the Retrieval Augmented Generation (RAG) (Ji et al., 2023; Sharma et al., 2024), our framework integrates innovative techniques tailored specifically for the medical domain. It leverages open-source LangChain[1] to generate multiple queries from the same question, expanding the range of keywords and covering wider aspects of the query. Additionally, we integrate more relevant and trustworthy medical knowledge from PubMed[2] to generate context on the question. PubMed is used for medical knowledge because it provides access to a vast and reputable database of peer-reviewed medical research, ensuring that the information used is accurate, up-to-date, and based on the latest scientific evidence (Umapathi et al., 2023). To create high-quality context, we employ a maximum marginal relevance (Carbonell and Goldstein, 1998) scoring mechanism to balance the trade-off between relevance and diversity of retrieved PubMed data. The top-scored context is then provided to LLMs, along with a few-shot prompting strategy to guide the model's understanding of how to answer the question step by step. Finally, we evaluate the effectiveness of our framework on the MEDMCQA dataset (Pal et al., 2022), which contains 194k high-quality AIIMS & NEET PG entrance exam multiple-choice questions covering 2.4k healthcare topics and 21 medical subjects. Our experimental results demonstrate that *HALO* can boost the trustworthiness and reliability of medical question answering tasks.

Aiming to enhance clinical decision-making and ensure the reliability of information in medical QA scenarios, the key contributions of this work are summarized as follows:

- We propose and implement *HALO*, an open-source framework to enhance the accuracy and reliability of medical QA systems by focusing on hallucination detection and mitigation.
- *HALO* boosts LLMs' capability of providing more contextually grounded responses by incorporating external knowledge with RAG and refining the LLMs' understanding of the query with carefully designed prompt engineering, such as few-shot prompting, chain-of-thought reasoning, and appropriate prioritization.
- Extensive experiments on 194k medical questions across 21 subjects with both commercial LLM (ChatGPT) and open-source LLMs (Llama-3.1 and Mistral) demonstrate the generalizability and adaptability of the proposed *HALO*. Additionally, a case study on neurological disorders highlights *HALO*'s effectiveness in addressing specific disease-related queries.

2 Related Work

Hallucinations in LLMs have been identified as a significant concern across various domains, such as medicine (Umapathi et al., 2023; Pal et al., 2022; Ponce-López, 2024; Rehman et al., 2023), finance (Kang and Liu, 2023), education (Sharma et al., 2024), and law (Yu et al., 2022). The potential consequences of these hallucinations, particularly in critical domains like healthcare, underscore the importance of addressing this issue.

Prompt engineering and instruction techniques like few-shot learning (Kang and Liu, 2023; Yu et al., 2022) and chain-of-thought (CoT) (Yu et al., 2022; Braunschweiler et al., 2023) prompts can guide LLMs to improve task performance and reasoning accuracy. Few-shot learning uses a few examples to enhance understanding, while CoT prompts encourage step-by-step reasoning to reduce errors. However, these methods increase complexity and computational cost, and their success is highly dependent on carefully designed prompts. Kang and Liu (2023) showed limitations of these methods due to heavy reliance on the prompt quality. Additionally, Yu et al. (2022) reported significant performance variability across test sets, and Braunschweiler et al. (2023) found the effectiveness of prompts was inconsistent and responses were often verbose.

Recent efforts to integrate external knowledge sources into LLMs have shown promise in improving factual accuracy. Approaches such as retrieval-augmented generation (RAG) (Sharma et al., 2024; Thorne et al., 2018) and prompt-based tool learning (Kang and Liu, 2023) leverage external knowledge to enhance the reliability of generated responses. While these methods have demonstrated

[1] https://github.com/langchain-ai/langchain
[2] https://pubmed.ncbi.nlm.nih.gov/

success, they require carefully designed prompts and significant computational resources. Additionally, their effectiveness is often constrained by their domain-specific nature and reliance on up-to-date data sources, which may not comprehensively address general real-world scenarios.

Validation and explanation techniques like Verification-of-Choice (VoC) (Li et al., 2023a; Zhu et al., 2023b) and self-reflection (Zhang et al., 2023; Zhu et al., 2023b) can enhance reasoning and self-correction in LLMs to mitigate hallucinations. VoC generates and verifies explanations for each choice, while self-reflection allows models to correct their errors. However, these methods are computationally expensive and their success is highly dependent on the initial accuracy of the model. Their effectiveness is constrained by the model's initial error rate and requires a robust base model, which is also limited by context length restrictions.

To improve the reliability and accuracy of LLMs in specific domains, various domain-specific prompt tuning and model fine-tuning techniques have been proposed. For instance, Zhu et al. (2023a) evaluated fine-tuning strategies for Chinese biomedical tasks, while Jeong et al. (2024) fine-tuned LLMs using biomedical knowledge to improve factual accuracy. However, these approaches rely on pre-existing knowledge with fixed timestamps, limiting access to the most recent information and reducing their generalizability to evolving real-world scenarios.

While existing techniques for mitigating hallucinations in LLMs have made significant progress, each has notable limitations. Prompt engineering and instruction methods, such as few-shot learning and CoT prompts, are highly dependent on prompt quality and introduce additional complexity. External knowledge integration methods like RAG demand substantial computational resources, while validation techniques, including VoC and self-reflection, are resource-intensive and rely on the initial accuracy of the model. Domain-specific fine-tuning, although effective, often lacks generalizability and requires considerable resources. We propose *HALO* to harness the strengths of these hallucination mitigation methods while reducing their limitations. *HALO* improves LLM reliability performance by retrieving and prioritizing relevant information from domain-specific datasets, applying a maximum marginal relevance score, and integrating this context with CoT reasoning and few-shot prompting. This approach reduces reliance on prompt quality while efficiently leveraging external knowledge without the need for excessive computational resources, as the integration of external knowledge with CoT prompts does not require LLM fine-tuning.

3 Methodology

3.1 *HALO* Framework Overview

To mitigate hallucination in LLMs during medical decision-making, we introduce *HALO*, which integrates many optimization techniques, including query expansion, RAG, MMR-based document selection, few-shot learning, and CoT. As depicted in Figure 1, *HALO* comprises three core components: multiquery generation, contextual knowledge integration, and prompt engineering. An example of utilizing HALO is in the Section A.1.

First, to enhance the contextual understanding (Andriopoulos and Pouwelse, 2023) of each query, we collect supplementary information from PubMed articles and Wikipedia. Specifically, by generating multiple queries for each question and retrieving information from PubMed, we obtain diverse perspectives and keywords, facilitating a more nuanced understanding of the query. Second, we leverage LangChain for document retrieval (Chase, 2023), utilizing embeddings generated by the Instructor-XL model from Hugging Face (Su et al., 2022). This optimization improves the efficiency of retrieval and comparison, ensuring access to relevant documents. Moreover, we employ the maximum marginal relevance technique to enhance the diversity of the retrieved documents, reducing redundancy and improving the overall quality of information. This ranking process is integrated to refine retrieval results and prioritize the most relevant documents, thereby enhancing the relevance of responses. Additionally, we implement few-shot and CoT prompting to improve the model's understanding and coherence in responding to the questions, enabling it to generalize more effectively from limited examples and thereby improving performance.

3.2 Multiquery Generation

Multiquery involves generating multiple queries Q for a single query q, broadening the scope of perspectives and keywords explored compared to relying solely on a single question and keyword. This approach, exemplified by LangChain, facilitates a more exhaustive exploration of the subject matter, thereby enhancing the effectiveness of in-

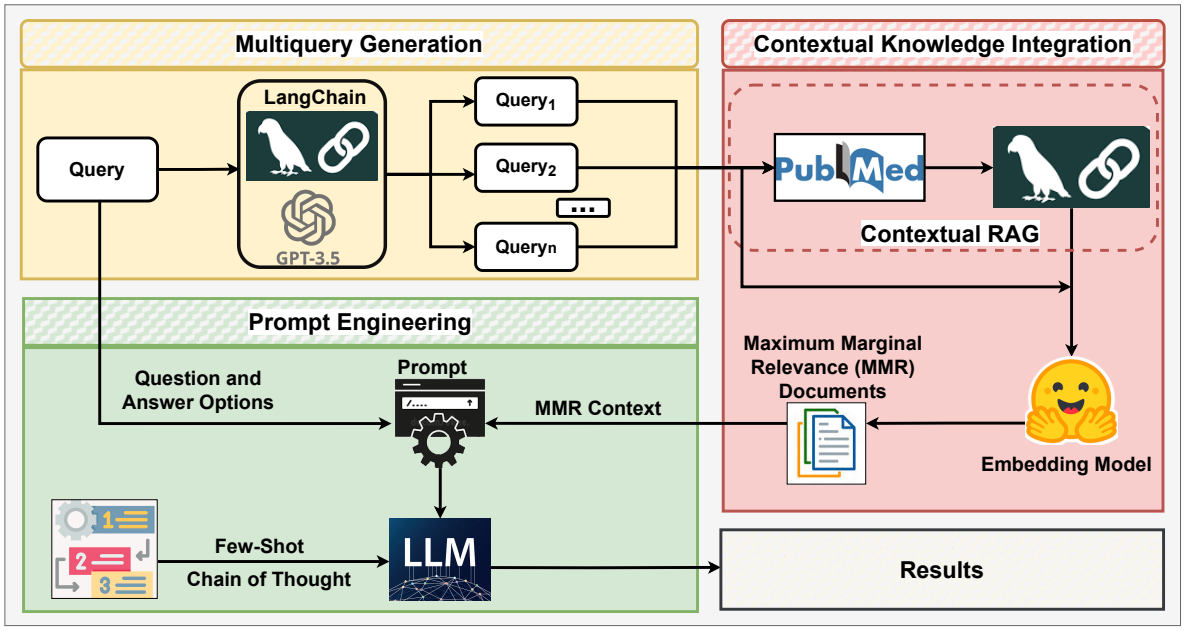

Figure 1: *HALO* framework. *HALO* comprises three key components: multiquery generation, contextual knowledge integration through the maximum marginal relevance-optimized RAG, and few-shot and CoT-based prompt.

formation retrieval. Creating multiple queries for a single question becomes crucial as it expands the search parameters, thereby increasing the likelihood of capturing relevant context and information. Given any LLM's challenges in accurately responding to intricate medical queries due to constraints in its training data and domain-specific understanding, generating multiple queries enables the extraction of diverse perspectives and keywords from various sources. This process significantly augments the probability of uncovering pertinent information to effectively tackle the question at hand.

In LangChain, multiquery generation (Chase, 2023) uses an LLM to generate multiple alternative phrasings of a query, expanding the scope of document retrieval. The user integrates the LLM via an API (in our case, ChatGPT-3.5), and LangChain leverages its *runnable interface*, a standardized framework for executing components like LLMs, to carry out the process. A prompt template is predefined to control the structure of the task, ensuring the generated queries maintain the original query's semantic meaning while altering the wording. After the LLM generates these query variations, a parser called *LineListOutputParser* is used to break down the output into a list of distinct queries, each treated as a separate input for document retrieval. The number and diversity of the queries are customizable by adjusting the prompt or retriever parameters, providing flexibility to generate more focused or diverse query sets. Additionally, the LLM used for this task can easily be changed based on the user's preference, allowing flexibility for different models or integrations. For instance, consider a medical query about a drug like Remifentanyl. A single query might focus on its mechanism of action, while multiple queries allow *HALO* to explore various aspects like pharmacokinetics, side effects, and clinical applications(see the gray block below). This broadens insights from diverse sources, offering a more comprehensive understanding.

Multiquery Generation Example

- **Original Single Query** q:
  - q: What are the characteristics of Remifentanyl?
- **Generated Multiple Queries** Q=$\{Q_1, Q_2, ..., Q_n\}$:
  - Q_1: What are the distinguishing features of Remifentanil?
  - Q_2: How does Remifentanil differ from other opioids, such as Alfentanil?
  - Q_n: What are the unique characteristics of Remifentanil, particularly regarding metabolism, half-life, potency, and dosage adjustments in hepatic and renal disease?

3.3 Contextual Knowledge Integration

Retrieval augmented generation (RAG) is a novel framework that combines retrieval and generation processes to produce more accurate and contextually relevant responses. It leverages pre-existing knowledge sources, such as databases or documents, to enhance the generation of responses by incorporating retrieved information into the generated text. *HALO* adopts RAG to the next level

by integrating it with a multiquery approach. This advanced integration involves employing a multiquery RAG approach, optimizing document retrieval from PubMed by generating diverse queries for each user input. By generating multiple queries, each targeting different aspects of the input question, *HALO* can retrieve a broader range of relevant documents from PubMed. By tapping into PubMed's vast repository of peer-reviewed articles and research papers, *HALO* enriches its comprehension of medical topics.

RAG is capable of retrieving multiple pieces of information, but the quality and relevance of these results can vary significantly (Yu et al., 2024). *HALO* incorporates a maximum marginal relevance (MMR) scheme to mitigate redundancy and prioritize diversity among the retrieved relevant results R. Note that *HALO* first expands a single query q into multiple queries Q. Instead of using the traditional single-query MMR, We design a multiquery MMR as shown in Equation 1. By evaluating the relevance of each document $D_i \in R$ based on the new information it contributes compared to previous results, the multiquery MMR effectively reduces duplication while preserving query relevance.

$$\text{MMR} \stackrel{\text{def}}{=} \arg\max_{D_i \in R \setminus S} \left[\lambda \cdot \frac{1}{n} \sum_{k=1}^{n} \text{Rel}(D_i, Q_k) - (1-\lambda) \cdot \max_{D_j \in S} \text{Sim}(D_i, D_j) \right] \tag{1}$$

where D_i is a document in the set of retrieved results R from PubMed; S is the set of already-selected documents; Q_k is the k-th query in the generated multiple queries Q; $\text{Rel}(D_i, Q_k)$ is the relevance of document D_i to the k-th query; $\text{Sim}(D_i, D_j)$ is the similarity between two documents D_i and D_j; and λ is a parameter that controls the balance between relevance and diversity.

To accurately calculate $\text{Rel}(D_i, Q_k)$ and $\text{Sim}(D_i, D_j)$ in Equation 1, we generate embeddings for both the retrieved external documents R and the medical queries Q, along with their associated answer options. To accomplish this, *HALO* utilizes the Hugging Face Instructor-XL model (Su et al., 2022), an instruction-finetuned text embedding model designed to create task-specific embeddings for diverse applications. This allows *HALO* to accurately select and create the most contextually relevant information, retrieved through the multiquery, MMR-optimized RAG process, for the given question and its associated answer choices.

This careful contextual knowledge integration ensures that the responses of *HALO* are comprehensive, avoiding unnecessary repetition and leading to higher-quality outputs. The combination of the multiquery RAG approach with MMR optimization significantly improves the efficacy of *HALO* retrieval and ranking processes, resulting in reliable responses from our framework.

3.4 Prompt Engineering

HALO leverages few-shot learning and CoT reasoning to refine context, ensuring coherent and accurate responses from LLMs. Few-shot prompting helps LLMs generalize from limited examples, enhancing performance with minimal training data. CoT reasoning breaks down complex medical questions into smaller parts, encouraging step-by-step, human-like reasoning. This approach improves LLMs' handling of nuanced queries, resulting in more accurate, detailed answers. By combining few-shot examples with CoT reasoning, *HALO* guides LLMs to produce well-structured, contextually relevant, and logical responses.

4 Evaluation

4.1 Evaluation Metrics

The evaluation of hallucination detection can be approached in various ways, depending on the specific task, including fact-based metrics, classifier-based metrics, QA-based metrics, uncertainty estimation, and prompting-based metrics (Huang et al., 2023). **Fact-based metrics** (Thorne et al., 2018), checks if the facts in the model's responses match the facts in the original content. It is like comparing two lists to see how many items they have in common. While this technique is useful, it often misses the nuances and context-specific details with poor correlation. **Classifier-based metrics** (Jeong et al., 2024) involve using natural language inference models that can determine if the model's response logically follows from the original content. Consequently, classifier-based metrics require extensive training data to be effective. **QA-based metrics** (Sharma et al., 2024) involves asking questions based on the original content and comparing the model's answers to the correct ones. **Uncertainty estimation** technique measures how confident the model is in its answers, this type of detection is used in zero-resource settings (Huang et al., 2023). If the model is unsure, there is a higher chance it might be making things up. This method can be

useful but often needs to be combined with other metrics for comprehensive evaluation. **Prompting-based metrics** (Yu et al., 2022; Zhu et al., 2023a) uses specific prompts to guide the model in evaluating its own responses for accuracy. The model is given examples or specific instructions to check if its answers are faithful to the source content. This method can enhance the accuracy of responses but depends heavily on the quality of the prompts.

HALO was evaluated by QA-based metrics to measure understanding and response accuracy, making it ideal for the MedMCQA dataset. We employed few-shot prompts with answer format examples and cleared history after each question to prevent prior influence. This approach helped identify hallucinations by ensuring responses were directly comparable to expected answers.

4.2 MedMCQA Dataset

The MedMCQA dataset is designed for real-world medical entrance exam, making it ideal for training and evaluating QA models in the medical domain. It reflects the types of questions and challenges medical practitioners face in their practice.

The MedMCQA dataset consists of multiple-choice questions (MCQs), which are a common format in medical exams (Pal et al., 2022; Umapathi et al., 2023). This format requires models to identify the correct answer, reason and differentiate it from the other distractors, adding an additional layer of complexity to the task. The large size of the dataset, with over 194,000 MCQs, provides a rich and diverse set of examples for training and evaluation. The dataset is divided with approximately 53% for DEV and 47% for TEST categories, with medical subjects and topics evenly distributed across both sets for balanced evaluation. This helps in capturing a wide range of medical knowledge, scenarios, and variations in question styles, ensuring that models are exposed to a comprehensive set of challenges. The questions in the MedMCQA dataset are created by human experts, ensuring their quality and reliability. This expert curation guarantees that the dataset represents accurate and credible medical knowledge, enhancing the validity of the evaluation and the trustworthiness of the results. MedMCQA's 21 medical subjects demonstrate comprehensive coverage across a wide range of knowledge areas, encompassing 2,400 internal medical topics. The distribution of subjects in the dataset indicates a fair allocation of focus across different areas of medical expertise.

4.3 LLM Models

We evaluated the performance of *HALO* on both commercial and open-source LLMs, including ChatGPT-3.5-16K, Llama-3.1 8B, and Mistral 7B.

4.3.1 ChatGPT-3.5-16K

We have chosen to use the ChatGPT-3.5-16K model in our evaluation for several reasons. First, this model has a maximum context length of 16,384 tokens (OpenAI, 2023), which allows it to process and understand larger inputs and provide more detailed outputs, making it well-suited for handling the intricacies of medical knowledge and providing contextually relevant responses (Rosoł et al., 2023). The ChatGPT-3.5-16K model has also been specifically fine-tuned for instruction-based tasks (OpenAI, 2023; AI, 2023), making it well-suited for enhancing medical question answering. By utilizing the ChatGPT-3.5-16K model, we aim to leverage its advanced capabilities and specialized training in providing comprehensive information to medical professionals and users seeking medical insights.

4.3.2 Llama-3.1 8B

Llama-3.1 8B is an open-source powerful and adaptable NLP model, optimized for a wide range of tasks with high efficiency and accuracy. As one of the most recently released Llama models, which has been optimized for a wide variety of instructions, it excels at following instructions precisely (Raza, 2024). As an open-source model, it offers transparency and flexibility, allowing us to customize and optimize it for our specific needs.

4.3.3 Mistral 7B

Mistral 7B (Jiang et al., 2023) is an efficient, open-source language model that excels in complex NLP tasks with a smaller parameter size, making it lightweight and fast. It offers a balance between computational efficiency and performance in resource-constrained environments. After evaluating ChatGPT-3.5-16k and Llama-3.1 8B for medical QA tasks, we selected Mistral 7B to deliver robust performance even in environments with limited resources.

4.4 Results on 21 Medical Subjects

Using the QA-based evaluation metric for hallucination detection, which aligns with Faithfulness Hallucination Detection as detailed in (Huang et al., 2023; Ji et al., 2023), we first evaluated the performance of BERT-based (Devlin, 2018) and Codex-based (Chen et al., 2021) language models on the

MedMCQA dataset. As the dataset is split by the authors into two subsets, DEV (53%) and TEST (47%), we report the accuracies for both. As shown in Table 1, BERT Base, BioBERT, SciBERT, and PubMedBERT exhibit very low accuracies on both the DEV and TEST subsets (Pal et al., 2022). In contrast, Codex 5-shot CoT and MQ-SequenceBERT (Liévin et al., 2023) demonstrate improvements due to advanced techniques (Umapathi et al., 2023; Ponce-López, 2024), such as few-shot learning and sequence-based enhancements.

Models	DEV Acc.	TEST Acc.
BERT Base	0.35	0.33
BioBERT	0.38	0.37
SciBERT	0.39	0.39
PubMedBERT	0.40	0.41
Codex 5-shot CoT	0.63	0.60
MQ-SequenceBERT	0.68	0.60
ChatGPT-3.5	0.56	0.44
Llama-3.1 8B	0.44	0.38
Mistral 7B	0.37	0.28
***HALO* + ChatGPT 3.5**	**0.70**	**0.65**
***HALO* + Llama-3.1 8B**	**0.66**	**0.60**
***HALO* + Mistral 7B**	**0.58**	**0.55**

Table 1: Performance of LLMs on the MedMCQA DEV (53%) and MedMCQA TEST (47%) subsets

ChatGPT-3.5-turbo-16K achieved only 56% accuracy—significantly lower than the typical 90% or higher accuracy achieved by human medical professionals (Umapathi et al., 2023). After applying *HALO*, accuracy improved from 44% to 65% on the TEST subset (and from 56% to 70% on the DEV subset). Similarly, the accuracy of the Llama-3.1 8B model improved from 38% to 60% on the TEST subset (and from 44% to 66% on DEV subset). The Mistral 7B model also showed a boost in accuracy, increasing from 28% to 55% on the TEST subset (and 37% to 58% on DEV subset), though it still fell short of expectations. Notably, our *HALO* framework achieved the highest accuracies—70% on the DEV set and 65% on the TEST set—demonstrating its superior performance and effectiveness in addressing medical QA tasks.

To evaluate the generalizability of *HALO* across different medical subjects, we evaluated its performance on all 21 subjects (e.g., anatomy and dental) covered by the MedMCQA dataset. Figure 2 compares the accuracy of LLMs and *HALO*, showing consistent improvements by ChatGPT-3.5, Llama-3.1 8B, and Mistral 7B. These results highlight *HALO*'s effectiveness in generating and prioritizing relevant context, demonstrating its robustness and superiority over strong baselines.

4.5 Case Study on Neurological Disorders

Neurological disorders present diagnostic and treatment challenges due to their complex and variable symptoms, necessitating broad medical knowledge for informed decision-making. We began by filtering relevant questions from the MedMCQA dataset, which is organized by subjects like Medicine, Psychiatry, General, and Preventive Medicine using keywords such as "aphasia", "dementia", and "mild cognitive impairment". Then we applied *HALO* on these questions. The sample question with the multiquery generation is shown in the following gray box, where multiple queries are generated to explore different aspects of the aphasia diagnosis, enhancing the model's ability to understand and retrieve relevant information.

Example Question and Multiquery Generation

Aphasia Diagnosis Question q:
A man comes with aphasia. He is unable to name things and repetition is poor. However, comprehension, fluency, and understanding of written words are unaffected. He is probably suffering from:
A. Anomic aphasia
B. Broca's aphasia
C. Transcortical sensory aphasia
D. Conduction aphasia
E. Global aphasia

Multiquery Generation
Q_1: What are the common types of aphasia, and how do their symptoms differ?
Q_2: What type of aphasia is characterized by the inability to name things and poor repetition, but with unaffected comprehension, fluency, and understanding of written words?
Q_3: How do different aphasia types impact a patient's ability to communicate, and what are the key signs to look for?

The CoT reasoning of the aphasia diagnosis example question is outlined in the above box, detailing the logical steps to reach the correct answer taken by the LLM. The CoT process begins with identifying the main subject (a patient with aphasia) and understanding the objective by determining the type of aphasia. Next, each option is analyzed based on the described symptoms. The MMR content extracted from PubMed is reviewed by the ChatGPT-3.5 to compare the characteristics of different aphasia types. After this analysis, it is determined that option "D. Conduction aphasia", is the correct choice, as it aligns best with the patient's symptoms of poor repetition, difficulty with naming, but preserved fluency and comprehension.

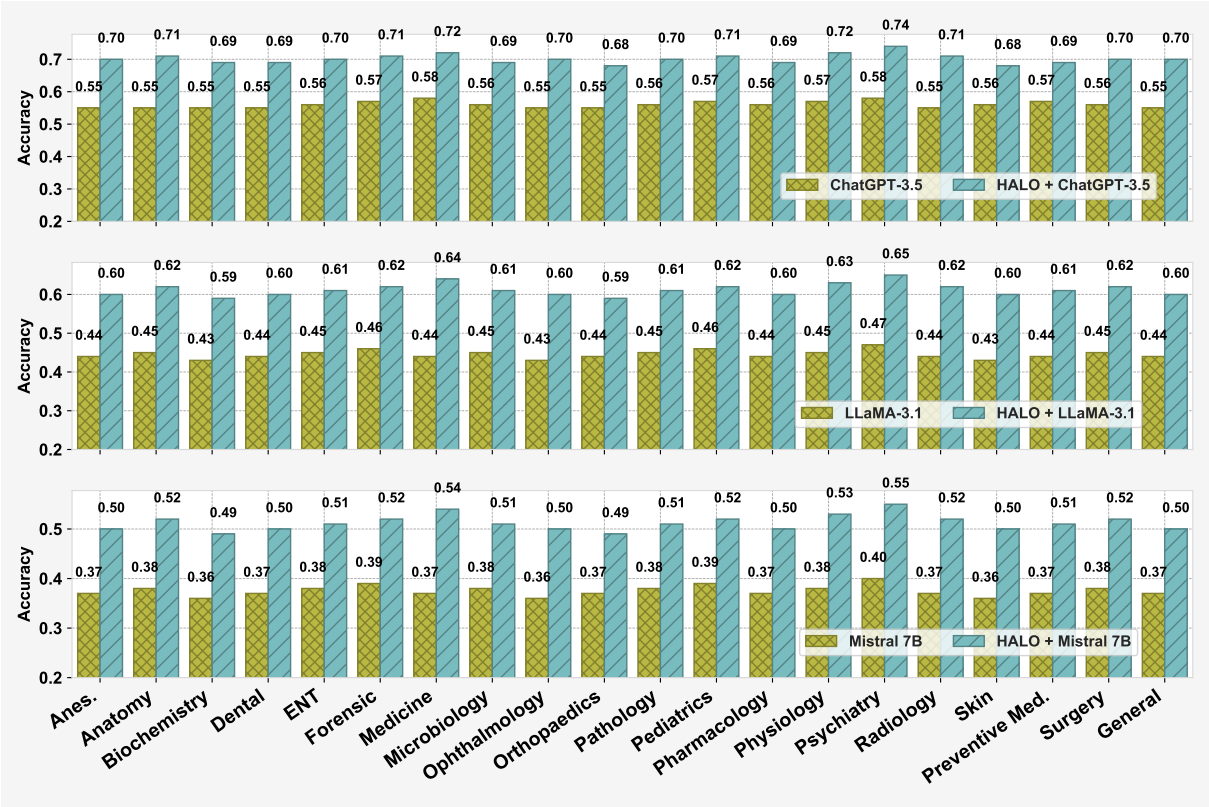

Figure 2: Comparative analysis of the LLMs (ChatGPT-3.5, Llama-3.1 8B, Mistral 7B) and the *HALO* framework on LLMs accuracies across 21 subjects in MedMCQA dataset

Finally, ChatGPT-3.5 generates the output by adhering to these CoT steps and utilizing the content.

Chain of Thought (CoT) Reasoning

1. Identify the main subject or keyword in the question: The subject is a patient with aphasia exhibiting poor repetition and naming ability but preserved fluency and comprehension.
2. Understand the relationship between the subject and the question asked: The goal is to identify which type of aphasia corresponds to these specific symptoms.
3. Extract relevant information from the context:

Contextual documents/information retrieved.
4. Analyze each option:
A. Anomic aphasia: Naming difficulty fits, but repetition is unaffected, so this is incorrect.
B. C. D. E. ...
5. Provide the answer: The correct answer is "D" Conduction aphasia, as it aligns perfectly with the described symptoms.

Models	w/o *HALO* Acc.	w/ *HALO* Acc.
ChatGPT-3.5-16K	0.49	**0.65**
Llama-3.1 8B	0.40	**0.58**
Mistral 7B	0.34	**0.48**

Table 2: The performance of *HALO* for neurological disorder related questions from MedMCQA dataset

4.5.1 Case Study Results

HALO improved baseline accuracies of 49%, 40%, and 34% for ChatGPT-3.5, Llama-3.1 8B, and Mistral 7B to 65%, 58%, and 48%, demonstrating reduced hallucinations in neurological decisions.

5 Conclusion

This paper introduces *HALO*, a novel open-source framework designed to identify and mitigate hallucinations in LLMs, with a specific emphasis on the medical domain. By incorporating advanced techniques such as multiquery generation, contextual knowledge integration, and prompt engineering, *HALO* demonstrates significant improvements in mitigating hallucinations. To thoroughly evaluate the framework, we evaluated *HALO* across multiple commercial and open-source LLMs, including ChatGPT-3.5, Llama-3.1 8B, and Mistral 7B, as well as six other BERT- and Codex-based models. These evaluations on the MedMCQA dataset highlighted notable enhancements in both accuracy and reliability when *HALO* was integrated, achieving consistent improvements across all 21 medical subjects. In addition, a case study on neurological disorders further underscored *HALO*'s effectiveness in addressing specific disease-related queries.

6 Limitations

While PubMed provides highly reputable information, *HALO* could be improved by integrating multimodal inputs, such as videos, images, and diverse journal sources, to broaden its scope. Medical decision-making often relies on more than just textual data, so incorporating other forms of media could enhance the richness and accuracy of the context provided. Currently, *HALO* primarily relies on text-based sources to ensure the validity and reliability of the information it retrieves. Although the evaluation focuses on reducing hallucinations through chain-of-thought reasoning, which is essential in the medical domain, this approach may limit its adaptability to free-form question-answering or dynamic contexts in other fields.

7 Ethical Implications

The ethical considerations of integration of *HALO* for clinical decision making include data privacy, bias, transparency and its impact on the LLMs used. Privacy concerns, particularly around the use of sensitive medical data, were mitigated by strictly using publicly available sources like PubMed. The MedMCQA dataset, which consists of a diverse range of high quality, expertly curated multiple choice questions across various medical subjects, ensures that *HALO* is evaluated in an unbiased manner. Transparency is crucial, and our framework enhances it by incorporating a chain of thought processes, guiding the model through logical steps to improve accuracy and decision making clarity. *HALO*'s processes are optimized by using selective data retrieval, filtering redundant information, and employing efficient prompt techniques to reduce computational load, energy consumption and reduced environmental impact. Finally, *HALO* is designed to assist, not replace, medical professionals, with clear safeguards ensuring that AI-driven recommendations always complement human judgment, preserving the critical role of healthcare providers. By being open-source, *HALO* will encourage engagement from the LLM community, allowing for collaborative improvements, shared learnings, and collective problem-solving around common challenges such as hallucination.

References

Anakin AI. 2023. What is gpt-3.5 turbo instruct?

Konstantinos Andriopoulos and Johan Pouwelse. 2023. Augmenting llms with knowledge: A survey on hallucination prevention. *arXiv preprint arXiv:2309.16459*.

Norbert Braunschweiler, Rama Doddipatla, Simon Keizer, and Svetlana Stoyanchev. 2023. Evaluating large language models for document-grounded response generation in information-seeking dialogues. *arXiv preprint arXiv:2309.11838*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Harrison Chase. 2023. Multi query retriever. Online. Available at https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/MultiQueryRetriever/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024. Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548*.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Qiang Li, Xiaoyan Yang, Haowen Wang, Qin Wang, Lei Liu, Junjie Wang, Yang Zhang, Mingyuan Chu, Sen Hu, Yicheng Chen, et al. 2023a. From beginner to expert: Modeling medical knowledge into general llms. *arXiv preprint arXiv:2312.01040*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning*, pages 20950–20977. PMLR.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What lms know about unseen entities. *arXiv preprint arXiv:2205.02832*.

OpenAI. 2023. Gpt-3.5 turbo.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Víctor Ponce-López. 2024. Large language models for multi-choice question classification of medical subjects. *arXiv preprint arXiv:2403.14582*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Wamiq Raza. 2024. Leveraging meta's llama 3.1 for efficient question-answering on google colab.

Tohida Rehman, Ronit Mandal, Abhishek Agarwal, and Debarshi Kumar Sanyal. 2023. Hallucination reduction in long input text summarization. *arXiv preprint arXiv:2309.16781*.

Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. 2023. Evaluation of the performance of gpt-3.5 and gpt-4 on the medical final examination. *medRxiv*, pages 2023–06.

Sanat Sharma, David Seunghyun Yoon, Franck Dernoncourt, Dewang Sultania, Karishma Bagga, Mengjiao Zhang, Trung Bui, and Varun Kotte. 2024. Retrieval augmented generation for domain-specific question answering. *arXiv preprint arXiv:2404.14760*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Victoria Lin, Noah A Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2023. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv:2405.07437*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

W Zhu, X Wang, H Zheng, M Chen, and B Tang. 2023a. Promptcblue: A chinese prompt tuning benchmark for the medical domain. arxiv e-prints. *arXiv preprint arXiv:2310.14151*.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023b. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*.

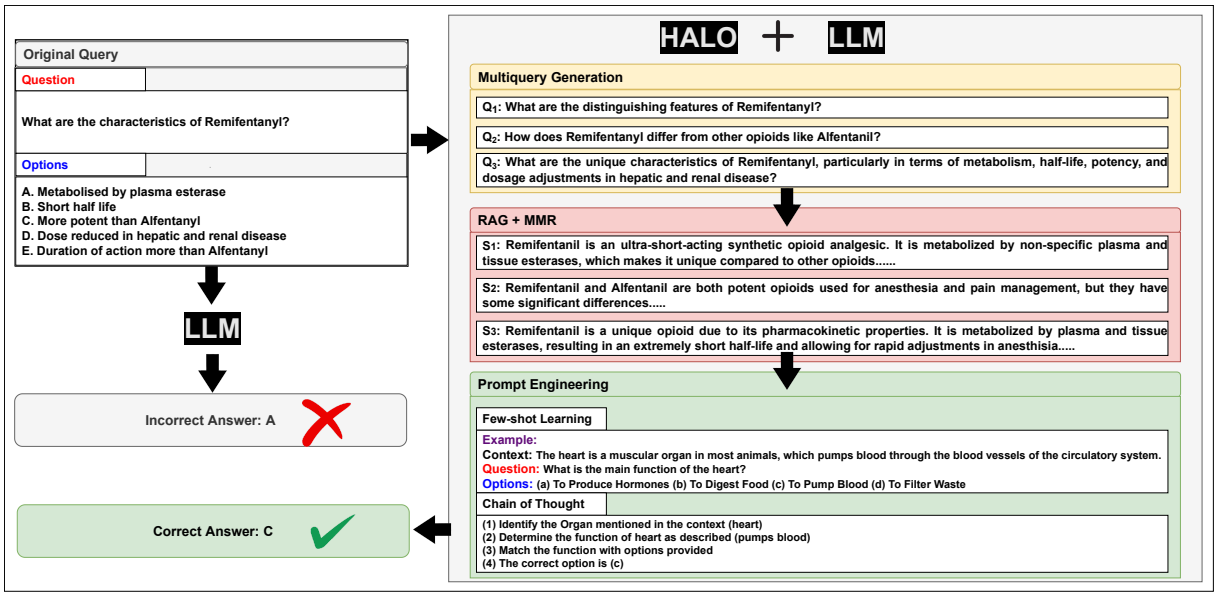

Figure 3: Example of *HALO* framework applied to a medical question from the MedMCQA dataset

A Appendix

A.1 An Example of How *HALO* Works

Figure 3 illustrates the *HALO* framework applied to enhance the accuracy of medical question answering for a sample query. In this example, a question regarding the characteristics of Remifentanyl is posed, and the baseline model incorrectly selects the answer “A" instead of the correct answer “C". Using our framework, three alternative queries (highlighted in yellow) are generated through the LangChain Multiquery Retriever to expand the search scope. Relevant information is then retrieved from PubMed (highlighted in red), enriching the context and providing diverse perspectives. The framework further incorporates few-shot learning and CoT reasoning by offering structured prompts and examples (highlighted in green) to guide the model towards more accurate responses.