
RESEARCH PROPOSAL

AI800 – Research Methodology in Computer Science

Mitigating Hallucinations in Large Language Models

Using Retrieval-Augmented Generation: A Framework for Domain-Specific Factual Accuracy

Student Name:

Rayn Suliman

Course Code: AI800

Lecturer: Dr. Mohammed Anbar

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, enabling breakthroughs in machine translation, question answering, summarization, and conversational AI. However, a critical limitation—hallucination—poses a significant barrier to their deployment in high-stakes, domain-specific environments such as healthcare, law, and education. Hallucination refers to the generation of text that appears linguistically fluent and confident yet is factually incorrect, fabricated, or unsupported by external knowledge. This research proposes a novel framework that leverages Retrieval-Augmented Generation (RAG) to systematically reduce hallucinations in LLMs for domain-specific question-answering tasks.

This research aims to investigate and classify different types of hallucinations that occur in various contexts, and to develop a more advanced pipeline for retrieving and ranking information. The pipeline will combine dense and sparse retrieval methods with a re-ranker that takes into account the faithfulness of the information. The goal is to test this framework using standard benchmarks, such as TruthfulQA and MedQA, as well as datasets that are specific to certain domains. By doing so, the study hopes to achieve a significant reduction in hallucination rates - by at least 30% - compared to current large language models. Additionally, the research will result in a published taxonomy of hallucination patterns that are specific to certain domains. This work is important for ensuring that large language models are used responsibly and for advancing the field of trustworthy artificial intelligence. The expected outcome will not only improve the factual accuracy of these models but also contribute to a better understanding of how they can be used in a reliable and trustworthy manner.

Keywords: Large Language Models, Hallucination, Retrieval-Augmented Generation, Natural Language Processing, Factual Accuracy, Domain-Specific AI

CHAPTER 1

Introduction, Background & Research Design

Introduction, Background & Research Design

1.1 Introduction

The rapid evolution of artificial intelligence has ushered in a new era of natural language processing (NLP), with Large Language Models (LLMs) emerging as the cornerstone of modern AI-driven communication systems. Models such as OpenAI's GPT-4, Meta's Llama 3, Google's Gemini, and Mistral have demonstrated an extraordinary capacity to generate human-like text across a vast range of tasks, including question answering, text summarization, code generation, and multi-turn dialogue. Their widespread adoption across industries—from customer service and healthcare diagnosis support to legal research and educational tutoring—reflects their transformative potential.

These big language models have some major limitations. They rely on what they learned during training, which can be outdated or just plain wrong. When they're asked something that their training data doesn't cover well, they might give a confident-sounding answer that's actually false. This is called hallucination, and it can be really bad for people's trust in the model. In some cases, like in healthcare or law, it can even cause serious harm. For example, if a language model is used to help doctors make decisions and it suggests the wrong medication dosage, that's a big problem. Or if a legal assistant cites a law that doesn't exist, that's not just wrong, it's also unhelpful. These kinds of mistakes show that we still have a lot of work to do to make AI reliable.

This research addresses the hallucination problem by investigating Retrieval-Augmented Generation (RAG) as a principled solution. RAG supplements LLMs with a retrieval component that dynamically fetches relevant documents from an external knowledge base at inference time, grounding generation in verifiable evidence. This approach reduces the model's dependence on potentially flawed parametric memory and has shown promise in open-domain question answering. However, existing RAG systems face challenges including retrieval noise, faithfulness failures, and performance degradation in specialized

domains. This study proposes an enhanced RAG framework specifically designed to address these shortcomings in domain-specific applications.

1.2 Background

The trajectory of language modeling traces back to early statistical approaches—n-gram models, Hidden Markov Models (HMMs), and later, recurrent neural networks (RNNs) such as LSTMs and GRUs. The pivotal shift occurred with the introduction of the Transformer architecture by Vaswani et al. (2017), which enabled parallelizable attention-based processing of sequential data at unprecedented scale. Pre-trained transformer models such as BERT (Devlin et al., 2019) established the paradigm of pretraining on massive corpora followed by fine-tuning on downstream tasks, while GPT-2 (Radford et al., 2019) demonstrated the viability of autoregressive language modeling at scale.

When we made these systems really big, with billions and billions of pieces of information, they started to do things that were really different from smaller systems. They could learn from just a few examples, think in a chain of thoughts, and follow instructions. One system, called GPT-3, started a lot of research into how these systems can learn from the context they are given, and other systems like InstructGPT, ChatGPT, and GPT-4 showed they could be used for many things without needing to be specially trained for each task. Other companies have also made their own systems, like Llama, PaLM 2, and Gemini, which has made the group of strong language systems even bigger and more varied.

Concurrent with this scaling trend, researchers began documenting systematic failure modes. Hallucination—defined as the generation of content that is unfaithful to the source (extrinsic) or internally inconsistent (intrinsic)—has been observed across tasks including summarization, translation, and question answering (Ji et al., 2023). Several contributing factors have been identified: exposure bias during training, overconfident token probability distributions, conflicting training data, and the model's inability to express calibrated uncertainty. These failures are not merely theoretical; documented cases include medical advice errors, legal misquotations, and fabricated academic citations.

Retrieval-Augmented Generation is a new way of building models that was introduced by Lewis and his team in 2020. It's a promising approach that combines a part of the model that

retrieves information with a part that generates text. This allows the model to use real-time evidence when generating text, which is a big advantage. Some newer versions of this approach, like FiD, Atlas, REALM, and Self-RAG, have improved its performance in tasks that need a lot of knowledge. However, there are still some challenges to overcome, such as making sure the retrieved information is accurate, using the retrieved passages in a faithful way, and adapting to areas with limited training data.

1.3 Problem Statement

The deployment of Large Language Models in domain-specific, knowledge-intensive applications is severely hampered by the hallucination problem. Despite achieving high performance on general benchmarks, LLMs frequently generate factually incorrect, fabricated, or misleading information when tasked with queries requiring precise, current, or domain-specific knowledge. This limitation is particularly dangerous in high-stakes fields such as healthcare, law, and finance, where factual accuracy is not merely preferable but essential for user safety and legal compliance.

Current methods for reducing hallucinations have some limitations. These methods include fine-tuning models on specific datasets, using reinforcement learning from human feedback, and chain-of-thought prompting. Fine-tuning uses a lot of resources and doesn't completely stop hallucinations. Reinforcement learning from human feedback is better at improving tone and helpfulness, but it's not as good at ensuring factual accuracy. Chain-of-thought prompting can actually make things worse by spreading reasoning errors. The standard RAG approach has some conceptual flaws. For example, the retrieval part might bring up passages that are not relevant or even contradict each other. The generative model might not properly use the retrieved content, essentially ignoring the context. Moreover, there's not enough evaluation on specific domains that reflect how professionals use these models in real-world scenarios. This means we need to find better ways to address these issues and make these models more reliable and accurate. We need to think about how to improve these methods, so they can better handle real-world use cases. This could involve making the retrieval component more accurate, ensuring the generative model uses the context correctly, and testing these models on a wider range of professional datasets. By doing so, we can develop more effective solutions for mitigating hallucinations and make these models more trustworthy.

There is therefore an identifiable gap in the literature: a systematic, end-to-end framework that combines advanced retrieval strategies, faithfulness-aware re-ranking, and grounded generation specifically tailored to reduce hallucinations in domain-specific LLM deployments. This research targets that gap, positing that an architecture integrating hybrid retrieval (dense + sparse), contextual re-ranking, and generation-side faithfulness constraints can achieve substantially lower hallucination rates compared to baseline LLMs and standard RAG systems in specialized domains.

1.4 Research Objectives

This study is led by four main goals:

To systematically analyze and classify hallucination types occurring in LLMs during domain-specific question-answering tasks, establishing a comprehensive taxonomy grounded in empirical evidence.

To design and implement an enhanced RAG framework incorporating hybrid retrieval (BM25 + dense vector search), a faithfulness-aware re-ranking module, and a constrained generation pipeline tailored for domain-specific knowledge bases.

To empirically evaluate the proposed framework's effectiveness in reducing hallucination rates using standardized benchmarks (TruthfulQA, Natural Questions, MedQA) and purpose-constructed domain-specific evaluation datasets.

To compare the performance of the proposed framework against established baselines—including vanilla LLM (GPT-3.5), standard RAG (Lewis et al., 2020), and Self-RAG (Asai et al., 2023)—across factual accuracy, retrieval precision, and generation faithfulness metrics.

1.5 Research Scope

The scope of this research is defined by the following boundaries and parameters:

Model Coverage: The study will focus on publicly accessible transformer-based LLMs, specifically GPT-3.5-turbo, GPT-4 (via API), and Llama 3 (8B and 70B variants), to ensure reproducibility and broad applicability.

We're going to look at two main areas - healthcare and legal text. These are important because they have a big impact and we have good datasets to work with. In healthcare, we'll focus on answering clinical questions and finding drug information. For legal text, we'll work on understanding laws and finding relevant case law.

We'll be doing all our experiments in English. We know that dealing with multiple languages at the same time can be a problem, but that's not what we're focusing on here.

This project is about finding answers to specific questions and checking if the information is true. It's not about creating new stories or writing in a creative way. The goal is to get accurate answers to questions and make sure the facts are correct.

Evaluation Datasets: TruthfulQA (Lin et al., 2022), MedQA-USMLE (Jin et al., 2021), Natural Questions (Kwiatkowski et al., 2019), and a purpose-built domain dataset will serve as the evaluation suite.

Timeframe: The research will be completed within a six-month period as detailed in the Gantt chart in Chapter 3.

This proposal doesn't cover a few things: making models run faster on hardware, teaching models from other models, and getting information in real-time from multiple sources.

CHAPTER 2

Literature Review

2.1 Overview

This chapter brings together research from three main areas: what causes hallucinations in large language models, how retrieval-augmented generation architectures work, and how NLP is used in specific domains. It looks at important studies from the past and present, giving a solid base for this research. At least ten key studies are examined closely to find gaps that make this proposal necessary. The goal is to understand the basics and the latest developments in these areas, and to see how they relate to each other. By reviewing the literature, we can identify what we already know and what we still need to learn, which will help us move forward with our research.

2.2 Hallucination in Large Language Models

2.2.1 Taxonomy and Causes of Hallucination

Researchers have been looking at a big problem in natural language generation called hallucination. Ji and his team did a really thorough study in 2023, where they looked at over 200 other studies and came up with a way to categorize hallucinations. They found that there are two main types: intrinsic hallucinations, which are when the generated text contradicts the original source, and extrinsic hallucinations, which are when the text introduces new information that can't be verified. The team also identified some key factors that contribute to hallucinations. These include problems with the data, such as noise or bias, issues with how the models are trained, like when they prioritize fluency over truth, and problems with how the models make predictions, like when they rely too heavily on low-confidence tokens. What's really important is that Ji's study shows that hallucination isn't just one thing, but rather a range of failures that can happen in different tasks, domains, and models. This means that we need to come up with specific strategies to address hallucinations in each of these areas. This study is really important for our own research, especially when it comes to our first goal.

Researchers Zhang and colleagues took a closer look at this issue in their 2023 study, "Siren's Song in the AI Ocean". They focused on a problem called hallucination in large language models that follow instructions. Specifically, they talked about "sycophantic hallucination",

where models produce answers that sound good but aren't necessarily true, just to agree with what the user said. This can be especially bad in situations where professionals are giving advice. To address this, Zhang and colleagues suggested a way to evaluate these models, which includes five key areas: factuality, faithfulness, reasoning coherence, knowledge grounding, and calibration. This framework is also used in our study to evaluate the models.

Tonmoy and his team did a really thorough job in 2024 of looking at all the different ways to reduce mistakes in AI models. They grouped these methods into three main categories: changing the model's architecture to better handle uncertainty, using special training techniques like curated data and reinforcement learning, and improving the model's output with things like retrieval augmentation and self-consistency checking. What they found was that no one method works perfectly in all situations, but methods that use retrieval - which is like looking up information in a big database - seem to work really well for tasks that need a lot of knowledge. This is why we're focusing on a RAG-centric approach in our proposal, which is all about using retrieval to improve the model's performance.

2.2.2 Hallucination in Domain-Specific Contexts

Researchers like Mallen and others looked into how well language models work in 2023. They wanted to know if these models can be trusted to give accurate information. Their study, called "When Not to Trust Language Models," tested these models using a dataset called PopQA. What they found was that language models are good at answering questions about popular things that are talked about a lot online. But when it comes to less popular topics, or what they call "tail entities," these models don't do very well. This is a problem because specialized fields like medicine or law have a lot of these less popular topics. The study shows that language models need help from outside knowledge to get these answers right. Just relying on the model's memory isn't enough, especially when it comes to facts that don't come up very often.

In healthcare, a study by Omiye et al. in 2023 tested GPT-4 and two other large language models on medical exam questions. They found that while the models were generally accurate, they made mistakes on questions that required exact knowledge of medicines - they came up with believable but wrong drug names, doses, and warnings. This is a big problem in healthcare, and it shows how important it is to make sure language models

have accurate information when they're used in medical settings. This is what motivated the part of our research that focuses on testing language models in specific areas.

2.3 Retrieval-Augmented Generation

2.3.1 Foundational RAG Architecture

Researchers Lewis and his team came up with a new idea in 2020 called Retrieval-Augmented Generation, or RAG for short. This is a way to make language models better at understanding and using knowledge. They put together two different parts: a tool that finds relevant information in big chunks of text, and a generator that can create new text based on what it's learned. When the model is trying to answer a question, it looks at the top pieces of information it's found and uses those to help it generate a response. The team tested RAG on a few different tasks, like answering natural questions, trivia, and questions from the web. They found that it did better than other models that didn't have access to external information. This work laid the groundwork for what we're doing now, which is building on their ideas to make language models even smarter.

Izcard and Grave came up with something called Fusion-in-Decoder, or FiD for short, back in 2021. This idea builds on something else called the RAG paradigm, but it does things a bit differently. Instead of looking at all the information it retrieves as one big chunk, FiD looks at each piece of information on its own and then combines them in a way that makes sense. This approach turned out to be really effective - it did a lot better than RAG when it came to answering questions on a wide range of topics. One of the reasons it was so successful is that it could look at a lot of different pieces of information at the same time, without having to cut any of them short. This is important because it shows that being able to find and use the right information is crucial for getting good results. In fact, when they increased the number of pieces of information they looked at from 10 to 100, the accuracy just kept getting better and better. This tells us that having a good strategy for finding the right information is key, and that's something we're going to keep in mind as we move forward with this proposal.

2.3.2 Advanced and Adaptive RAG Systems

Researchers Asai and their team came up with something called Self-RAG in 2023. It's a way to help a large language model, or LLM, figure out when to look up information, what to look

up, and how to use what it finds. The LLM does this by creating special tokens that help it think about what it's doing. When they tested Self-RAG, it worked a lot better than the standard way of doing things, and even better than other LLMs, on tasks that need a lot of knowledge. This shows that it's better to look up information in a way that depends on how hard the question is, rather than always doing it the same way. Self-RAG is an important comparison for this study.

Researchers like Shuster and his team looked into how retrieval augmentation works in conversations. They found out that models using retrieval to back up their conversations make fewer mistakes, or "hallucinations", than models that just rely on their own parameters. In their study of the Wizard of Wikipedia and Wizard of Internet datasets, they saw that using retrieval to ground conversations in facts directly reduces the number of hallucinations - it's not just about making the conversations more informative. This discovery supports the main idea behind this research, and it's based on actual data.

Gao and his team recently did a big study on something called the RAG landscape. They came up with a way to understand it better by breaking it down into three parts: Naive RAG, Advanced RAG, and Modular RAG. The Advanced RAG is an improvement on the simple approach because it does things like expand queries before searching, re-rank results after searching, and search again to get better results. The Modular RAG is even more flexible, letting you mix and match different parts like searching, rewriting, filtering, and generating text. This study is really helpful because it gives us a clear picture of how all these things fit together and points out some big challenges we still need to solve - like getting generated text to be accurate and working well in different areas.

2.3.3 Retrieval Quality and Faithfulness

Researchers like Nakano and his team made a system called WebGPT in 2022. This system uses a special browser to find information on the web in real-time, and then it uses that information to answer questions. The system was trained with feedback from humans to make sure it was finding the best information. When WebGPT was tested, it showed that using information from other sources made its answers more accurate and trustworthy. However, it also had a problem - sometimes it would pick information that sounded good but was actually wrong. This is called "cherry-picking" and it's like taking a piece of information out of context to support a wrong answer. To fix this problem, the researchers

came up with a new idea called the faithfulness re-ranking module. This module checks the information that WebGPT finds and makes sure it doesn't conflict with other facts that are known to be true. If it does conflict, then the module lowers the ranking of that information so that WebGPT is less likely to use it. The goal is to make WebGPT's answers more accurate and trustworthy by using only the best information.

Researchers found that even the best language models can be easily confused by irrelevant information. In a study, Shi and his team showed that when useless passages are included in the context, it can hurt the model's performance. They call this problem "distraction hallucination." This discovery goes against the idea that having more information is always better. Instead, it highlights the importance of filtering out useless information and re-ranking what's left, which is a key part of the approach we're proposing.

Researchers Min and their team came up with a new way to evaluate how well machines can generate long pieces of text without making things up. They called it FActScoring, and it works by breaking down the text into tiny facts and checking each one against a trusted source. This approach is more detailed and scalable than other methods, which only look at how accurate the text is overall. It can even measure how often the machine makes things up at the sentence or clause level. We're going to use FActScoring, along with a few other methods like TruthfulQA and ROUGE, to assess how well our own system works.

2.4 Gaps in the Literature

The review of existing literature reveals several critical gaps that this research addresses:

Most tests of RAG systems are done using general information, but there's not much research on how well they work in specific areas like healthcare or law, where they could be really useful.

Existing frameworks do not comprehensively integrate hybrid retrieval (sparse + dense), faithfulness-aware re-ranking, and generation constraints within a single end-to-end architecture.

Most of the time, when people talk about hallucinations in research, they just describe what they are. But we're missing a clear way to group them based on what actually goes wrong when

we're trying to retrieve or generate information. We need a system that's based on real data and applies to specific areas, so we can understand how different types of hallucinations happen.

The interaction between retrieval quality and hallucination rate has not been systematically modeled across multiple LLM sizes and architectures in domain-specific settings.

These gaps give us a good reason to do this research and show why it's important from a scientific point of view.

CHAPTER 3

Research Methodology & Planning

3.1 Research Methodology

3.1.1 Research Design

This study uses a quantitative approach and an experimental design, which is part of the applied computing field. We're following a design science research method, which was first talked about by Hevner and others in 2004. This method involves creating something new, like our improved RAG framework, and then testing it thoroughly against existing standards. Our research is divided into five phases that happen one after the other, but also repeat and build on each other: first, we analyze the problem based on what others have written; then, we design the system; next, we implement it; after that, we evaluate how well it works; and finally, we report our findings.

3.1.2 Phase 1 — Hallucination Analysis and Taxonomy Construction

The initial step is to carefully examine the patterns of hallucinations in basic language models when they're used for specific question-answering tasks. To do this, 500 questions will be selected from a few different datasets, including TruthfulQA, MedQA-USMLE, and a special legal QA dataset that was created for this purpose. Two models, GPT-3.5-turbo and Llama 3-8B, will be asked these questions without any prior training, and their answers will be compared to the correct answers using both automatic checks, like FactScore, ROUGE-L, and BERTScore, and manual reviews. The incorrect answers, or hallucinations, will be categorized using a system proposed by Ji et al. in 2023, with some adjustments made based on what's learned from the manual reviews. This step will help create a detailed list of the types of hallucinations that occur, which will then be used to inform the design of the models in the next step.

3.1.3 Phase 2 — Framework Design

The proposed framework consists of three integrated modules:

Hybrid Retrieval Module: Combines BM25 sparse retrieval (Elasticsearch backend) with dense vector search (FAISS index using E5-large embeddings) via a reciprocal rank fusion (RRF)

merging strategy. Domain-specific knowledge bases will be constructed from curated corpora: PubMed abstracts (healthcare) and US Code + case law digests (legal).

Faithfulness-Aware Re-Ranking Module: Retrieved passages will be re-scored using a cross-encoder reranker (MiniLM-L6) fine-tuned on natural language inference (NLI) data to score passage-query relevance and passage-answer consistency. Passages scoring below a calibrated faithfulness threshold will be filtered out before context assembly.

We're working on a special module that helps our language model generate better answers.

This module gets the model to use specific information from the context and cite the right sources. It also flags when the model is unsure about something. After the model gives an answer, we use a smaller model to check if the answer makes sense based on the information we have. This helps us make sure the answers are accurate and trustworthy.

3.1.4 Phase 3 — Implementation

We'll be using Python to put our framework into action, and we'll be relying on some helpful libraries like LangChain and Haystack to get everything working together smoothly. For the really smart stuff, we'll be using HuggingFace Transformers to make predictions with our models. To keep track of all our experiments, we'll be using MLflow, which will help us stay organized. All of this will be running on a powerful GPU cluster, thanks to our institution's HPC allocation - we're talking NVIDIA A100 40GB, so it's got some serious muscle. And the best part? We'll be making all of this available as an open-source repository, so anyone can jump in and reproduce our results.

3.1.5 Phase 4 — Evaluation

The framework will be evaluated against four baselines: (1) Vanilla LLM (GPT-3.5-turbo, no retrieval), (2) Standard RAG (Lewis et al., 2020), (3) Self-RAG (Asai et al., 2023), and (4) BM25-only RAG. Evaluation metrics include:

Hallucination Rate (HR): Proportion of atomic facts in generated answers classified as unsupported or contradicted (measured via FActScore).

Exact Match (EM) and F1: Standard QA metrics against ground-truth answers.

Retrieval Precision@k: Proportion of top-k retrieved passages that are relevant to the query.

Faithfulness Score: NLI-based entailment score between generated answer and retrieved context.

Latency: Average end-to-end response time (seconds) to assess practical deployability.

Statistical significance of improvements will be assessed using paired t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.05$). Error analysis will identify residual hallucination patterns to guide future refinement.

3.1.6 Phase 5 — Reporting and Dissemination

We will write up our results in a detailed report that follows the guidelines set out for the assignment. The results will be prepared for submission to a top conference in the field, such as ACL, EMNLP, or NAACL. At the same time, we will make the implementation of our project available to everyone on GitHub, using an MIT license to allow others to use and build on our work.

3.2 Expected Outcomes

The following tangible outcomes are anticipated upon completion of this research:

A domain-specific hallucination taxonomy categorizing hallucination types across healthcare and legal LLM question-answering, with empirical frequency data for each category.

A fully implemented and documented enhanced RAG framework (codebase + API) integrating hybrid retrieval, faithfulness re-ranking, and constrained generation.

Empirical evidence demonstrating a reduction in hallucination rate of at least 30% compared to baseline LLMs and at least 15% compared to standard RAG on domain-specific benchmarks.

A comparative evaluation report benchmarking the proposed framework against four baselines across all defined metrics, with statistical significance analysis.

A peer-reviewed conference paper or preprint submitted to ACL/EMNLP/NAACL presenting the framework and evaluation results.

An open-source repository enabling researchers to replicate experiments and extend the framework to other domains

3.3 Gantt Chart

The following Gantt chart illustrates the research timeline across a six-month period (May – October 2026).

Activity / Month	May	Jun	Jul	Aug	Sep	Oct
Phase 1: Literature Review & Taxonomy						
Phase 2: Framework Design						
Phase 3: Implementation						
Phase 4: Evaluation & Experimentation						
Phase 5: Analysis & Write-up						
Manuscript Submission						

Shaded cells indicate active phases for each month. The six-month timeline is designed to allow iterative overlap between phases, enabling early experimental findings to feed back into the framework design before full-scale evaluation.

References

References are formatted in IEEE citation style.

- [1] R. Nakano, J. Hilton, S. Balwit, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders et al., "WebGPT: Browser-assisted question-answering with human feedback," arXiv preprint arXiv:2112.09332, 2022.
- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.
- [3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.
- [4] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3784–3803, 2021.
- [5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.
- [6] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 9802–9822, 2023.
- [7] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proceedings of EACL 2021, pp. 874–880, 2021.
- [8] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," arXiv preprint arXiv:2309.01219, 2023.
- [9] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in International Conference on Learning Representations (ICLR), 2024.
- [10] S. M. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," arXiv preprint arXiv:2401.01313, 2024.

- [11] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FactScoring: Fine-grained atomic evaluation of factual precision in long form text generation," in Proceedings of EMNLP 2023, pp. 12076–12100, 2023.
- [12] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. Chi, N. Scharli, and D. Zhou, "Large language models can be easily distracted by irrelevant context," in Proceedings of ICML 2023, pp. 31210–31227, 2023.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
- [14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
- [15] J. A. R. Omiye, J. C. Lester, S. T. Spichak, V. Rotemberg, and R. Daneshjou, "Large language models propagate race-based medicine," npj Digital Medicine, vol. 6, no. 1, p. 195, 2023.