Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 03:2026
DOI: 10.5281/zenodo.18998811

EXPLORING HALLUCINATIONS IN LARGE LANGUAGE MODELS (LLMs): A SYSTEMATIC REVIEW OF TYPOLOGIES, ORIGINS, AND MITIGATION APPROACHES

AMAL ALTALHI
Center of Information Systems and Technology, Claremont Graduate University, Claremont, CA, USA. Department of Management Information Systems, Faculty of Business Administration, University of Tabuk, 71491 Saudi Arabia. Email: aaltalhi@ut.edu.sa, ORCID: 0009-0006-6107-8620

ITAMAR E. SHABTAI
Center of Information Systems and Technology, Claremont Graduate University, Claremont, CA, 91711, USA. Email: itamar.shabtai@cgu.edu, ORCID: 0009-0006-4771-0466

Abstract

This systematic literature review pull out the understanding of large language models (LLMs) by thoroughly examining hallucination situations, including the types, causes, and reduce approaches to enhance LLM usefulness in natural language processing (NLP). Electronic databases (Web of Science, IEEE Xplore, Open Review, Google Scholar) were queried by a comprehensive search, generating 1136 records. Of these, 27 met the inclusion criteria and were included. A meta-aggregative approach was used to analyze and synthesize the articles. The research questions formed significant themes for organizing the findings and results section. LLM's ordinary taxonomy includes fact hallucination, honesty hallucination, lack of alignment, conflict in ideas, nonsensical hallucination, random hallucination, object hallucination, and intrinsic and external hallucination. Hallucination causes were training data issues, model limitation/ overfitting, limited context window/ knowledge cutoff, and nuanced language understanding. Effective mitigative approaches were domain-specific fine-tuning, prompting, model reprogramming, and grounding.

Keywords: Artificial Intelligence, Generative AI, Hallucination, Large Language Model, LLMs, Fine-Tuning, Overfitting, Object Hallucination, Prompting, Grounding.

1. INTRODUCTION

Large language models (LLMs) represent a major advancement in artificial intelligence and natural language processing (NLP). GPT-4,and GPT-3 (Brown, 2020), PaLM (Anil et al., 2023), LLaMA and LLaMA2 (Touvron et al., 2023) demonstrate high-performance levels in understanding, generation, summary, and prediction of the content, and their performance has been improved in NLP tasks (Wang et al., 2019). LLMs can generate fluent and realistic answers using pre-training in the context of supervised fine-tuning and reinforcement learning (Onoe et al., 2022) (Zhao et al., 2023).

However, LLMs also generate untruthful, illogical, fabricated, and unfaithful outputs (Zheng et al., 2023; Jones & Steinhardt, 2022). This hallucination phenomenon (Lee et al., 2023; Yu et al., 2023) reduces reliability. Hallucination appears as intrinsic or extrinsic (Yu et al., 2023), faithful hallucination from contextual, instructional, or logical inconsistencies (Huang et al., 2023), and “silver lining” or “factual mirage” errors (Rawte et al., 2023).

Its causes include poor-quality, biased, outdated, or misleading data, and training issues such as heuristic data collection, imperfect representation learning, innate divergence, exposure bias, error imperceptibility, erroneous decoding, and parametric knowledge bias (Lee et al., 2023). Architectural limitations, attention mechanisms, RLHF (Ouyang et al., 2022), black-box model behavior (Lin et al., 2023), vague knowledge boundaries, insufficient context, sampling randomness, and softmax bottlenecks (Chang & McCallum, 2022; Dhingra et al., 2018) also contribute. Mitigation approaches like benchmark methods (Li et al., 2023), factual-centered metrics, retrieval-based techniques, and prompting for reasoning and self-verification (Shuster et al., 2021) remain underexplored.

This review examines hallucination types, causes, and mitigation strategies, identifies and categorizes hallucinations, analyzes contributing mechanisms, and compares existing techniques. It addresses what types of hallucinations appear in LLMs, what factors cause them, and what approaches mitigate them. The study aims to improve LLM reliability and applicability by offering a systematic framework for understanding and addressing LLM hallucinations.The Hallucination Cascade Model sequences these factors as triggers (data-level biases) amplifying through training (overfitting) to inference (stochasticity), revealing propagations like modality gaps cascading across tasks.

1.1 Background

Large language models (LLMs), primarily characterized by the potential to execute complex tasks such as understanding, generating, summarizing, and predicting novel content, have marked a promising milestone in artificial intelligence, particularly in natural language processing (NPL). For instance, GPT-4, GPT-3 (Brown, 2020) PaLM (Anil et al., 2023), LLaMA (Touvron et al., 2023), and LLaMA2 models exhibit significant efficiency improvements on various NPL tasks (Wang et al., 2019). With appropriate alignments, such as pre-training on massive text corpora preceding supervised fine-tuning and reinforcement learning (Onoe et al., 2022), the LLMs are programmed to understand the natural language and generate fluent and realistic responses following human interactions (Zhao et al., 2023). Despite their plausible capabilities and success, the LLMs occasionally produce untruthful content with illogical, false, fabricated texts(Zheng et al., 2023) or unfaithful output responses (Jones & Steinhardt, 2022). This phenomenon is referred to as hallucination (Lee, et al., 2023; Yu, et al., 2023), known to undermine LLMs' reliability and applicability.

1.2 Problem Statement

The devastating phenomena demands comprehensiveness in LLM hallucination quantifying (by type and causes) and identifying appropriate mitigative strategies. However, the latter remains a challenge because hallucination is, by default, a composition set of phenomena (Lee, et al., 2023). Hallucinations are conventionally classified as intrinsic or extrinsic (Yu, et al., 2023). Huang et al. (2023) mention faithful hallucination, which accounts for LLMs' user cases through contextual, instructional, and logical inconsistencies. Rawte et al. (2023) perceptively classify hallucination into “silver lining and “factual mirage” based on the erroneous outputs of factual input.

On LLM hallucination causes, data issues such as poor quality, bias, misinformation, and outdated knowledge are critical. Lee, et al. (2023) also mentioned heuristic data collection, imperfect representation learning or massive data training, innate divergence, exposure bias, imperceptibility of errors, erroneous decoding, and parametric knowledge bias as confounders. In the training phase, architectural and strategic deficiencies could hamper proper model learning, causing inconsistency (Ranzato et al., 2015). Attention mechanisms, including the RLHF process (Ouyang et al., 2022), black-box LLMs property (Lin et al., 2023), and the vague knowledge boundary, also limit hallucination detection. At the inference level, insufficient context attention sampling randomness causes LLM hallucination and softmax bottleneck (Chang & McCallum, 2022; Dhingra et al., 2018). Among the mitigative approach options, scholars enlisted benchmark methods (Li et al., 2023), factual-centered metrics, retrieval-based methods, and prompting the model to reason and verify their answers, which also have potential hallucination mitigators (Shuster et al., 2021).

1.3 Objectives

This review aims to offer a comprehensive, structured understanding of LLMs' hallucinations, particularly the types, causes, and mitigation approaches, to bolster large language model applicability in natural language processing (NPL) and the larger artificial intelligence field. Therefore, the specific study objectives included identifying and categorizing different types of hallucinations experienced in LLMs, analyzing the underlying causes and mechanisms leading to hallucination in LLMs, and summarizing and comparing existing techniques and strategies for addressing hallucinations in LLMs.

Research Questions

1) What are the common types of hallucination observed in large language models?

2) What are the factors contributing to hallucination in LLMs?

3) What approaches have been proposed to mitigate hallucination in LLMs?

2. METHODS

This present study adopted the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) frameworks and the SLR methodology outlined by Kitchenham et al. (2009), a rule-driven comprehensive approach to find and analyze prior knowledge on a topic of interest through a rigorous and transparent method of identifying, evaluating, and interpreting facts and findings from the available research (Brereton et al., 2007).

2.1 Search Strategy

After delineating our research questions, various search engines and databases (Web of Science, Google Scholar, IEEE Xplore, and OpenReview) were consulted to produce relevant articles. The search included the following search terms and keywords: large language model (LLMs), generative pre-trained transformer (GPT), large vision-language models (LVLM), hallucination, LLM types/taxonomy, LLM causes/factors, and mitigation.

2.2 Eligibility Criteria

After articles were retrieved from the databases, a relevance assessment ensued based on the study's inclusion and exclusion criteria. Studies were included if they were peer-reviewed, including academic journals, surveys, conferences, and papers; discussed LLM hallucinations, contextualized type or classification, causes/ problems aligned to LLM hallucination, and or mitigation strategies/approaches/ with or without frameworks; and published English language papers between 2020 and 2024; and with full-time access.

Otherwise, studies were excluded if they were non-peer-reviewed articles, opinion papers, grey literature, irrelevant articles, including book chapters and editorials or duplicates, and non-English language articles.

2.3 Data Extraction and Management

Two authors independently screened the titles and abstracts of 20% of the retrieved articles to test and refine the preset eligibility criteria. Following a joint discussion of the screened records, the first reviewer screened the rest of the studies. All the eligible studies were retrieved after full-text screening and pooled for data abstraction. The two reviewers independently extracted data into a self-designed extraction form.

They then iteratively discussed the extracted data, discussing any anomalies and inconsistencies to a consensus. The following data were collected: author and publication year, type of study, the taxonomy of LLM, task categories, factors or issues likely to cause hallucination, and any mitigating ideas, including methods or approaches. In the extraction form, issues and factors related to hallucination, and mitigation methods, the contents were purely verbatim results from each of the studies included in the review.

2.4 Analysis & Synthesis

Data analysis and synthesis were done following a meta-aggregative approach. Based on the information fields collected, preliminary categories were developed and refined iteratively based on the research questions.

Due to constraints in conducting a quality appraisal for the included studies, a hybrid approach was opted. Descriptive reporting on study procedures was done to distinguish records that passed through quality control (peer-reviewed articles) from other articles. Findings were then critically evaluated during reporting to appraise the validity and comprehensiveness of their information.

3. RESULTS & SYNTHESIS

3.1 Description of Studies/Search Results

The search process yielded 1136 articles. After removing duplicates, 756 records were screened in the title and abstract. Unsuitable articles were then eliminated, leaving only 156 for full-text assessment. The final result included only 27 records in the dataset in full compliance with the inclusion criteria. The search process result is illustrated in **Figure 1**.

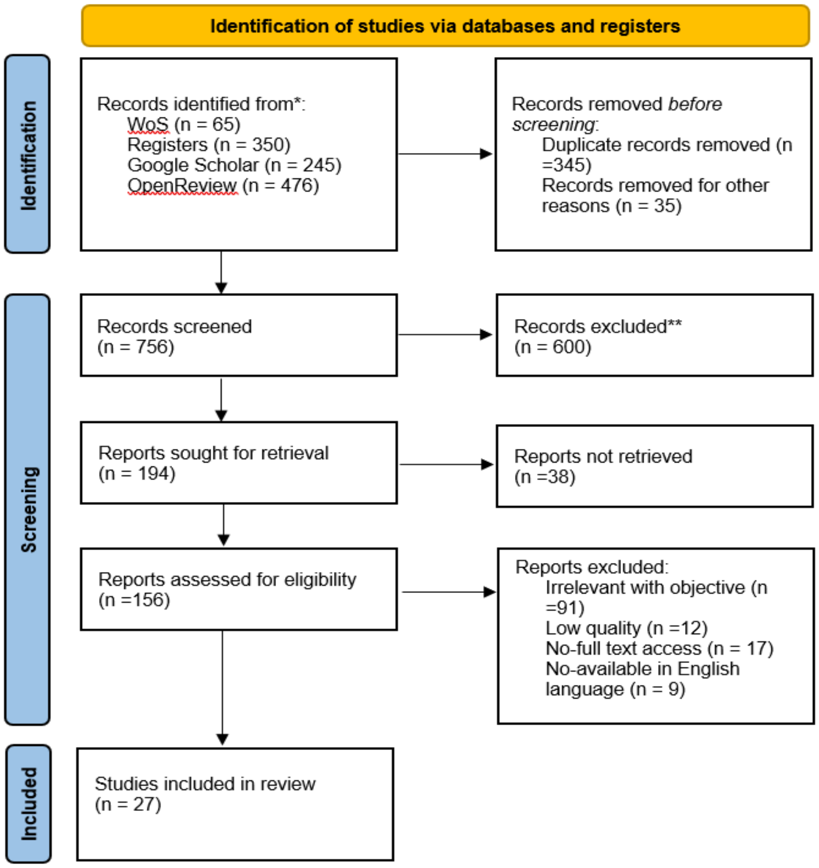
Identification of studies via databases and registers
Identification
Records identified from*:
WoS (n = 65)
Registers (n = 350)
Google Scholar (n = 245)
OpenReview (n = 476)
Records removed *before screening*:
Duplicate records removed (n =345)
Records removed for other reasons (n = 35)
Screening
Records screened
(n = 756)
Records excluded**
(n = 600)
Reports sought for retrieval
(n = 194)
Reports not retrieved
(n =38)
Reports assessed for eligibility
(n =156)
Reports excluded:
Irrelevant with objective (n =91)
Low quality (n =12)
No-full text access (n = 17)
No-available in English language (n = 9)
Included
Studies included in review
(n = 27)

Figure 1: PRISMA Flow Diagram

The included studies were published between 2021 and 2024, reporting various LLM hallucination phenomena. Three major themes in the dataset aligned with the intended research questions during the literature analysis. These themes were used to structure the reporting and synthesis of this review. The main themes included common types or taxonomy of LLM hallucinations, causes or factors contributing to LLM hallucinations, and mitigation frameworks and approaches to LLMs. A detailed summary of information collected from the included studies is reported in **Tables 1 and 2.**

3.2 Common Types (Taxonomy) of Large Language Model Hallucinations

The main LLM task applications reported included task summarization, machine translation, multilingual sequence-to-sequence, knowledge-grounded dialogue generation, and question-and-answer tasks, which formed the central classification of LLM hallucinations in this review, as shown in **Table 1.**

Within the machine translation, the study noted trustworthy hallucination, possibly induced by text perturbations (e.g., spelling or capital errors). Natural hallucination was also identified under corpus-level noise, with subtypes including detached and oscillatory outputs (Raunak et al., 2021). Dale et al. (2023) alludes to full and partial hallucinations, which occur at the sentence and word levels. Machine translation tasks also exhibited largely fluent types of hallucinations (Guerreiro et al., 2022), premediated by off-target, over-generated, or failed translation prompts. For questions and answers, comprehension hallucination, specificity, inference, and factuality hallucinations were notable (Wang, 2023), possibly due to the imperfect responses by LLMs resulting from flawed external knowledge, reasoning instructions, or knowledge recall cues. Studies also pointed to reasoning hallucinations and memory-based hallucinations, which were consequences of the external knowledge of the model and memorized content without reliable, accurate, and accessible knowledge sources (Umapathi et al., 2023). Lin et al. (2021) added imitative falsehood and factorial errors, a type of hallucination emanating from scaling up models (Cheng et al., 2023; Lin et al., 2021). This taxonomy also witnessed semantic equivalence, intrinsic ambiguity, symbolic equivalence, incomplete enumeration and granularity discrepancies (Adlakha et al., 2023). These arise when LLMs cannot avoid answering even when provided with irrelevant data and instead proceed to provide incomplete and plausible answers. Object hallucination is common in question and answering tasks in large language models (Li et al., 2023).

Regarding dialogue system tasks, uncooperativeness (Dziri, Rashkin, et al., 2022), generic or historic corrupted contents, and attributable and general hallucinations were notable(Das et al., 2022). Arguably, studies perceive dialogue LLM as unobtrusive imitators simulating the data distributional properties rather than generating faithful output content. The former can well be described as unfaithful hallucinations. Reiterating the history-corrupted contents, Das et al. 2023) further mentioned extrinsic and intrinsic or repetitive hallucinations. In the knowledge graph, chatbots Mihindukulasooriya et al. (2023) reported that subject hallucination, object hallucination, and relation hallucination are inherent based on the fidelity of the knowledge source of the models. Yu et al. (2023) attributed to knowledge hallucination, which is usually realized during knowledge creation.

Nevertheless, for task summarization and reasoning (text and visuals), a common hallucination is faithfulness hallucination, which is the source data document's lack of faithfulness. This could be in the form of intrinsic hallucination, which corrupts or distorts the information in the document, and extrinsic hallucination, which adds more non-attributable information to the content source document (Qiu et al., 2023). Cao et al. (2021) identified factual and non-factual hallucinations, which are subtypes of extrinsic hallucinations. Factuality hallucination (Jha et al., 2023) is another common type, with subtypes including factual inconsistency (Tam et al., 2022), judgment or description hallucination (Liu et al., 2024), fine-grained object hallucination (Wang et al., 2024), and object hallucination (Zhou et al., 2023).

In the final taxon, the cross-model or multi-lingual systems allude to scenarios where the LVLM generates object descriptions that are not present in the target image (Wang et al., 2024), causing fine-grained or object hallucination (Wang et al., 2023). Pfeiffer et al. (2023) also reckon source language hallucination, realized when the multi-lingual sequence-to-sequence models are performing poorly due to the increasing inconsistency in generating correct text.

Table 1: Hallucination Taxonomy (classification) & Types

Research Method	LLMs models	Task Category	Hallucination types/ sub-types
Source perturbation	Neural Machine Translation (NMT)	Machine translation	Under perturbation, Natural hallucination (detached and oscilary).
Introduce pathology detection	Unspecified	Machine Transition	Full hallucination, Partial hallucination, Word-level hallucination
Consideration of a natural scenario	COMET-QE.	Machine Transition	Oscillatory hallucination, Largely fluent hallucination
Evaluate source language hallucination	mBERT, X-Mod, mT5.	Multilingual Seq2seq	Source language hallucination
Medical benchmark Med-HALT	Text Davinci, GPT-3.5, LlaMa-2, MPT, and Falcon.	Question and Answer	Reasoning hallucination, Memory-based hallucination
Manual analysis of responses	ChatGPT/ GPT-4	Question and Answer	Comprehension, Factualness, Specificity, Inference Hallucination
Cause imitative falsehoods	GPT-3, GPT-Neo/J, GPT-2	Question and Answer	Imitative falsehood
Evaluate retrieval augmented QA	Flan-T5, GPT-3.5, GPT-3.5	Question and Answer	Semantic equivalence, Intrinsic ambiguity, Granularity discrepancies, Enumeration, Satisfactory Subset
Caption hallucination assessment	mPLUG-Owl, MultiModal-GPT, MiniGPT-4, LLaVA,	Visual Question Answer	Object hallucination
Analyze entity-level fact hallucination	GPT2-KG	Dialog System	Extrinsic-Soft/Hard/ Grouped, Intrinsic-Soft/ Hard/Repetitive, History Corrupted
Hallucination-free benchmark FaithDial	GPT2, DIALOGPT, T5 (DoHA).	Dialog System	Hallucination, Generic, Uncooperativeness
Knowledge-grounded interaction benchmark Begin	T5, GPT2, DoHA, and CTRL-DIALOG	Dialog System	Fully attributable, Not attributable, Generic
Generate summaries from given models	BLOOM, GPT and OPT	Summarization System	Factually inconsistent summaries

Label factual entities from summarizations	K-Nearest Neighbors (KNN)	Summarization System	Non-hallucinated, Factual hallucination, Non-factual hallucination, Intrinsic hallucination
In a cross-lingual transfer setting	MAD-X	Summarization System	Intrinsic hallucination (unfaithfulness), Extrinsic hallucination
Ontology-driven KGC benchmark Text2KGBench	Wikidata, DBpedia	Knowledge-based text generation	Subject hallucination, relation hallucination, object hallucination
Evaluate knowledge-creating ability given known facts	GPT-4 (2.06) and GPT-3.5-turbo, GPT-J and BLOOM.	Knowledge-based text generation	Knowledge hallucination
Critical review with extensive experiments on models	GLM-130B and ChatGPT	Question and Answering	Imitative falsehoods and factual errors
Critical review with case studies	ChatGPT	Text summarization and questions and answering	Factually inaccurate hallucination
Critical review	MiniGPT-4.	Open-domain vision-language task	Judgement hallucination. Description hallucination
Critical reviewing	ChatGPT	Visual summarization	Fine-grained object hallucination

3.3 Causes/Factors Contributing to LLM Hallucinations

This review identified causes or factors leading to LLM hallucinations in distinct domains based on data source factors, factors from model alignment and vision encoder sources, and architecture-related LLM factors (**Table 2).**

Concerning data-related factors, studies attribute model training data quality issues to the model's efficiency and performance. For example, data bias exhibited by the distribution imbalance of training data causes hallucinations. Data bias can also be revealed via data homogeneity, hindering the model's ability to understand texts or visual information for accurate execution (Liu et al., 2024). Lin et al. (2021) corroborates that training data on false answers led to imitative falsehood. Zhang et al. (2023) further connotes knowledge recall, memorization, data issues, and premeditated factuality hallucination. Annotation irrelevance is also a significant factor. This is where LLMs synthesize a large amount of instruction data from primary source documents, but due to model unreliability, annotation irrelevance surfaces (Liu et al., 2024). Adlakha et al. (2023) summarized that inaccurate judgment of knowledge relevance causes semantic hallucinations and intrinsic ambiguity. In relation to model alignment and vision encoder factors, the vision encoder includes limited visual resolutions and fine-grained visual semantics. Wang et al. (2024) connote that object inconsistency with the target image in the description attributes object hallucination. High image resolution arguably enhances the accuracy of visual encoders in object recognition. Thus, limiting visual LVLM resolution while handling a more comprehensive range of images is a possible cause of fine-grain and object hallucinations.

Nonetheless, hallucinations caused by modality alignment tend to be catalyzed by the LVLM connection module, for instance, which connects the module project's visual features into the word embedding space of the model. Thus, model misalignment could be a critical confounder for hallucination. Zhou et al. (2023) respectively assert that object hallucination is a consequence of non-existing object imaging, possibly due to modality misalignment, as earlier connoted by Wang et al. Wang et al. (2024). For the architecture-related factors, insufficient context attention, capability misalignment based mainly on the model's inherent capabilities, and stochastic sample decoding that introduces randomness into decoding are hallucination-related factors. Insufficient context attention, for example, occurs when the LLM focuses only on particular or partial information during text decoding. In accordance, Jha et al. (2023) sentiments that lack of real-world knowledge is a strong factor for factuality hallucinations. Similar observations are the nexus for data training issues premediating general hallucination (Li et al., 2023) generic hallucination (Dziri, Kamalloo, et al., 2022).

3.4 Mitigative Approaches to LLMs Hallucinations

The inherent complexity and training processes of the LLMs mean that hallucinations are almost inevitable. Due to this, there has been a rapid shift from trying to eliminate them to effectively mitigating their impact and frequencies. Hence, strategies have been proposed and implemented, ranging from technical solutions in the model training phase to procedural safeguarding of the implementation at the deployment and use phase, as detailed in **Table 2.**

At the foundational stages of the use and product design case, the product design and user integration approaches have been vital, focusing on configuring the models to diminish hallucination risks intrinsically. Fine-tuning, for example, an approach involving transforming the general-purpose model into a specialized LLM model, is considered a top priority. According to Wang et al. (2024), fine-tuning, particularly fine-grained probing, immensely minimized hallucination risks in fine-grained objects. During the deployment and implementation stages, prompt engineering and meta-prompt design are alluded to as effective in mitigating hallucination risks. Accordingly, the nuanced practice of prompt engineering, or construction of meta-prompts, which involves transforming the input text data based on the specific module templates and restructuring the tasks into formats that fully utilize the pre-trained language models, could be crucial for optimizing LLM functionality and performance. Jha et al. (2023) reiterates the effectiveness of iterative prompting to alleviate factually inaccurate hallucinations. Literature also recognized the zero-shot, one-shot, and few-shot approaches, where the models are enabled to predict unseen classes without necessarily being trained on these contents (Pfeiffer et al., 2023).

Nevertheless, reprogramming or adversarial reprogramming, which involves modifying the model inputs to enable the application of new tasks, could also be an ideal mitigation method. Zhou et al. (2023) resonate to a reconstructive description of the LLM model to alleviate hallucinations. On the same note, retrieval augmented generation (RAG), an architectural approach to incorporating specific context or data into the LLMs to provide an accurate and domain-specific response, is also a valuable mitigation strategy. A more refined concept in RAG is the LLM grounding, which enhanced the separation of the query logic from linguistics for boosted debugging (Es et al., 2023). Also, data management and continuous improvement approaches could be the cornerstones to effectively mitigating hallucination. These approaches ensure meticulous model data management and the continuous improvement of the LLMs. Cumulatively, included studies suggest scaling up (Liu et al., 2024), confidence versus uncertainty measures (Lin et al., 2023), knowledge alignment (Zhang et al., 2023), knowledge presentation harmonization (Li et al., 2023), constructive learning, and standard log probability (Dale et al., 2023; Sun et al., 2023) as effective approaches to this cause.

Table 2: Hallucination Sources & Mitigation Approaches

Article Type	Tasks/ Applications	Hallucination Causes/ Sources	Mitigation approaches (Frameworks & Methods)	Conclusions
Survey	Visual summarization and reasoning	Descriptions with non-existing image objects	Reconstructing less hallucinatory descriptions, the LURE model was proposed to post-hoc rectify object hallucination.	LURE was effective in general object hallucination evaluation metrics, GPT, and human evaluations
Survey	Knowledge-grounded dialogue generation	Limited topics covered in the training data.	Unified knowledge presentation: The study proposed a PLUG model for harmonizing different knowledge sources	PLUG generalizes well across different knowledge-grounded dialogue tasks
Survey	Knowledge-grounded dialogue generation	Contents contradiction with previous texts, content misalignment.	Knowledge alignment: Experiment using the MixAlign benchmark, which produces high-quality user-centered clarifications.	Knowledge alignment crucially enhances model performance.
Survey	Questions & Answering	Content divergence	Confidence vs uncertainty measures applying selective natural language generation	Semantic dispersion measures enhance reliably to predict the quality of LLM responses.
Survey	Text summarization and questions and answering	Lacking real-world knowledge and inaccurate	Iterative prompting (counterexample-guided abstraction refinement)	Iterative prompting architecture can formally detect errors in

		training data/ responses.		LLM responses automatically.
Survey	Open-domain vision-language task summarization	Annotation irrelevance, limited-visual, and resolution.	Scaling up vision resolution, perceptual enhancement, connection module enhancement, alignment training optimization	Post-processing or output editing via an additional module or operations.
Experimental Survey	Visual summarization	Non-existing objective image in the input image	Fine-Tuning: Caption Rewrites/fine-grained probing-based evaluation method. ReCaption	ReCaption effectively reduces fine-grained object LVLM hallucination and improves generated text quality.
Experimental	Multilingual Seq2seq	Representation drift during fine-tuning	Prompting (one-shot and few/zero-shot) using mmT5, a modular multilingual sequence-to-sequence model.	mmT5 raises the rate of generating text in the correct language, alleviating the source language hallucination.
Experimental learning	Contrastive learning; MixCL	Unspecified	Contrastive learning using a learning scheme (MixCL).	MixCL optimized the knowledge elicitation process of LMs and thus reduced their hallucination.
Experimental	Knowledge-grounded dialogue generation	Unspecified	Reconstructing descriptions using Knowledge Graphs; Text2KGBench, a benchmark to evaluate the capabilities of language models to generate KGs from natural language text.	There is room for improvement in the semantic web and natural language processing techniques.
Survey	Questions & Answering	Lack of faithfulness to the original document	Augmentation for domain specialization (RAG): RAG encompasses a retrieval and LLM-based generation module and provides LLMs with knowledge from a reference textual database, enabling them to act as a natural language layer between a user and textual databases.	The approach can effectively evaluate varied domain dimensions without relying on ground truth human annotations.

4. DISCUSSION

4.1 Main Types

Based on the findings, factuality hallucination is among the most commonly observed in LLMs, with subtypes including factual inaccuracies, inconsistencies, and fabrication (Zhang et al., 2024). Factual inaccuracy occurs when the models generate misleading or incorrect information, such as inaccurate historical texts or inconsistent scientific facts. Factual contradiction, on the other hand, manifests when the LLMs generate fabricated or fictional content and falsely present it as factual.

Faithfulness hallucination is also popular, presented through instruction, context, or logical inconsistency (Xu et al., 2024). It mainly happens when the model produces texts or contents that could be more consistent or unfaithful to its source data. Instruction inconsistency occurs when the model ignores specific user instructions.

Similarly, context inconsistency arises when the model output includes data not presented in the source context or contradicts the provided context (Benrimoh et al., 2019). Logical inconsistency also happens when the output of the LLM model contains logical errors. Nonsensical responses are another common type of LLM hallucination that occurs when the model generates irrelevant responses or texts in the input prompt (Liu et al., 2024), usually due to a limited understanding of the context or loss of the logical conversation thread.

Nevertheless, LLMs are also prone to generating hallucinations in the form of random or irrelevant responses that are not pertinent to the input or desired content output. Within the domains of vision-language generation, object hallucination is the primary, encompassing visual, auditory, olfactory, tactile gustatory, and general somatic hallucinations. These hallucinations can alternatively be classified as intrinsic or extrinsic hallucinations.

4.2 Contributing Factors

From the findings, this study extrapolates some of the common causes or factors of large language hallucinations. Broadly, insufficient or biased training data limits the quality and diversity of data used to train large language models, leading to hallucinations (Huang et al., 2023). By default, the large language models require comprehensive and diverse datasets to learn accurate and consistent language presentations. Insufficient data exposes the model to a lack of valuable information for generating consistent and accurate outputs.

Nevertheless, if there are inaccuracies, misinformation, inadequacy, and biases in the training data, the LLM model tends to learn and perpetuate the biases and inaccuracies, causing hallucinations. Training data issues can also include noise, inconsistencies, errors, and irrelevant information, which mediates factuality hallucination (Yu et al., 2024). Model limitation, or overfitting, refers to when the LLM is highly accurate with the information it was initially trained on but struggles with new datasets (Yin et al., 2023).

In most cases, LLMs are trained to generalize based on the training data and, consequently, to handle new contexts sufficiently. Mostly common in the inference stage of LLM development, model limitation can also happen due to inherent randomness in the data source sampling and decoding.

Studies also allude to limited context window or prompt engineering as a hallucination factor. This occurs when the model is designed or trained to simultaneously consider a particular word(tokens) volume (Liu et al., 2023). This leads to misunderstanding and omission of critical information when the model is tasked with larger documents. When prompted, the model generates responses based on a partial understanding of the context. For nuanced language understanding, the large language model struggles to interpret subtleties of human language, such as cultural references and sarcasm, consequently generating irrelevant or outdated information in pertinent times when nuance is crucial for understanding the LLM prompt.

4.3 Mitigating Strategies

Mitigating LLM hallucinations still poses a significant threat due to their increasing worldwide adoption. The findings of this study highlight but are not limited to domain-specific fine-tuning, prompting, and model reprogramming as potential mitigative approaches. Training data issues have been a concern in the model pre-training and training phase; hence, improving training data quality and diversity is deemed ideal to minimize hallucination (Tonmoy et al., 2024). Curating the training datasets to make them balanced, comprehensive, and representative of wide topic ranges and perspectives can reduce the chances of LLMs learning inaccurate or biased language patterns (Amatriain, 2024). In prompting, particularly one-shot and few/zero-shot, demonstrations restrict the model’s output behavior to a specific response length. The one-shot prompting, for instance, involves framing the prompt to a single instruction or sentence to limit model responses to minimize hallucination risks. Contrarily, the few-shot prompting prompts the model to use a series of instructions or contextual examples to build the model output context so that the model generates anticipated or desired responses.

Findings also reiterate the significance of fine-tuning, which involves teaching the model new knowledge while retaining its existing capabilities. Scholars corroborate that domain-specific fine-tuning shapes the LLM’s responses, preventing it from hallucinating in the form of plausible reactions (Zheng et al., 2024). Another highlighted strategy is the RAG (Lewis et al., 2020), which involves appending the prompt with embeddings generated from the domain-specific knowledge dataset to allow additional context while generating output text for downstream tasks.

Comprehensively, RAG includes domain knowledge augmentation (grounding) and domain tool augmentation (Ling et al., 2023). Grounding broadly is the comprehensive understanding of concepts, patterns, and facts unique to a specific subject area and domain. The domain tool augmentation, on similar accounts, integrates or supplements the LLM's responses with external information to expand its capabilities to handle more tasks (Li et al., 2024).

5. CONCLUSION

This systematic literature review explores the burgeoning phenomenon of large language model hallucinations. As exhibited, LLMs have immense potential to transform the language domains by providing sophisticated tools for parsing large datasets that enhance natural language processing tasks. However, this journey is impeded by hallucination, a phenomenon that compromises LLM performance. This review has analyzed and summarized the types, causes, and mitigative approaches of LLM hallucinations in varied task levels. Despite the immense contribution to the LLM field, this study was not short of a limitation. The failure to perform a quality assessment for the included articles was a significant exception. This was due to the nature of the studies, which could not inform a standardized quality appraisal metric. Lack of quality assessment means that quality issues such as missing data, biases, and risks, among others, cannot be established for comparison. Even so, there were significant implications for future research. For instance, fine-tuning, advancement, knowledge transfer, and meta-learning present a crucial future avenue to make large language models highly adaptable with accuracy and precision. Such future development holds the potential to enhance model efficacy in the rapidly changing environment invaluably. On scaling up, the future of large language models will likely experience an inevitable shift towards integrated multimodal or multilingual information sources.

References

1) Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., & Reddy, S. (2023). *Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering*. https://doi.org/10.48550/ARXIV.2307.16877

2) Amatriain, X. (2024). *Measuring and Mitigating Hallucinations in Large Language Models: A multifaceted Approach*.

3) Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. El, Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). *PaLM 2 Technical Report*. https://doi.org/10.48550/ARXIV.2305.10403

4) Benrimoh, D., Parr, T., Adams, R. A., & Friston, K. (2019). Hallucinations both in and out of context: An active inference account. *PLOS ONE*, *14*(8), e0212379. https://doi.org/10.1371/journal.pone.0212379

5) Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

6) Brown, T. (2020). Pronunciation and good language learners. In *Lessons from Good Language Learners* (Vol. 33, pp. 197–207). Cambridge University Press. https://doi.org/10.1017/cbo9780511497667.018

7) Cao, M., Dong, Y., & Cheung, J. C. K. (2021). *Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization*. https://doi.org/10.48550/ARXIV.2109.09784

8) Chang, H.-S., & McCallum, A. (2022). Softmax bottleneck makes language models unable to represent multi-mode word distributions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*.

9) Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., He, J., Huang, M., Yin, Z., Chen, K., & Qiu, X. (2023). *Evaluating Hallucinations in Chinese Large Language Models*. https://doi.org/10.48550/ARXIV.2310.03368

10) Dale, D., Voita, E., Lam, J., Hansanti, P., Ropers, C., Kalbassi, E., Gao, C., Barrault, L., & Costa-Jussà, M. (2023). HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 638–653. https://doi.org/10.18653/v1/2023.

11) Das, S., Saha, S., & Srihari, R. (2023). *Diving Deep into Modes of Fact Hallucinations in Dialogue Systems*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.

12) Dhingra, B., Jin, Q., Yang, Z., Cohen, W. W., & Salakhutdinov, R. (2018). *Neural Models for Reasoning over Multiple Mentions using Coreference*. https://doi.org/10.48550/ARXIV.1804.05922

13) Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E., & Reddy, S. (2022). <scp>FaithDial</scp>: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, *10*, 1473–1490. https://doi.org/10.1162/tacl_a_00529

14) Dziri, N., Rashkin, H., Linzen, T., & Reitter, D. (2022). Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. *Transactions of the Association for Computational Linguistics*, *10*, 1066–1083. https://doi.org/10.1162/tacl_a_00506

15) Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. https://doi.org/10.48550/ARXIV.2309.15217

16) Guerreiro, N., Voita, E., & Martins, A. (2022). *Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023

17) Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. https://doi.org/10.48550/ARXIV.2311.05232

18) Jha, S., Jha, S. K., Lincoln, P., Bastian, N. D., Velasquez, A., & Neema, S. (2023). Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. *2023 IEEE International Conference on Assured Autonomy (ICAA)*, 149–152. https://doi.org/10.1109/ICAA58325.2023.00029

19) Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

20) Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1827–1843). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.

21) Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, *35*, 11785–11799.

22) Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. In *Information and Software Technology* (Vol. 51, Issue 1). https://doi.org/10.1016/j.infsof.2008.09.009

23) Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc.

24) Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. https://doi.org/10.48550/ARXIV.2305.11747

25) Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., … Zhao, L. (2023). *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey*. https://doi.org/10.48550/ARXIV.2305.18703

26) Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. https://doi.org/10.48550/ARXIV.2109.07958

27) Lin, Z., Trivedi, S., & Sun, J. (2023). *Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models*. https://doi.org/10.48550/ARXIV.2305.19187

28) Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., & Zhang, L. (2024). *Exploring and Evaluating Hallucinations in LLM-Powered Code Generation*. https://doi.org/10.48550/ARXIV.2404.00971

29) Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). *A Survey on Hallucination in Large Vision-Language Models*. https://doi.org/10.48550/ARXIV.2402.00253

30) Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, *55*(9). https://doi.org/10.1145/3560815

31) Li, X., Liu, M., & Gao, S. (2024). *GRAMMAR: Grounded and Modular Methodology for Assessment of Domain-Specific Retrieval-Augmented Language Model*. https://doi.org/10.48550/ARXIV.2404.19232

32) Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). *Evaluating Object Hallucination in Large Vision-Language Models*. https://doi.org/10.48550/ARXIV.2305.10355

33) Mihindukulasooriya, N., Tiwari, S., Enguix, C., & Lata, K. (2023). Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. In *The Semantic Web – ISWC 2023* (pp. 247–265). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47243-5_14

34) Onoe, Y., Zhang, M. J. Q., Choi, E., & Durrett, G. (2022). *Entity Cloze by Date: What LMs Know About Unseen Entities*. https://doi.org/10.48550/ARXIV.2205.02832

35) Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Lowe, & R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

36) Pfeiffer, J., Piccinno, F., Nicosia, M., Wang, X., Reid, M., & Ruder, S. (2023). *mmT5: Modular Multilingual Pre-Training Solves Source Language Hallucinations*. https://doi.org/10.48550/ARXIV.2305.14224

37) Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E. M., & Cohen, S. B. (2023). *Detecting and Mitigating Hallucinations in Multilingual Summarisation*. https://doi.org/10.48550/ARXIV.2305.13632

38) Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2015). *Sequence Level Training with Recurrent Neural Networks*. https://doi.org/10.48550/ARXIV.1511.06732

39) Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). *The Curious Case of Hallucinations in Neural Machine Translation*. https://doi.org/10.48550/ARXIV.2104.06683

40) Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A. P., & Das, A. (2023). *The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations*. https://doi.org/10.48550/ARXIV.2310.04988

41) Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval Augmentation Reduces Hallucination in Conversation*. https://doi.org/10.48550/ARXIV.2104.07567

42) Sun, W., Shi, Z., Gao, S., Ren, P., De Rijke, M., & Ren, Z. (2023). Contrastive Learning Reduces Hallucination in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(11), 13618–13626. https://doi.org/10.1609/aaai.v37i11.26596

43) Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., & Raffel, C. (2022). *Evaluating the Factual Consistency of Large Language Models Through News Summarization*. https://doi.org/10.48550/ARXIV.2211.08412

44) Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. https://doi.org/10.48550/ARXIV.2401.01313

45) Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://doi.org/10.48550/ARXIV.2302.13971

46) Umapathi, L., Pal, A., & Sankarasubbu, M. (2023). *Med-halt: Medical domain hallucination test for large language models*.

47) Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Bowman, & S. (2019). General Purpose Embedded Operating Systems. In *Embedded and Real-Time Operating Systems* (pp. 293–355). Springer International Publishing. https://doi.org/10.1007/978-3-031-28701-5_8

48) Wang, L., He, J., Li, S., Liu, N., & Lim, E.-P. (2024). Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. In *MultiMedia Modeling* (pp. 32–45). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53302-0_3

49) Wang, Y. (2023). *Aligning Large Language Models with Human: A Survey*. https://doi.org/10.48550/ARXIV.2307.12966

50) Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). *Aligning Large Language Models with Human: A Survey*. https://doi.org/10.48550/ARXIV.2307.12966

51) Xu, Z., Jain, S., & Kankanhalli, M. (2024). *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. https://doi.org/10.48550/ARXIV.2401.11817

52) Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., BAI, L. E. I., Shao, J., & Ouyang, W. (2023). LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 26650–26685).

53) Yu, J., Wang, X., Tu, S., Cao, S., Zhang-Li, D., Lv, X., Peng, H., Yao, Z., Zhang, X., Li, H., Li, C., Zhang, Z., Bai, Y., Liu, Y., Xin, A., Lin, N., Yun, K., Gong, L., Chen, J., … Li, J. (2023). *KoLA: Carefully Benchmarking World Knowledge of Large Language Models*. https://doi.org/10.48550/ARXIV.2306.09296

54) Yu, J., Zhang, X., Xu, Y., Lei, X., Yao, Z., Zhang, J., Hou, L., & Li, J. (2024). *A Cause-Effect Look at Alleviating Hallucination of Knowledge-grounded Dialogue Generation*. https://doi.org/10.48550/ARXIV.2404.03491

55) Zhang, J., Xu, C., Gai, Y., Lecue, F., Song, D., & Li, B. (2024). *KnowHalu: Hallucination Detection via Multi-Form Knowledge Based Factual Checking*. https://doi.org/10.48550/ARXIV.2404.02935

56) Zhang, S., Pan, L., Zhao, J., & Wang, W. Y. (2023). *The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models*. https://doi.org/10.48550/ARXIV.2305.13669

57) Zhao, R., Li, X., Joty, S., Qin, C., & Bing, L. (2023). *Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework*. https://doi.org/10.48550/ARXIV.2305.03268

58) Zheng, C., Sun, K., Wu, H., Xi, C., & Zhou, X. (2024). *Balancing Enhancement, Harmlessness, and General Capabilities: Enhancing Conversational LLMs with Direct RLHF*. https://doi.org/10.48550/ARXIV.2403.02513

59) Zheng, S., Huang, J., & Chang, K. C.-C. (2023). *Why Does ChatGPT Fall Short in Providing Truthful Answers?* https://doi.org/10.48550/ARXIV.2304.10513

60) Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., & Yao, H. (2023). *Analyzing and Mitigating Object Hallucination in Large Vision-Language Models*. https://doi.org/10.48550/ARXIV.2310.00754