A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

LEI HUANG, Harbin Institute of Technology, Harbin, China
WEIJIANG YU, Huawei Inc., Shenzhen, China
WEITAO MA, WEIHONG ZHONG, ZHANGYIN FENG, and HAOTIAN WANG, Harbin Institute of Technology, Harbin, China
QIANGLONG CHEN and WEIHUA PENG, Huawei Inc., Shenzhen, China
XIAOCHENG FENG, BING QIN, and TING LIU, Harbin Institute of Technology, Harbin, China

The emergence of large language models (LLMs) has marked a significant breakthrough in natural language processing (NLP), fueling a paradigm shift in information acquisition. Nevertheless, LLMs are prone to hallucination, generating plausible yet nonfactual content. This phenomenon raises significant concerns over the reliability of LLMs in real-world information retrieval (IR) systems and has attracted intensive research to detect and mitigate such hallucinations. Given the open-ended general-purpose attributes inherent to LLMs, LLM hallucinations present distinct challenges that diverge from prior task-specific models. This divergence highlights the urgency for a nuanced understanding and comprehensive overview of recent advances in LLM hallucinations. In this survey, we begin with an innovative taxonomy of hallucination in the era of LLM and then delve into the factors contributing to hallucinations. Subsequently, we present a thorough overview of hallucination detection methods and benchmarks. Our discussion then transfers to representative methodologies for mitigating LLM hallucinations. Additionally, we delve into the current limitations faced by retrieval-augmented LLMs in combating hallucinations, offering insights for developing more robust IR systems. Finally, we highlight the promising research directions on LLM hallucinations, including hallucination in large vision-language models and understanding of knowledge boundaries in LLM hallucinations.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; • **General and reference** → **Surveys and overviews**;

Additional Key Words and Phrases: Large Language Models, Hallucination, Factuality, Faithfulness

Authors' Contact Information: Lei Huang, Harbin Institute of Technology, Harbin, China; e-mail: lhuang@ir.hit.edu.cn; Weijiang Yu, Huawei Inc., Shenzhen, China; e-mail: weijiangyu8@gmail.com; Weitao Ma, Harbin Institute of Technology, Harbin, China; e-mail: wtma@ir.hit.edu.cn; Weihong Zhong, Harbin Institute of Technology, Harbin, China; e-mail: whzhong@ir.hit.edu.cn; Zhangyin Feng, Harbin Institute of Technology, Harbin, China; e-mail: zyfeng@ir.hit.edu.cn; Haotian Wang, Harbin Institute of Technology, Harbin, China; e-mail: wanght1998@gmail.com; Qianglong Chen, Huawei Inc., Shenzhen, China; e-mail: chenqianglong.ai@gmail.com; Weihua Peng, Huawei Inc., Shenzhen, China; e-mail: pengwh.hit@gmail.com; Xiaocheng Feng (corresponding author), Harbin Institute of Technology, Harbin, China; e-mail: xcfeng@ir.hit.edu.cn; Bing Qin, Harbin Institute of Technology, Harbin, China; e-mail: qinb@ir.hit.edu.cn; Ting Liu, Harbin Institute of Technology, Harbin, China; e-mail: tliu@ir.hit.edu.cn.

ACM 1558-2868/2025/1-ART42
https://doi.org/10.1145/3703155

ACM Reference format:
Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (January 2025), 55 pages.
https://doi.org/10.1145/3703155

1 Introduction

Recently, the emergence of **large language models (LLMs)** [379], exemplified by LLaMA [295, 296], Claude [9], Gemini [7, 256] and GPT-4 [229], has ushered in a significant paradigm shift in **natural language processing (NLP)**, achieving unprecedented progress in language understanding [115, 123], generation [369, 389] and reasoning [57, 148, 247, 322, 350]. Furthermore, the extensive factual knowledge encoded within LLMs has demonstrated considerable advancements in leveraging LLMs for information seeking [6, 243], potentially reshaping the landscape of **information retrieval (IR)** systems [390]. Nevertheless, in tandem with these remarkable advancements, concerns have arisen about the tendency of LLMs to generate hallucinations [15, 104], resulting in seemingly plausible yet factually unsupported content. Further compounding this issue is the capability of LLMs to generate highly convincing and human-like responses [261], which makes detecting these hallucinations particularly challenging, thereby complicating the practical deployment of LLMs, especially real-world IR systems that have integrated into our daily lives like chatbots [8, 228], search engines [4, 211], and recommender systems [96, 168]. Given that the information provided by these systems can directly influence decision-making, any misleading information has the potential to spread false beliefs, or even cause harm.

Notably, hallucinations in conventional **natural language generation (NLG)** tasks have been extensively studied [124, 134], with hallucinations defined as generated content that is either nonsensical or unfaithful to the provided source content. These hallucinations are categorized into two types: *intrinsic hallucination*, where the generated output contradicts the source content, and *extrinsic hallucination*, where the generated output cannot be verified from the source. However, given their remarkable versatility across tasks [15, 30], understanding hallucinations in LLMs presents a unique challenge compared to models tailored for specific tasks. Besides, as LLMs typically function as open-ended systems, the scope of hallucination encompasses a broader concept, predominantly manifesting factual errors. This shift necessitates a reevaluation and adjustment of the existing taxonomy of hallucinations, aiming to enhance its adaptability in the evolving landscape of LLMs.

In this survey, we propose a redefined taxonomy of hallucination tailored specifically for applications involving LLMs. We categorize hallucination into two primary types: *factuality hallucination* and *faithfulness hallucination*. *Factuality hallucination* emphasizes the discrepancy between generated content and verifiable real-world facts, typically manifesting as factual inconsistencies. Conversely, *faithfulness hallucination* captures the divergence of generated content from user input or the lack of self-consistency within the generated content. This category is further subdivided into instruction inconsistency, where the content deviates from the user's original instruction; context inconsistency, highlighting discrepancies from the provided context; and logical inconsistency, pointing out internal contradictions within the content. Such categorization refines our understanding of hallucinations in LLMs, aligning it closely with their contemporary usage.

Delving into the underlying causes of hallucinations in LLMs is essential not merely for enhancing the comprehension of these phenomena but also for informing strategies aimed at alleviating them. Recognizing the multifaceted sources of LLM hallucinations, our survey identifies potential

contributors into three main aspects: data, training, and inference stages. This categorization allows us to span a broad spectrum of factors, providing a holistic view of the origins and mechanisms by which hallucinations may arise within LLM systems. Furthermore, we comprehensively outline a variety of effective detection methods specifically devised for detecting hallucinations in LLMs, as well as an exhaustive overview of benchmarks related to LLM hallucinations, serving as appropriate testbeds to assess the extent of hallucinations generated by LLMs and the efficacy of detection methods. Beyond evaluation, significant efforts have been undertaken to mitigate hallucinations of LLMs. These initiatives are comprehensively surveyed in our study, in accordance with the corresponding causes, spanning from data-related, training-related, and inference-related approaches. In addition, the effectiveness of **retrieval-augmented generation (RAG)** in mitigating hallucinations has garnered tremendous attention within the field. Despite the considerable potential of RAG, current systems inherently face limitations and even suffer from hallucinations. Accordingly, our survey undertakes an in-depth analysis of these challenges, aiming to provide valuable insights aimed at developing more robust RAG systems. We also highlight several promising avenues for future research, such as hallucinations in **large vision-language models (LVLMs)** and understanding of knowledge boundaries in LLM hallucinations, paving the way for forthcoming research in the field.

Comparing with Existing Surveys. As hallucination stands out as a major challenge in generative AI, numerous research [134, 189, 255, 294, 308, 372] have been directed towards hallucinations. While these contributions have explored LLM hallucination from various perspectives and provided valuable insights, our survey seeks to delineate their distinct contributions and the comprehensive scope they encompass. Ji et al. [134] primarily shed light on hallucinations in pre-trained models for NLG tasks, leaving LLMs outside their discussion purview. Tonmoy et al. [294] mainly focused on discussing the mitigation strategies combating LLM hallucinations. Besides, Liu et al. [189] took a broader view of LLM trustworthiness without delving into specific hallucination phenomena, whereas Wang et al. [308] provided an in-depth look at factuality in LLMs. However, our work narrows down to a critical subset of trustworthiness challenges, specifically addressing factuality and extending the discussion to include faithfulness hallucinations. To the best of our knowledge, Zhang et al. [372] presented research closely aligned with ours, detailing LLM hallucination taxonomies, evaluation benchmarks, and mitigation strategies. However, our survey sets itself apart through a unique taxonomy and organizational structure. We present a detailed, layered classification of hallucinations and conduct a more comprehensive analysis of the causes of hallucinations. Crucially, our proposed mitigation strategies are directly tied to these causes, offering a targeted and coherent framework for addressing LLM hallucinations.

Organization of this Survey. In this survey, we present a comprehensive overview of the latest developments in LLM hallucinations, as shown in Figure 1. We commence by constructing a taxonomy of hallucinations in the realm of LLM (Section 2). Subsequently, we analyze factors contributing to LLM hallucinations in depth (Section 3), followed by a review of various strategies and benchmarks employed for the reliable detection of hallucinations in LLMs (Section 4). We then detail a spectrum of approaches designed to mitigate these hallucinations (Section 5). Concluding, we delve into the challenges faced by current RAG systems (Section 6) and delineate potential pathways for forthcoming research (Section 7).

2 Definitions

For the sake of a comprehensive understanding of hallucinations in LLMs, we commence with a succinct introduction to LLMs (Section 2.1), delineating the scope of this survey. Subsequently, we delve into the training stages of LLMs (Section 2.2), as a thorough understanding of the training

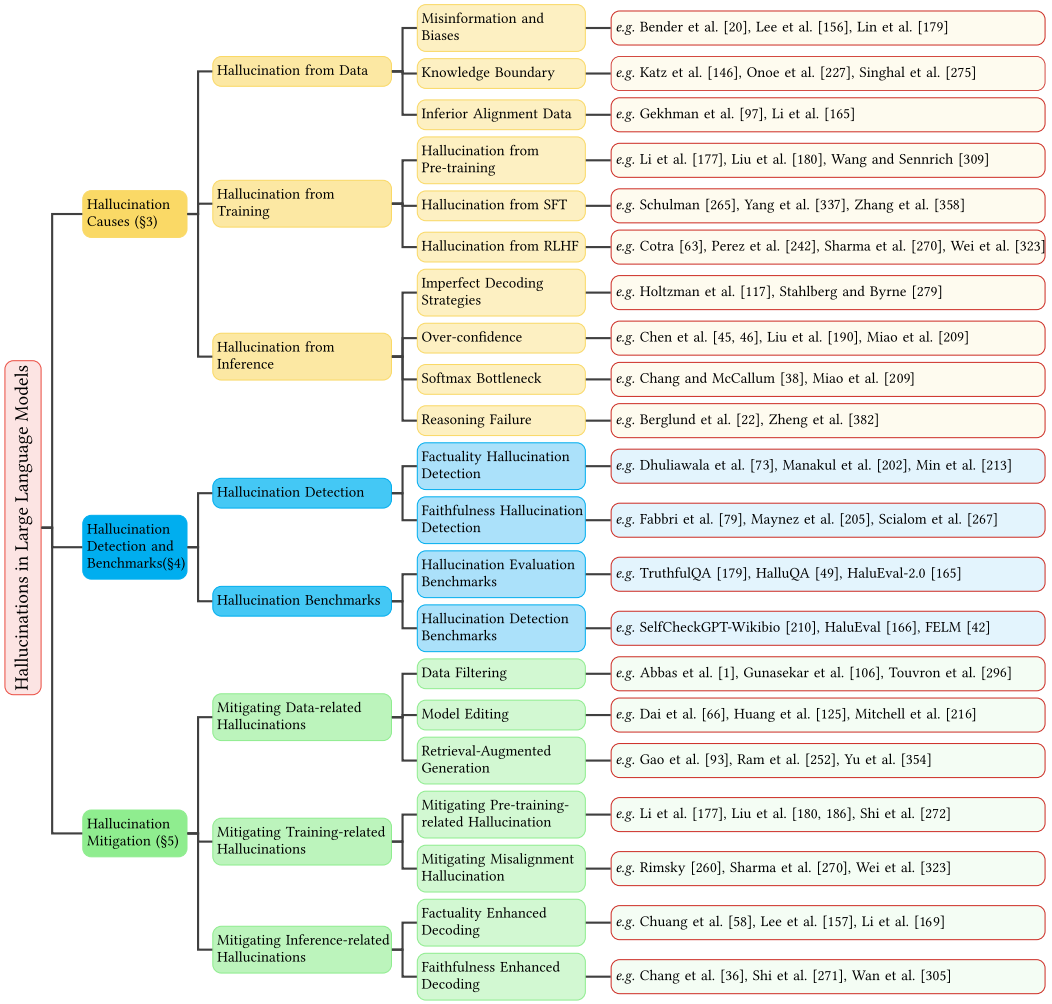

Fig. 1. The main content flow and categorization of this survey.

mechanisms contributes significantly to elucidating the origins of hallucinations. Lastly, we expound upon the concept of hallucinations in LLMs (Section 2.3), further categorizing it into two distinct types.

2.1 LLMs

Before delving into the causes of hallucination, we first introduce the concept of LLMs. Typically, LLMs refer to a series of general-purpose models that leverage the transformer-based language model architecture and undergo extensive training on massive textual corpora with notable examples including GPT-3 [29], PaLM [54], LLaMA [296], GPT-4 [229] and Gemini [256]. By scaling the amount of data and model capacity, LLMs raise amazing emergent abilities, typically including in-context learning [29], chain-of-thought prompting [322] and instruction following [241].

2.2 Training Stages of LLMs

The attributes and behaviors of LLMs are deeply intertwined with their training processes. LLMs undergo three primary training stages: pre-training, **supervised fine-tuning (SFT)**, and **reinforcement learning from human feedback (RLHF)**. Analyzing these stages provides insight into hallucination origins in LLMs, as each stage equips the model with specific capabilities.

2.2.1 Pre-training. Pre-training is widely acknowledged as a foundational stage for LLM to acquire knowledge and capabilities [384]. During this phase, LLMs engage in autoregressive prediction of subsequent tokens within sequences. Through self-supervised training on extensive textual corpora, LLMs acquire knowledge of language syntax, world knowledge, and reasoning abilities, thereby laying a solid groundwork for further fine-tuning. Besides, recent research [71, 287] suggests that predicting subsequent words is akin to losslessly compressing significant information. The essence of LLMs lies in predicting the probability distribution for upcoming words. Accurate predictions indicate a profound grasp of knowledge, translating to a nuanced understanding of the world.

2.2.2 SFT. While LLMs acquire substantial knowledge and capabilities during the pre-training stage, it's crucial to recognize that pre-training primarily optimizes for completion. Consequently, pre-trained LLMs fundamentally serve as completion machines, which can lead to a misalignment between the next-word prediction objective of LLMs and the user's objective of obtaining desired responses. To bridge this gap, SFT [366] has been introduced, which involves further training LLMs using a meticulously annotated set of (instruction, response) pairs, resulting in enhanced capabilities and improved controllability of LLMs. Furthermore, recent studies [59, 127] have confirmed the effectiveness of SFT to achieve exceptional performance on unseen tasks, showcasing their remarkable generalization abilities.

2.2.3 RLHF. While the SFT process successfully enables LLMs to follow user instructions, there is still room for them to better align with human preferences. Among various methods that utilize human feedback, RLHF stands out as a representative solution for aligning with human preferences through reinforcement learning [55, 230, 281]. Typically, RLHF employs a preference model [26] trained to predict preference rankings given a prompt alongside a pair of human-labeled responses. To align with human preferences, RLHF optimizes the LLM to generate outputs that maximize the reward provided by the trained preference model, typically employing a reinforcement learning algorithm, such as Proximal Policy Optimization [266]. Such integration of human feedback into the training loop has proven effective in enhancing the alignment of LLMs, guiding them toward producing high-quality and harmless responses.

2.3 Hallucinations in LLMs

The concept of hallucination traces its roots to the fields of pathology and psychology and is defined as *the perception of an entity or event that is absent in reality* [199]. Within the realm of NLP, hallucination is typically referred to as a phenomenon in which the generated content appears nonsensical or unfaithful to the provided source content [88, 205]. This concept bears a loose resemblance to the phenomenon of hallucination observed in human psychology. Generally, hallucinations in NLG tasks can be categorized into two primary types: *intrinsic hallucination* and *extrinsic hallucination* [134, 171].

Specifically, *intrinsic hallucinations* pertain to the model outputs that directly conflict with the provided source context. On the other hand, *extrinsic hallucinations* involve outputs that cannot be verified using the provided source context or external knowledge bases. This means the generated

Table 1. Examples of Each Category of LLM Hallucinations

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	*Instruction*: What are Thomas Edison's main contributions to science and technology? *Response*: Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, *Edison improved the light bulb, building on earlier designs by others*, and *Alexander Graham Bell invented the telephone.*
	Factual Fabrication	*Instruction*: What are the major environmental impacts of the construction of the Eiffel Tower? *Response*: The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the "Parisian tiger" and sparked the global green architecture movement. In fact, *there is no evidence of a "Parisian tiger" ever existing, making this a fabricated claim.* Moreover, *attributing the origins of green architecture to the Eiffel Tower is an exaggeration*, as this movement has diverse roots and cannot be traced back to a single event.
Faithfulness Hallucination	Instruction Inconsistency	*Instruction*: Translate the English question into Spanish: "What is the capital of France?" *Response*: The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. *Instead of performing the translation as instructed, the LLM chose to answer the question.*
	Context Inconsistency	*Retrieved context*: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. *Instruction*: Summarize the retrieval document above. *Response*: The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the *paragraph mentions that the Nile originates in the Great Lakes region of central Africa.*
	Logical Inconsistency	*Instruction*: Solve the equation $2x + 3 = 11$ step by step. *Response*: Step 1: Subtract 3 from both sides to get: $2x = 8$. Step 2: Divide both sides by 2 to get: $x = 3$. So the result of the equation is $x = 3$.	The first step correctly isolates the term with x to get $2x = 8$. However, *the next step inaccurately divides 8 by 2 to yield a result of $x = 3$*, which is inconsistent with the earlier reasoning.

Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination.

text is neither supported by nor directly contradicts the available information, rendering the output unverifiable and potentially misleading.

However, in the era of LLMs, the versatile capabilities of these models have facilitated their widespread use across diverse fields, highlighting limitations in existing task-specific categorization paradigms. Considering that LLMs place a significant emphasis on user-centric interactions and prioritize alignment with user directives, coupled with the fact that their hallucinations predominantly surface at factual levels, we introduce a more granular taxonomy building upon the foundational work by Ji et al. [134]. This refined taxonomy seeks to encapsulate the distinct intricacies associated with LLM hallucinations. To provide a more intuitive illustration of our definition of LLM hallucination, we present examples for each type of hallucination in Table 1, namely *factuality hallucination* and *faithfulness hallucination.*

2.3.1 Factuality Hallucination. The emergence of LLMs marks a significant shift from traditional task-specific toolkits to AI assistants that have a heightened focus on open-domain interactions. This shift is primarily attributed to their vast parametric factual knowledge. However, existing LLMs occasionally exhibit tendencies to produce outputs that are either inconsistent with real-world facts or unverifiable [165], posing challenges to the trustworthiness of artificial intelligence. In this context, we categorize these factuality hallucinations into two primary types:

Factual Contradiction refers to situations where the LLM's output contains facts that can be grounded in real-world information but present contradictions. This type of hallucination occurs most frequently and arises from diverse sources, encompassing the LLM's capture, storage, and expression of factual knowledge. Depending on the error type of contradictions, it can be further divided into two subcategories: *entity-error hallucination* and *relation-error hallucination*.

— *Entity-error hallucination* refers to the situations where the generated text of LLMs contains erroneous entities. As shown in Table 1, when asked about "*the inventor of the telephone*," the model erroneously states "*Thomas Edison*," conflicting with the real fact that it was "*Alexander Graham Bell*."
— *Relation-error hallucination* refers to instances where the generated text of LLMs contains wrong relations between entities. As shown in Table 1, when inquired about "*the inventor of the light bulb*," the model incorrectly claims "*Thomas Edison*," despite the fact that *he improved upon existing designs and did not invent it*.

Factual Fabrication refers to instances where the LLM's output contains facts that are unverifiable against established real-world knowledge. This can be further divided into *unverifiability hallucination* and *overclaim hallucination*.

— *Unverifiability hallucination* pertains to statements that are entirely non-existent or cannot be verified using available sources. As shown in Table 1, when asked about "*the major environmental impacts of the construction of the Eiffel Tower*," the model incorrectly states that "*the construction led to the extinction of the Parisian tiger*," a species that does not exist and thus, this claim cannot be substantiated by any historical or biological record.
— *Overclaim hallucination* involves claims that lack universal validity due to subjective biases. As shown in Table 1, the model claims that "*the Eiffel Tower's construction is widely recognized as the event that sparked the global green architecture movement*." This is an overclaim, as there is no broad consensus or substantial evidence to support the statement.

2.3.2 Faithfulness Hallucination. LLMs are inherently trained to align with user instructions. As the use of LLMs shifts towards more user-centric applications, ensuring their consistency with user-provided instructions and contextual information becomes increasingly vital. Furthermore, LLM's faithfulness is also reflected in the logical consistency of its generated content. From this perspective, we categorize three subtypes of faithfulness hallucinations:

Instruction inconsistency refers to the LLM's outputs that deviate from a user's directive. While some deviations might serve safety guidelines, the inconsistencies here signify unintentional misalignment with non-malicious user instructions. As described in Table 1, the user's actual intention is translation. However, the LLM erroneously deviated from the user's instruction and performed a question-answering task instead.

Context inconsistency points to instances where the LLM's output is unfaithful with the user's provided contextual information. For example, as shown in Table 1, the user mentioned the Nile's source being in the Great Lakes region of central Africa, yet the LLM's response contradicted the context.

Logical inconsistency underscores when LLM outputs exhibit internal logical contradictions, often observed in reasoning tasks. This manifests as inconsistency both among the reasoning steps themselves and between the steps and the final answer. For example, as shown in Table 1, while the reasoning step of dividing both sides of the equation by 2 is correct, the final answer of $x = 4$ is inconsistent with the reasoning chain, leading to an incorrect result.

3 Hallucination Causes

LLM hallucinations have multifaceted origins, spanning the entire spectrum of LLMs' capability acquisition process. In this section, we delve into the root causes of hallucinations in LLMs, primarily categorized into three key aspects: (1) *Data* (Section 3.1), (2) *Training* (Section 3.2), and (3) *Inference* (Section 3.3).

3.1 Hallucination from Data

Data for training LLMs are comprised of two primary components: (1) pre-training data, through which LLMs acquire their general capabilities and factual knowledge [384], and (2) alignment data, which teach LLMs to follow user instructions and align with human preferences [318]. Although these data constantly expand the capability boundaries of LLMs, they inadvertently become the principal contributors to LLM hallucinations. This primarily manifests in three aspects: the presence of misinformation and biases in the flawed pre-training data sources (Section 3.1.1), the knowledge boundary inherently bounded by the scope of the pre-training data (Section 3.1.2), and the hallucinations induced by inferior alignment data (Section 3.1.3).

3.1.1 Misinformation and Biases. Neural networks possess an intrinsic tendency to memorize training data [35], and this memorization tendency grows with model size [34, 54]. In general, the inherent memorization capability is a double-edged sword in the fight against hallucinations. On the one hand, the capacities of LLMs to memorize suggests their potential to capture profound world knowledge. On the other hand, it becomes problematic in the context of misinformation and biases present within pre-training data and may inadvertently be amplified, manifesting as *imitative falsehood* [179] and the reinforcement of societal biases. For a more comprehensive understanding, detailed examples are presented in Table 2.

Imitative Falsehood. Misinformation such as fake news and unfounded rumors has been widely spread among social media platforms and gradually serves as a significant contributor to LLM hallucinations. The increasing demand for large-scale corpora for pre-training necessitates the employment of heuristic data collection methods. While facilitating the acquisition of extensive data, challenges arise in maintaining consistent data quality, which inevitably introduces such misinformation in pre-training data [20, 324]. In this situation, LLMs will increase the likelihood of generating such false statements due to their remarkable memorization capabilities, leading to *imitative falsehoods*. The issue is further exacerbated by the fact that LLMs have drastically lowered the barriers to content creation, posing considerable risks to the trustworthiness of public discourse and internet ecosystems.

Societal Biases. In addition to misinformation, biases are also deeply rooted in social media platforms, showing up variously, like biased hiring, prejudiced news, and hate-spewing. As a purely negative phenomenon, biases and hallucinations have obvious differences, mainly manifested as stereotypes spreading and social inequalities. However, certain biases are intrinsically tied to hallucinations, especially those related to gender [239] and nationality [152, 223]. For instance, LLMs might associate the profession of nursing with females, even when gender isn't explicitly mentioned in the user-provided context, exemplifying context inconsistency as discussed in Section 2.3. Such biases can be inadvertently acquired from internet-based texts, which are rife with diverse and biased viewpoints and subsequently be propagated into the generated content [152].

3.1.2 Knowledge Boundary. While the vast pre-training corpora have empowered LLMs with extensive factual knowledge, they inherently possess knowledge boundaries. These boundaries arise primarily from two sources: (1) the inability of LLMs to memorize all factual knowledge encountered

Table 2. Examples of Hallucinations from Imitative Falsehoods and Social Biases

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	*While Edison improved the light bulb's design, he was not its sole inventor.* The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Societal Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts.	Dr. Kim from South Korea recently published a paper on quantum computing, detailing new algorithms and their applications.	*The model erroneously added "South Korea" based on the name "Kim"* revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

Each category in the table is accompanied by a premise outlining the data issue, user input, and the LLM's hallucinatory output marked in Red (displayed in red font), and an explanation for the occurrence, aiding comprehension of these complex phenomena.

during pre-training, especially the less frequent long-tail knowledge, and (2) the intrinsic boundary of the pre-training data itself, which does not include rapidly evolving world knowledge or content restricted by copyright laws. Consequently, when LLMs encounter information that falls outside their limited knowledge boundaries, they are more susceptible to generating hallucinations. We present detailed examples for clear illustration in Table 3.

Long-tail Knowledge. The distribution of knowledge within the pre-training corpora is inherently non-uniform, which results in LLMs demonstrating varying levels of proficiency across different types of knowledge. Recent studies have highlighted a strong correlation between the model's accuracy on general domain questions and the volume of relevant documents [143] or entity popularity [201] within the pre-training corpora. Furthermore, given that LLMs are predominantly trained on extensive general domain corpora [92, 240, 251], they may exhibit deficits in domain-specific knowledge. This limitation becomes particularly evident when LLMs are confronted with tasks that require domain-specific expertise, such as medical [176, 275] and legal [146, 349] questions; these models may exhibit pronounced hallucinations, often manifesting as factual fabrication.

Up-to-date Knowledge. Beyond the shortfall in long-tail knowledge, another intrinsic limitation concerning the knowledge boundaries within LLMs is their constrained capacity for up-to-date knowledge. The factual knowledge embedded within LLMs exhibits clear temporal boundaries and can become outdated over time [145, 163, 227]. Once these models are trained, their internal knowledge is never updated. This poses a challenge given the dynamic and ever-evolving nature of our world. When confronted with queries that transcend their temporal scope, LLMs often resort to fabricating facts or providing answers that might have been correct in the past but are now outdated.

Table 3. Example of Knowledge Boundary

Type	Input	Model Ouput	Explanation
Long-tail Knowledge	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	The correct cause of phenylketonuria is mutations in the PAH gene. *Such specialized medical information is less frequently encountered in the training corpus, making LLMs struggle to memorize.*
Up-to-date Knowledge	Who will be the Democratic Party candidate for the 2024 US presidential election?	Michelle Obama has announced her candidacy and is leading in the polls, with significant endorsements from key Democratic figures such as Barack Obama and Bernie Sanders. She has also outlined a policy platform focusing on healthcare reform and climate change initiatives.	After President Joe Biden announced his withdrawal from the race on July 21, 2024, he endorsed Kamala Harris as the Democratic Party candidate. *Such hallucinations arise because LLMs' parametric knowledge cannot be updated in real-time, leading to potential fabrication regarding current events.*
Copyright-sensitive Knowledge	Describe the ending of the novel "The Secrets of the Alchemist" that was published last year.	In the dramatic conclusion of "The Secrets of the Alchemist," the protagonist discovers the true secret behind the philosopher's stone, leading to a peaceful resolution with all characters sharing in the wisdom.	*The novel "The Secrets of the Alchemist" is under copyright protection, and LLMs have not been trained directly on such copyrighted materials.* Thus, the model's output fabricates details about the book's ending.

Content marked in Red represents the hallucinatory output.

Copyright-sensitive Knowledge. Due to licensing restrictions [258], existing LLMs are legally constrained to training on corpora that are publicly licensed [62, 92] or otherwise available for use without infringing copyright laws [10, 114]. This limitation significantly impacts the breadth and diversity of knowledge that LLMs can legally acquire. A significant portion of valuable knowledge, encapsulated in copyrighted materials such as recent scientific research, proprietary data, and copyrighted literary works, remains inaccessible to LLMs. This exclusion creates a knowledge gap, leading to potential hallucinations when LLMs attempt to generate information in domains where their training data is inaccessible [212].

3.1.3 Inferior Alignment Data. After the pre-training stage, LLMs have embedded substantial factual knowledge within their parameters, thereby establishing obvious knowledge boundaries. During the SFT stage, LLMs are typically trained on instruction pairs labeled by human annotators, potentially introducing new factual knowledge that extends beyond the knowledge boundary established during pre-training. Gekhman et al. [97] analyzed the training dynamics of incorporating new factual knowledge during the SFT process and found that LLMs struggle to acquire such new knowledge effectively. Most importantly, they discovered a correlation between the acquisition of new knowledge through SFT and increased hallucinations, suggesting that introducing new factual knowledge encourages LLMs to hallucinate. Additionally, Li et al. [165] conducted extensive analysis on the effect of instructions in producing hallucinations. Findings indicated that task-specific instructions which primarily focus on task format learning, tend to yield a higher proportion of hallucinatory responses. Moreover, overly complex and diverse instructions also lead to increased hallucinations.

3.2 Hallucination from Training

As detailed in Section 2.2, the distinct stages of training impart various capabilities to LLMs, with pre-training focusing on acquiring general-purpose representations and world knowledge, and alignment enables LLMs to better align with user instructions and preferences. While these stages are critical for equipping LLMs with remarkable capabilities, shortfalls in either stage can inadvertently pave the way for hallucinations.

3.2.1 Hallucination from Pre-training. Pre-training constitutes the foundational stage for LLMs, predominantly utilizing a transformer-based architecture following the paradigm established by GPT [29, 248, 249], and further developed by OPT [368], Falcon [240], and Llama-2 [296]. This stage employs a causal language modeling objective, where models learn to predict subsequent tokens solely based on preceding ones in a unidirectional, left-to-right manner. While facilitating efficient training, it inherently limits the ability to capture intricate contextual dependencies, potentially increasing risks for the emergence of hallucination [177]. Moreover, recent research has exposed that LLMs can occasionally exhibit unpredictable reasoning hallucinations spanning both long-range and short-range dependencies, which potentially arise from the limitations of soft attention [52, 110], where attention becomes diluted across positions as sequence length increases. Notably, the phenomenon of exposure bias [21, 253] has been a longstanding and serious contribution to hallucinations, resulting from the disparity between training and inference in the auto-regressive generative model. Such inconsistency can result in hallucinations [309], especially when an erroneous token generated by the model cascades errors throughout the subsequent sequence, akin to a snowball effect [364].

3.2.2 Hallucination from SFT. LLMs have inherent capability boundaries established during pre-training. SFT seeks to utilize instruction data and corresponding responses to unlock these pre-acquired abilities. However, challenges arise when the demands of annotated instructions exceed the model's pre-defined capability boundaries. In such cases, LLMs are trained to fit responses beyond their actual knowledge boundaries. As discussed in Section 3.1.3, over-fitting on new factual knowledge encourages LLMs prone to fabricating content, amplifying the risk of hallucinations [97, 265]. Moreover, another significant reason lies in the models' inability to reject. Traditional SFT methods typically force models to complete each response, without allowing them to accurately express uncertainty [337, 358]. Consequently, when faced with queries that exceed their knowledge boundaries, these models are more likely to fabricate content rather than reject it. This misalignment of knowledge boundaries, coupled with the inability to express uncertainty, are critical factors that contribute to the occurrence of hallucinations during the SFT stage.

3.2.3 Hallucination from RLHF. Several studies [13, 31] have demonstrated that LLM's activations encapsulate an internal belief related to the truthfulness of its generated statements. Nevertheless, misalignment can occasionally arise between these internal beliefs and the generated outputs. Even when LLMs are refined with human feedback [230], they can sometimes produce outputs that diverge from their internal beliefs. Such behaviors, termed as sycophancy [63], underscore the model's inclination to appease human evaluators, often at the cost of truthfulness. Recent studies indicate that models trained via RLHF exhibit pronounced behaviors of pandering to user opinions. Such sycophantic behaviors are not restricted to ambiguous questions without definitive answers [242], like political stances, but can also arise when the model chooses a clearly incorrect answer, despite being aware of its inaccuracy [323]. Delving into this phenomenon, Sharma et al. [270] suggested that the root of sycophancy may lie in the training process of RLHF models. By further exploring the role of human preferences in this behavior, the research indicates that the tendency

for sycophancy is likely driven by both humans and preference models showing a bias towards sycophantic responses over truthful ones.

3.3 Hallucination from Inference

Decoding plays an important role in manifesting the capabilities of LLMs after pretraining and alignment. However, certain shortcomings in decoding strategies can lead to LLM hallucinations.

3.3.1 Imperfect Decoding Strategies. LLMs have demonstrated a remarkable aptitude for generating highly creative and diverse content, a proficiency that is critically dependent on the pivotal role of *randomness* in their decoding strategies. Stochastic sampling [83, 117] is currently the prevailing decoding strategy employed by these LLMs. The rationale for incorporating randomness into decoding strategies stems from the realization that high likelihood sequences often result in surprisingly low-quality text, which is called *likelihood trap* [117, 206, 279, 359]. The diversity introduced by the randomness in decoding strategies comes at a cost, as it is positively correlated with an increased risk of hallucinations [58, 77]. An elevation in the sampling temperature results in a more uniform token probability distribution, increasing the likelihood of sampling tokens with lower frequencies from the tail of the distribution. Consequently, this heightened tendency to sample infrequently occurring tokens exacerbates the risk of hallucinations [5].

3.3.2 Over-confidence. Prior studies in conditional text generation [45, 209] have highlighted the issue of *over-confidence* which stems from an excessive focus on the partially generated content, often prioritizing fluency at the expense of faithfully adhering to the source context. While LLMs, primarily adopting the causal language model architecture, have gained widespread usage, the *over-confidence* phenomenon continues to persist. During the generation process, the prediction of the next word is conditioned on both the language model context and the partially generated text. However, as demonstrated in prior studies [19, 186, 303], language models often exhibit a localized focus within their attention mechanisms, giving priority to nearby words and resulting in a notable deficit in context attention [271]. Furthermore, this concern is further amplified in LLMs that exhibit a proclivity for generating lengthy and comprehensive responses. In such cases, there is even a heightened susceptibility to the risk of instruction forgetting [46, 190]. This insufficient attention can directly contribute to faithfulness hallucinations, wherein the model outputs content that deviates from the original context.

3.3.3 Softmax Bottleneck. The majority of language models utilize a softmax layer that operates on the final layer's representation within the language model, in conjunction with a word embedding, to compute the ultimate probability associated with word prediction. Nevertheless, the efficacy of Softmax-based language models is impeded by a recognized limitation known as the *Softmax bottleneck* [338], wherein the employment of softmax in tandem with distributed word embeddings constrains the expressivity of the output probability distributions given the context which prevents LMs from outputting the desired distribution. Additionally, Chang and McCallum [38] discovered that when the desired distribution within the output word embedding space exhibits multiple modes, language models face challenges in accurately prioritizing words from all the modes as the top next words, which also introduces the risk of hallucination.

3.3.4 Reasoning Failure. Beyond the challenges with long-tail knowledge, effective utilization of knowledge is inextricably linked with reasoning capabilities. For instance, in multi-hop question-answering scenarios, even if the LLM possesses the necessary knowledge, it may struggle to produce accurate results if multiple associations exist between questions, due to its limitations in reasoning [382]. Furthermore, Berglund et al. [22] unveiled a specific reasoning failure in LLMs termed the *Reversal Curse*. Specifically, while the model can correctly answer when the question is

formulated as "A is B," it exhibits a failed logical deduction when asked the converse "B is A." This discrepancy in reasoning extends beyond simple deductions.

4 Hallucination Detection and Benchmarks

The issue of hallucinations within LLMs has garnered considerable attention, raising concerns about the reliability of LLMs and their deployment in practical applications. As LLMs become increasingly adept at generating human-like text, accurately distinguishing between hallucinated versus factual content becomes increasingly vital. Moreover, effectively measuring the level of hallucination in LLM is crucial for improving their reliability. Thus, in this section, we delve into hallucination detection approaches (Section 4.1) and benchmarks for assessing LLM hallucinations (Section 4.2).

4.1 Hallucination Detection

Existing strategies for detecting hallucinations in LLMs can be categorized based on the type of hallucination: (1) factuality hallucination detection, which aims to identify factual inaccuracies in the model's outputs, and (2) faithfulness hallucination detection, which focuses on evaluating the faithfulness of model's outputs to the contextual information provided.

4.1.1 Factuality Hallucination Detection. Factuality hallucination detection involves assessing whether the output of LLMs aligns with real-world facts. Typical methods generally fall into two categories: *fact-checking*, which involves verifying the factuality of the generated response against trusted knowledge sources, and *uncertainty estimation*, which focuses on detecting factual inconsistency via internal uncertainty signals.

Fact-checking. Given that the output of LLMs is typically comprehensive and consists of multiple factual statements, the fact-checking approach is generally divided into two primary steps: (1) fact extraction, which involves extracting independent factual statements within the model's outputs, (2) fact verification, which aims at verifying the correctness of these factual statements against trusted knowledge sources. Depending on the type of knowledge sources employed for verification, fact-checking methodologies can be broadly categorized into two distinct parts: *external retrieval* and *internal checking*.

— *External retrieval*: The most intuitive strategy for fact verification is external retrieval. Min et al. [213] developed FACTSCORE, a fine-grained factual metric tailored for evaluating long-form text generation. It first decomposes the generation content into atomic facts and subsequently computes the percentage supported by reliable knowledge sources. Expanding on this concept, Chern et al. [50] proposed a unified framework that equips LLMs with the capability to identify factual inaccuracies by utilizing a collection of external tools dedicated to evidence gathering. In addition to retrieving supporting evidence solely based on decomposited claims, Huo et al. [126] improved the retrieval process through query expansion. By combining the original question with the LLM-generated answer, they effectively addressed the issue of topic drift, ensuring that the retrieved evidence aligns with both the question and the LLM's response.
— *Internal checking*: Given the extensive factual knowledge encoded in their parameters, LLMs have been explored as factual knowledge sources for fact-checking. Dhuliawala et al. [73] introduced the **Chain-of-Verification (CoVe)**, where an LLM first generates verification questions for a draft response and subsequently leverages its parametric knowledge to assess the consistency of the answer against the original response, thereby detecting potential inconsistencies. Kadavath et al. [141] and Zhang et al. [371] calculate the probability $p(True)$ to assess the factuality of the response to a boolean question, relying exclusively on the model'sinternal knowledge. Additionally, Li et al. [165] observed that most atomic statements

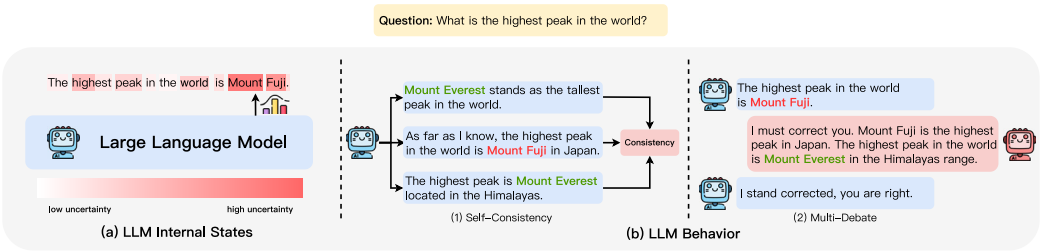

Fig. 2. Taxonomy of uncertainty estimation methods in factual hallucination detection, featuring (a) *LLM Internal States* and (b) *LLM Behavior*, with LLM behavior encompassing two main categories: self-consistency and multi-debate.

are interrelated, some may serve as contextual backgrounds for others, which potentially leads to incorrect judgments. Thus, they instruct the LLM to directly predict hallucination judgments considering all factual statements. However, as LLMs are not inherently reliable factual databases [381], solely relying on LLMs' parametric knowledge for fact-checking may result in inaccurate assessments.

Uncertainty Estimation. While many approaches to hallucination detection rely on external knowledge sources for fact-checking, several methods have been devised to address this issue in zero-resource settings, thus eliminating the need for retrieval. The foundational premise behind these strategies is that the origin of LLM hallucinations is inherently tied to the model's uncertainty. Therefore, by estimating the uncertainty of the factual content generated by the model, it becomes feasible to detect hallucinations. The methodologies in uncertainty estimation can broadly be categorized into two approaches: based on *LLM internal states* and *LLM behavior*, as shown in Figure 2.

—*LLM internal states*: The internal states of LLMs can serve as informative indicators of their uncertainty, often manifested through metrics like token probability or entropy. Varshney et al. [302] determined the model's uncertainty towards key concepts quantified by considering the minimal token probability within those concepts. The underlying rationale is that a low probability serves as a strong indicator of the model's uncertainty, with less influence from higher probability tokens present in the concept. Similarly, Luo et al. [195] employed a self-evaluation-based approach for uncertainty estimation by grounding in the rationale that a language model's ability to adeptly reconstruct an original concept from its generated explanation is indicative of its proficiency with that concept. By initially prompting the model to generate an explanation for a given concept and then employing constrained decoding to have the model recreate the original concept based on its generated explanation, the probability score from the response sequence can serve as a familiarity score for the concept. Furthermore, Yao et al. [341] interpreted hallucination through the lens of adversarial attacks. Utilizing gradient-based token replacement, they devised prompts to induce hallucinations. Notably, they observed that the first token generated from a raw prompt typically exhibits low entropy compared to those from adversarial attacks. Based on this observation, they proposed setting an entropy threshold to define such hallucination attacks.

—*LLM behavior*: However, when systems are only accessible via API calls [99, 211, 228], access to the output's token-level probability distribution might be unavailable. Given this constraint, several studies have shifted their focus to probing a model's uncertainty, either through natural language prompts [141, 331] or by examining its behavioral manifestations. For instance, by

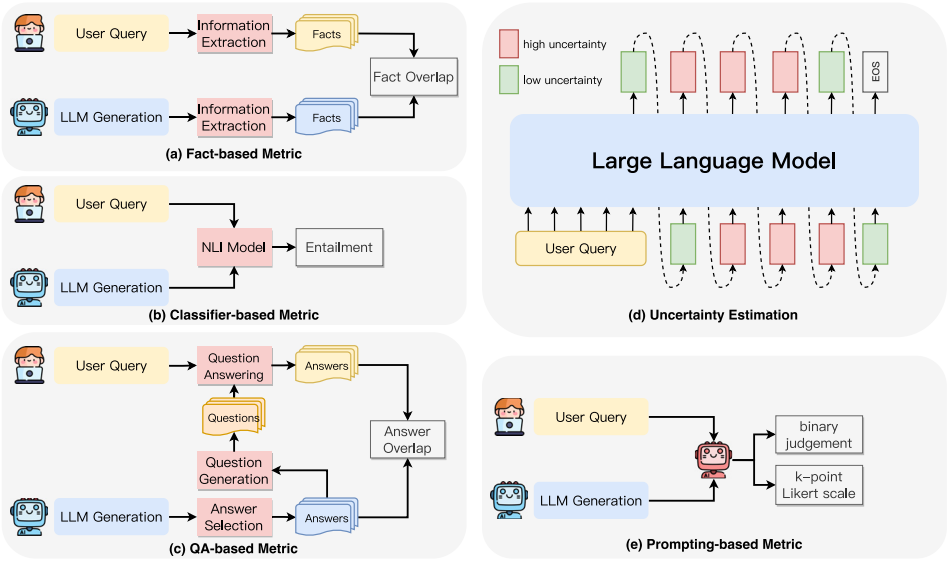

Fig. 3. The illustration of detection methods for faithfulness hallucinations: (a) *Fact-based Metrics*, which assesses faithfulness by measuring the overlap of facts between the generated content and the source content; (b) *Classifier-based Metrics*, utilizing trained classifiers to distinguish the level of entailment between the generated content and the source content; (c) *QA-based Metrics*, employing question-answering systems to validate the consistency of information between the source content and the generated content; (d) *Uncertainty Estimation*, which assesses faithfulness by measuring the model's confidence in its generated outputs; (e) *Prompting-based Metrics*, wherein LLMs are induced to serve as evaluators, assessing the faithfulness of generated content through specific prompting strategies.

sampling multiple responses from an LLM for the same prompt, Manakul et al. [202] detected hallucinations via evaluating the consistency among the factual statements. However, these methods predominantly rely on direct queries that explicitly solicit information or verification from the model. Agrawal et al. [3], inspired by investigative interviews, advocated for the use of indirect queries. Unlike direct ones, these indirect counterparts often pose open-ended questions to elicit specific information. By employing these indirect queries, consistency across multiple model generations can be better evaluated. Beyond assessing uncertainty from the self-consistency of a single LLM's multiple generations, one can embrace a multi-agent perspective by incorporating additional LLMs. Drawing inspiration from legal cross-examination practices, Cohen et al. [61] introduced the LMvLM approach. This strategy leverages an examiner LM to question an examinee LM, aiming to unveil inconsistencies of claims during multi-turn interaction.

4.1.2 Faithfulness Hallucination Detection. Ensuring the faithfulness of LLMs to provide context or user instructions is pivotal for their practical utility in IR applications, from conversational search to interactive dialogue systems. We categorize existing hallucination detection metrics tailored to faithfulness into the following groups, with an overview shown in Figure 3: (1) Fact-based, (2) Classifier-based, (3) QA-based, (4) Uncertainty-based, and (5) LLM-based.

Fact-based Metrics. In the realm of assessing faithfulness, one of the most intuitive methods involves measuring the overlap of pivotal facts between the generated content and the source content. Given the diverse manifestations of facts, faithfulness can be measured based on *n-gram*,

entities, and *relation triples*. Traditional *n-gram-based* metrics, such as BLEU [236], ROUGE [178], and PARENT-T [320], typically fall short in differentiating the nuanced discrepancies between the generated content and the source content [205]. *Entity-based* metrics [222] make a step further by calculating the overlap of entities, as any omission or inaccurate generation of these key entities could lead to an unfaithful response. Notably, even if entities match, the relations between them might be erroneous. Thus, *relation-based* metrics [98] focus on the overlap of relation tuples and introduce a metric that computes the overlap of relation tuples extracted using trained end-to-end fact extraction models.

Classifier-based Metrics. Beyond computing fact overlap, another straightforward approach to assessing the faithfulness of the model generation involves utilizing classifiers trained on data from related tasks such as **natural language inference (NLI)** and fact-checking, or data comprised of synthetically task-specific hallucinated and faithful content. A foundational principle for assessing the faithfulness of generated text is anchored on the idea that genuinely faithful content should inherently be entailed by its source content. In line with this, numerous studies [81, 205] have trained classifiers on NLI datasets to identify factual inaccuracies, especially in the context of abstract summarization. However, Mishra et al. [214] highlighted that the mismatch in input granularity between conventional NLI datasets and inconsistency detection datasets limits their applicability for effectively detecting inconsistencies. Building on this, more advanced studies have proposed methods such as fine-tuning on adversarial datasets [17], decomposing the entailment decisions at the dependency arc level [100], and segmenting documents into sentence units then aggregating scores between sentence pairs [151]. While using data from related tasks to fine-tune the classifier has shown promise in evaluating faithfulness, it's essential to recognize the inherent gap between related tasks and the downstream task. The scarcity of annotated data further constrains their applicability. In response to this challenge, a surge of research explores leveraging data-augmentation methods to construct synthetical data for fine-tuning the classifier, either by rule-based perturbation [78, 149, 262] or generation [385].

QA-based Metrics. In contrast to classifier-based metrics, QA-based metrics [76, 118, 267, 306] have recently garnered attention for their enhanced ability to capture information overlap between the model's generation and its source. These metrics operate by initially selecting target answers from the information units within the LLM's output, and then questions are generated by the question-generation module. The questions are subsequently used to generate source answers based on the user context. Finally, the faithfulness of the LLM's responses is calculated by comparing the matching scores between the source and target answers. Although these methodologies share a common thematic approach, they exhibit variability in aspects like answer selection, question generation, and answer overlap, leading to diverse performance outcomes. Building on this foundational work, Fabbri et al. [79] conducted an in-depth evaluation of the components within QA-based metrics, yielding further enhancements in faithfulness evaluation.

Uncertainty-based Metrics. Drawing parallels with the uncertainty-based approaches employed for detecting factuality hallucinations (Section 4.1.1), the application of uncertainty estimation in assessing faithfulness has been widely explored, typically characterized by entropy and log-probability. For entropy-based uncertainty, Xiao and Wang [329] has revealed a positive correlation between hallucination likelihood in data-to-text generation and predictive uncertainty, which is estimated by deep ensembles [153]. In a related vein, Guerreiro et al. [105] leveraged the variance in hypotheses yielded by Monte Carlo Dropout [91] as an uncertainty measure within neural machine translation. More recently, van der Poel et al. [301] employed conditional entropy [333] to assess model uncertainty in abstractive summarization. Regarding log-probability, it can be applied at different levels of granularity, such as word or sentence level. Notably, several studies [90, 105, 355] have adopted length-normalized sequence log-probability to measure model

confidence. Furthermore, considering the hallucinated token can be assigned high probability when the preceding context contains the same hallucinated information, Zhang et al. [370] focused on the most informative and important keywords and introduced a penalty mechanism to counteract the propagation of hallucinated content.

LLM-based Judgement. Recently, the remarkable instruction-following ability of LLMs has underscored their potential for automatic evaluation [51, 187, 310]. Exploiting this capability, researchers have ventured into novel paradigms for assessing the faithfulness of model-generated content [2, 94, 131, 150, 196]. By providing LLMs with concrete evaluation guidelines and feeding them both the model-generated and source content, they can effectively assess faithfulness. The final evaluation output can either be a binary judgment on faithfulness [196] or a k-point Likert scale indicating the degree of faithfulness [94]. For prompt selection, evaluation prompt can either be direct prompting, chain-of-thought prompting [2], using in-context-learning [131] or allowing the model to generate evaluation results accompanying with explanations [150].

4.2 Hallucination Benchmarks

In this section, we present a comprehensive overview of existing hallucination benchmarks (Table 4), which can be categorized into two primary domains: Hallucination Evaluation Benchmarks (Section 4.2.1), which assess the extent of hallucinations generated by existing cutting-edge LLMs, and Hallucination Detection Benchmarks (Section 4.2.2), designed specifically to evaluate the performance of existing hallucination detection methods. Collectively, these benchmarks establish a unified framework, enabling a nuanced and thorough exploration of hallucinatory patterns in LLMs.

4.2.1 Hallucination Evaluation Benchmarks. Hallucination evaluation benchmarks are devised to quantify the tendency of LLMs to generate hallucinations, particularly emphasizing factual inaccuracies and inconsistency from the given contexts. Given the adeptness of LLMs at memorizing high-frequency count knowledge, the primary focus of current hallucination evaluation benchmarks targets long-tailed knowledge and challenging questions that can easily elicit imitative falsehood. As for evaluating, these benchmarks typically utilize multiple choice QA, where performance is measured through accuracy metrics, or generative QA, evaluated either through human judgment or scores given by proxy models.

Long-tail Factual Knowledge. The selection criteria for gathering long-tail factual question-answering samples typically include the frequency of appearance, recency, and specific domains. Regarding the frequency of appearance, benchmarks such as PopQA [201] and Head-to-Tail [286] are constructed based on entity popularity derived directly from Wikipedia. Considering that world knowledge is constantly evolving, it becomes crucial to validate the LLM's factuality concerning the current world. Among benchmarks characterized by ever-changing, REALTIMEQA [145] and FreshQA [304] stand out. REALTIMEQA offers real-time, open-domain multiple-choice questions that are regularly updated to reflect the latest developments. These questions are derived from newly published news articles, encompassing a broad spectrum of topics, including politics, business, sports, and entertainment. Similarly, FreshQA challenges LLMs with questions designed to represent varying degrees of temporal change—categorized into never-changing, slow-changing, and fast-changing world knowledge. This benchmark is further enriched by including questions based on false premises, requiring debunking, thus comprising a total of 600 meticulously hand-crafted questions. Moreover, long-tail knowledge often pertains to specific domains. For instance, Med-HALT [299] is distinguished by its focus on the medical domain, challenging LLMs with multiple-choice questions derived from a variety of countries. Additionally, Malaviya et al. [200] collected expert-curated questions across 32 fields of study, resulting in a high-quality long-form QA dataset with 2,177 questions.

Table 4. An Overview of Existing Hallucination Benchmarks

Benchmark	Datasets	Data Size	Language	Attribute			Task			
				Factuality	Faithfulness	Manual	Task Type	Input	Label	Metric
TruthfulQA [179]	-	817	English	✔	✗	✔	Generative QA Multi-Choice QA	Question	Answer	LLM-Judge & Human
REALTIMEQA [145]	-	Dynamic	English	✔	✗	✔	Multi-Choice QA Generative QA	Question	Answer	Acc EM & F1
SelfCheckGPT-Wikibio [210]	-	1,908	English	✗	✔	✗	Detection	Paragraph & Concept	Passage	AUROC
HaluEval [166]	Task-specific	30,000	English	✗	✔	✗	Detection	Query	Response	Acc
	General	5,000	English	✗	✔	✗	Detection	Task Input	Response	Acc
Med-HALT [299]	-	4,916	Multilingual	✔	✗	✗	Multi-Choice QA	Question	Choice	Pointwise Score & Acc
FACTOR [220]	Wiki-FACTOR	2,994	English	✔	✗	✗	Multi-Choice QA	Question	Answer	likelihood
	News-FACTOR	1,036	English	✔	✗	✗	Multi-Choice QA	Question	Answer	likelihood
BAMBOO [75]	SenHallu	200	English	✗	✔	✗	Detection	Paper	Summary	P & R & F1
	AbsHallu	200	English	✗	✔	✗	Detection	Paper	Summary	P & R & F1
ChineseFactEval [307]	-	125	Chinese	✔	✗	✔	Generative QA	Question	-	Score
HaluQA [49]	Misleading	175	Chinese	✔	✗	✔	Generative QA	Question	Answer	LLM-Judge
	Misleading-hard	69	Chinese	✔	✗	✔	Generative QA	Question	Answer	LLM-Judge
	Knowledge	206	Chinese	✔	✗	✔	Generative QA	Question	Answer	LLM-Judge
FreshQA [304]	Never-changing	150	English	✔	✗	✔	Generative QA	Question	Answer	Human
	Slow-changing	150	English	✔	✗	✔	Generative QA	Question	Answer	Human
	Fast-changing	150	English	✔	✗	✔	Generative QA	Question	Answer	Human
	False-premise	150	English	✔	✗	✔	Generative QA	Question	Answer	Human
FELM [42]	-	3,948	English	✔	✔	✗	Detection	Question	Response	Balanced Acc & F1
PHD [336]	PHD-LOW	100	English	✗	✔	✗	Detection	Entity	Response	P & R & F1
	PHD-Meidum	100	English	✗	✔	✗	Detection	Entity	Response	P & R & F1
	PHD-High	100	English	✗	✔	✗	Detection	Entity	Response	P & R & F1
ScreenEval [155]	-	52	English	✗	✔	✗	Detection	Document	Summary	AUROC
RealHall [89]	COVID-QA	N/A	English	✗	✔	✗	Detection	Question	Answer	AUROC
	DROP	N/A	English	✗	✔	✗	Detection	Question	Answer	AUROC
	Open Assistant	N/A	English	✗	✔	✗	Detection	Question	Answer	AUROC
	TriviaQA	N/A	English	✗	✔	✗	Detection	Question	Answer	AUROC
LSum [84]	-	6,166	English	✗	✔	✗	Detection	Document	Summary	Balanced Acc
SAC^3 [360]	HotpotQA	250	English	✗	✔	✗	Detection	Question	Answer	AUROC
	NQ-Open	250	English	✗	✔	✗	Detection	Question	Answer	AUROC
HaluEval 2.0 [165]	Biomedicine	1,535	English	✔	✗	✗	Generative QA	Question	Answer	MiHR & MaHR
	Finance	1,125	English	✔	✗	✗	Generative QA	Question	Answer	MiHR & MaHR
	Science	1,409	English	✔	✗	✗	Generative QA	Question	Answer	MiHR & MaHR
	Education	1,701	English	✔	✗	✗	Generative QA	Question	Answer	MiHR & MaHR
	Open domain	3,000	English	✔	✗	✗	Generative QA	Question	Answer	MiHR & MaHR

For Attribute, *Factuality* and *Faithfulness* represent whether the benchmark is used to evaluate LLM's factuality or to detect faithfulness hallucination, and *Manual* represents whether the inputs in the data are handwritten.

Imitative Falsehood Knowledge. Imitative falsehood knowledge is specifically designed to challenge LLMs through adversarial prompting. This approach crafts questions in such a way that they are prone to misleading LLMs due to false beliefs or misconceptions. The two most representative benchmarks are TruthfulQA [179] and HalluQA [49]. TruthfulQA comprises 817 questions that span 38 diverse categories, such as health, law, finance, and politics. Crafted using an adversarial methodology, it aims to elicit "imitative falsehoods"—misleading responses that models might generate due to their frequent presence in training data. The benchmark is divided into two parts, one of which contains manually curated questions that were further refined by filtering out those correctly answered by GPT-3, resulting in 437 filtered questions. The other part includes 380 unfiltered non-adversarial questions. Drawing from the construction approach of TruthfulQA, HalluQA is crafted to specifically assess hallucinations in Chinese LLMs, focusing on imitative

falsehoods and factual errors. The benchmark comprises 450 handcrafted adversarial questions across 30 domains and is categorized into two parts. The misleading section captures questions that successfully deceive GLM-130B, while the knowledge section retains questions that both ChatGPT and Puyu consistently answer incorrectly. To comprehensively evaluate LLM hallucinations across various domains, Li et al. [165] constructed an upgraded hallucination evaluation benchmark, HaluEval 2.0, based on [166]. This benchmark includes 8,770 questions that LLMs are prone to hallucination across five domains: biomedicine, finance, science, education, and open domain.

4.2.2 Hallucination Detection Benchmarks. For hallucination detection benchmarks, most prior studies have primarily concentrated on task-specific hallucinations, such as abstractive summarization [80, 101, 149, 205, 233, 306], data-to-text [237, 292], and machine translation [385]. However, the content generated in these studies often originates from models with lesser capabilities, such as BART [161] and PEGASUS [362]. As a result, they may not accurately reflect the effectiveness of hallucination detection strategies, underlining the necessity for a significant shift toward developing benchmarks that encapsulate more complex scenarios reflective of the era of LLMs.

For example, SelfCheckGPT-Wikibio [210] offers a sentence-level dataset created by generating synthetic Wikipedia articles with GPT-3, manually annotated for factuality, highlighting the challenge of detecting hallucinations in the biography domain. Complementing this, HaluEval [166] combines automated generation with human annotation to evaluate LLMs' ability to recognize hallucinations across 5,000 general user queries and 30,000 task-specific samples, leveraging a "sampling-then-filtering" approach. Building upon existing research predominantly focused on short documents, BAMBOO [75] and ScreenEval [155] extend the scope in long-form hallucination detection. Further, FELM [42], distinguishes itself by assessing factuality across diverse domains including world knowledge, science, and mathematics, producing 817 samples annotated for various facets of factual accuracy, thereby addressing the need for cross-domain evaluation of factuality in LLM-generated content. On a different note, PHD [336], shifts the focus towards passage-level detection of non-factual content by analyzing entities from Wikipedia, thus offering a nuanced view on the knowledge depth of LLMs. RealHall [89] and SAC^3 [360] align closely with real-world applications focusing on open-domain question-answering, whereas LSum [84] concentrating on summarization tasks.

5 Hallucination Mitigation

In this section, we present a comprehensive review of contemporary methods aimed at mitigating hallucinations in LLMs. Drawing from insights discussed in *Hallucination Causes* (Section 3), we systematically categorize these methods based on the underlying causes of hallucinations. Specifically, we focus on approaches addressing *Data-related Hallucinations* (Section 5.1), *Training-related Hallucinations* (Section 5.2) and *Inference-related Hallucinations* (Section 5.3), each offering tailored solutions to tackle specific challenges inherent to their respective cause.

5.1 Mitigating Data-related Hallucinations

As analyzed in Section 3.1, data-related hallucinations generally emerge as a byproduct of misinformation, biases, and knowledge gaps, which are fundamentally rooted in the pre-training data. Several methods are proposed to mitigate such hallucinations, primarily categorized into three distinct parts: (1) *data filtering* aiming at selecting high-quality data to avoid introducing misinformation and biases, (2) *model editing* focusing on injecting up-to-date knowledge by editing model's parameters, and (3) RAG leveraging external non-parametric database for knowledge supplying.

5.1.1 Data Filtering. To reduce the presence of misinformation and biases, an intuitive approach involves the careful selection of high-quality pre-training data from reliable sources. In this way, we can ensure the factual correctness of data while also minimizing the introduction of social biases. As early as the advent of GPT-2, Radford et al. [249] underscored the significance of exclusively scraping web pages that had undergone rigorous curation and filtration by human experts. However, as pre-training datasets continue to scale, manual curation becomes a challenge. Given that academic or specialized domain data is typically factually accurate, gathering high-quality data emerges as a primary strategy. Notable examples include *the Pile* [92] and "textbook-like" data sources [106, 174]. Additionally, up-sampling factual data during the pre-training phase has been proven effective in enhancing the factual correctness of LLMs [296], thus alleviating hallucination.

In addition to strictly controlling the source of data, deduplication serves as a crucial procedure. Existing practices typically fall into two categories: exact duplicates and near-duplicates. For exact duplicates, the most straightforward method involves exact substring matching to identify identical strings. However, given the vastness of pre-training data, this process can be computationally intensive, a more efficient method utilizes the construction of a suffix array [203], enabling effective computation of numerous substring queries in linear time. Regarding near-duplicates, the identification often involves approximate full-text matching, typically utilizing hash-based techniques to identify document pairs with significant n-gram overlap. Furthermore, MinHash [28] stands out as a prevalent algorithm for large-scale deduplication tasks [109]. Additionally, SemDeDup [1] makes use of embeddings from pre-trained models to identify semantic duplicates, which refers to data pairs with semantic similarities but not identical.

Discussion. Since data filtering works directly at the source of hallucinations, it effectively mitigates hallucinations by ensuring the use of high-quality, factually accurate sources. Despite its effectiveness, the efficiency and scalability of current data filtering methods pose significant challenges as data volumes expand. Additionally, these methods often overlook the influence of LLM-generated content, which can introduce new risks and inaccuracies. To advance, future research must focus on developing more efficient, automated data filtering algorithms that can keep pace with the rapid expansion of datasets and the complexities of LLM-generated content.

5.1.2 Model Editing. Model editing [276, 316, 365] has garnered rising attention from researchers, which aims to rectify model behavior by incorporating additional knowledge. Current model editing techniques can be categorized into two classes: *locate-then-edit* and *meta-learning*.

Locate-then-edit. Locate-then-edit methods [65, 207] consist of two stages, which first locate the "buggy" part of the model parameters and then apply an update to them to alter the model's behavior. For example, ROME [207] located the edits-related layer by destroying and subsequently restoring the activations and then updates the parameters of FFN in a direct manner to edit knowledge. MEMIT [208] employed the same knowledge locating methods as ROME, enabling the concurrent updating of multiple layers to facilitate the simultaneous integration of thousands of editing knowledge. However, Yao et al. [343] found that these methods lack non-trivial generalization capabilities and varying performance and applicability to different model architectures. The best-performing methods ROME and MEMIT empirically only work well on decoder-only LLMs.

Meta-learning. Meta-learning methods [69, 215] train an external hyper-network to predict the weight update of the original model. Nevertheless, meta-learning methods often require additional training and memory cost, where MEND [215] utilized a low-rank decomposition with a specialized design to reduce the size of hyper-networks. Notably, MEND would exhibit a cancellation effect, where parameter shifts corresponding to different keys significantly counteract each other. MALMEN [288] further addressed this issue by framing the parameter shift aggregation as a least squares problem rather than a simple summation, thereby greatly enhancing its capacity for

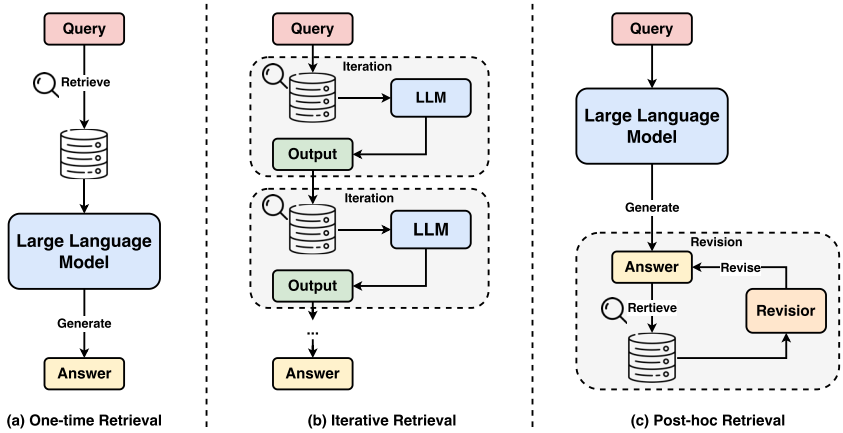

Fig. 4. The illustration of three distinct approaches for RAG: (a) *One-time Retrieval*, where relevant information is retrieved once before text generation; (b) *Iterative Retrieval*, involving multiple retrieval iterations during text generation for dynamic information integration; and (c) *Post hoc Retrieval*, where the retrieval process happens after an answer is generated, aiming to refine and fact-check the generated content.

extensive editing. While these methods can fine-grainedly adjust the behavior of the model, modifications to the parameters could have a potentially harmful impact on the inherent knowledge of the model.

Discussion. Model editing provides a precise way to mitigate hallucinations induced by specific misinformation without extensive retraining. However, these methods struggle with large-scale updates and can adversely affect the model's overall performance, particularly when continuous edits are applied. Consequently, future research should focus on improving model editing to handle large-scale knowledge updates more efficiently and address hallucinations caused by social biases.

5.1.3 RAG. Typically, RAG [108, 162, 274] follows a retrieve-then-read pipeline, where relevant knowledge is firstly retrieved by a *retriever* [144] from external sources, and then the final response is generated by a *generator* conditioning on both user query and retrieved documents. By decoupling external knowledge from LLM, RAG can effectively alleviate the hallucination caused by the knowledge gap without affecting the performance of LLM. Common practices can be divided into three parts, as shown in Figure 4: *one-time retrieval*, *iterative retrieval*, and *post hoc retrieval*, depending on the timing of retrieval.

One-time Retrieval. One-time retrieval aims to directly prepend the external knowledge obtained from a single retrieval to the LLMs' prompt. Ram et al. [252] introduced In-context RALM, which entails a straightforward yet effective strategy of prepending chosen documents to the input text of LLMs. Beyond conventional knowledge repositories such as Wikipedia, ongoing research endeavors have explored alternative avenues, specifically the utilization of **knowledge graphs (KGs)**. These KGs serve as a pivotal tool for prompting LLMs, facilitating their interaction with the most recent knowledge, and eliciting robust reasoning pathways [14, 246, 325]. Varshney et al. [302] introduce the **Parametric Knowledge Guiding (PKG)** framework, enhancing LLMs with domain-specific knowledge. PKG employs a trainable background knowledge module, aligning it with task knowledge and generating relevant contextual information.

Iterative Retrieval. When confronted with intricate challenges like multi-step reasoning [340] and long-form question answering [82, 280], traditional one-time retrieval may fall short. Addressing these demanding information needs, recent studies have proposed iterative retrieval, which allows

for continuously gathering knowledge throughout the generation process. Recognizing the substantial advancements chain-of-thought prompting [322] has brought to LLMs in multi-step reasoning, numerous studies [112, 297, 342] try to incorporate external knowledge at each reasoning step and further guide retrieval process based on ongoing reasoning, reducing factual errors in reasoning chains. Building upon chain-of-thought prompting, Press et al. [244] introduced *self-ask*. Diverging from the conventional continuous, undelineated chain-of-thought prompting, *self-ask* delineates the question it intends to address at each step, subsequently incorporating a search action based on the follow-up question. Instead of solely depending on chain-of-thought prompting for retrieval guidance, both Feng et al. [86] and Shao et al. [269] employed an iterative retrieval-generation collaborative framework, where a model's response serves as an insightful context to procure more relevant knowledge, subsequently refining the response in the succeeding iteration. Beyond multi-step reasoning tasks, Jiang et al. [138] shifted their emphasis to long-form generation. They proposed an active retrieval augmented generation framework, which iteratively treats the upcoming prediction as a query to retrieve relevant documents. If the prediction contains tokens of low confidence, the sentence undergoes regeneration. In addition to using iterative retrieval to improve intermediate generations, Zhang et al. [367] presented MixAlign, which iteratively refines user questions using model-based guidance and seeking clarifications from users, ultimately enhancing the alignment between questions and knowledge.

Post hoc Retrieval. Beyond the traditional *retrieve-then-read* paradigm, a line of work has delved into post hoc retrieval, refining LLM outputs through subsequent retrieval-based revisions. To enhance the trustworthiness and attribution of LLMs, Gao et al. [93] adopted the *research-then-revise* workflow, which initially research relevant evidence and subsequently revise the initial generation based on detected discrepancies with the evidence. Similarly, Zhao et al. [377] introduced the *verify-and-edit* framework to enhance the factual accuracy of reasoning chains by incorporating external knowledge. For reasoning chains that show lower-than-average consistency, the framework generates verifying questions and then refines the rationales based on retrieved knowledge, ensuring a more factual response. Yu et al. [354] enhanced the post hoc retrieval method through diverse answer generation. Instead of generating just a single answer, they sample various potential answers, allowing for a more comprehensive retrieval feedback. Additionally, by employing an ensembling technique that considers the likelihood of the answer before and after retrieval, they further mitigate the risk of misleading retrieval feedback.

Discussion. One crucial advantage of RAG methodology is its effectiveness in mitigating hallucinations caused by knowledge gaps, and their generality, which allows for application across any domain. This flexibility is further enhanced by the modularity of the approach, treating external knowledge bases like plug-ins that can be swapped or modified as needed. In terms of the drawbacks, it can be easily impacted by irrelevant retrievals, which may decrease the overall performance by introducing noise or incorrect information into the response generation process. Furthermore, the current paradigm exhibits shallow interactions between the retriever and generator components, leading to suboptimal knowledge utilization. Hence, future research should focus on developing a robust RAG system that minimizes the impact of irrelevant retrieval, as well as integrating adaptive learning components that can dynamically adjust retrieval strategies based on the context of the query and the performance of previous interactions.

5.2 Mitigating Training-related Hallucination

Training-related hallucinations typically arise from the intrinsic limitations of the architecture and training strategies adopted by LLMs. In this context, we discuss various optimization methods ranging from training stages (Section 5.2.1) and alignment stages (SFT and RLHF) (Section 5.2.2), aiming to mitigate hallucinations within the training process.

5.2.1 Mitigating Pretraining-related Hallucination. One significant avenue of research in mitigating pretraining-related hallucination centers on the limitations inherent in model architectures, especially *unidirectional representation* and *attention glitches*. In light of this, numerous studies have delved into designing novel model architectures specifically tailored to address these flaws. To address the limitations inherent in unidirectional representation, Li et al. [177] introduced BATGPT which employs a bidirectional autoregressive approach. This design allows the model to predict the next token based on all previously seen tokens, considering both past and future contexts, thus capturing dependencies in both directions. Building on this idea, Liu et al. [186] highlighted the potential of encoder-decoder models to make better use of their context windows, suggesting a promising direction for future LLMs architecture design. Besides, recognizing the limitations of soft attention within self-attention-based architecture, Liu et al. [180] proposed attention-sharpening regularizers. This plug-and-play approach specifies self-attention architectures using differentiable loss terms [361] to promote sparsity, leading to a significant reduction in reasoning hallucinations.

In the pre-training phase of LLMs, the choice of objective plays a pivotal role in determining the model's performance. However, conventional objectives can lead to fragmented representations and inconsistencies in model outputs. Recent advancements have sought to address these challenges by refining pre-training strategies, ensuring richer context comprehension, and circumventing biases. Addressing the inherent limitations in training LLMs, where unstructured factual knowledge at a document level often gets chunked due to GPU memory constraints and computational efficiency, leading to fragmented information and incorrect entity associations, Lee et al. [157] introduced a factuality-enhanced training method. By appending a TOPICPREFIX to each sentence in factual documents, the approach transforms them into standalone facts, significantly reducing factual errors and enhancing the model's comprehension of factual associations. Similarly, considering that randomly concatenating shorter documents during pre-training might introduce inconsistencies in model outputs, Shi et al. [272] proposed In-Context Pretraining, an innovative approach in which LLMs are trained on sequences of related documents. By altering the document order, this method aims to maximize similarity within the context windows. It explicitly encourages LLMs to reason across document boundaries, potentially bolstering the logical consistency between generations.

Discussion. Strategies designed to mitigate pretraining-related hallucinations typically are fundamental, potentially yielding significant improvements. However, they typically involve modifications to pre-training architectures and objectives, which are computationally intensive. Moreover, these integrations may lack broad applicability. Moving forward, the focus should be on developing adaptable and efficient strategies that can be universally applied without extensive system overhaul.

5.2.2 Mitigating Misalignment Hallucination. Hallucinations induced during alignment often stem from capability misalignment and belief misalignment. However, defining the knowledge boundary of LLMs proves challenging, making it difficult to bridge the gap between LLMs' inherent capabilities and the knowledge presented in human-annotated data. While limited research addresses capability misalignment, the focus mainly shifts toward belief misalignment.

Hallucinations stemming from belief misalignment often manifest as sycophancy, a tendency of LLMs to seek human approval in undesirable ways. This sycophantic behavior can be attributed to the fact that human preference judgments often favor sycophantic responses over more truthful ones [270], paving the way for reward hacking [264]. To address this, a straightforward strategy is to improve human preference judgments and, by extension, the preference model. Recent research [25, 264] has investigated the use of LLMs to assist human labelers in identifying overlooked flaws. Additionally, Sharma et al. [270] discovered that aggregating multiple human preferences enhances feedback quality, thereby reducing sycophancy.

Besides, modifications to LLMs' internal activations have also shown the potential to alter model behavior. This can be achieved through methods like fine-tuning [323] or activation steering during inference [68, 116, 285]. Specifically, Wei et al. [323] proposed a synthetic-data intervention, fine-tuning language models using synthetic data where the claim's ground truth is independent of a user's opinion, aiming to reduce sycophantic tendencies.

Another avenue of research [259, 260] has been to mitigate sycophancy through activation steering. This approach involves using pairs of sycophantic/non-sycophantic prompts to generate the sycophancy steering vector derived from averaging the differences in intermediate activations. During inference, subtracting this vector can produce less sycophantic LLM outputs.

Discussion. Mitigating hallucinations through post-training methods represents a direct and effective approach, bypassing the complexities associated with data sourcing and pre-training. However, a notable gap in current research is the limited attention given to capability misalignment within LLMs. Future research should prioritize understanding the knowledge boundaries in capability alignment to address hallucinations effectively.

5.3 Mitigating Inference-related Hallucination

Decoding strategies in LLMs play a pivotal role in determining the factuality and faithfulness of the generated content. However, as analyzed in Section 3.3, imperfect decoding often results in outputs that might lack factuality or stray from the original context. In this subsection, we explore two advanced strategies aimed at refining the decoding strategy to enhance both the factuality and faithfulness of the LLMs' outputs.

5.3.1 Factuality Enhanced Decoding. Factuality Enhanced Decoding aims to improve the reliability of outputs from LLMs by prioritizing the factuality of the information they generate. This line of methods focuses on aligning model outputs closely with established real-world facts, thereby minimizing the risk of disseminating false or misleading information.

Factuality Decoding. Considering the randomness in the sampling process can introduce non-factual content into open-ended text generation, Lee et al. [157] introduced the factual-nucleus sampling algorithm that dynamically adjusts the nucleus probability p throughout sentence generation. By dynamically adjusting the nucleus probability based on decay factors and lower boundaries and resetting the nucleus probability at the beginning of every new sentence, the decoding strategy strikes a balance between generating factual content and preserving output diversity. Moreover, some studies [31, 217] posit that the activation space of LLMs contains interpretable structures related to factuality. Building on this idea, Li et al. [169] introduced Inference-Time Intervention. This method first identifies a direction in the activation space associated with factually correct statements and then adjusts activations along the truth-correlated direction during inference. By repeatedly applying such intervention, LLMs can be steered towards producing more factual responses. Similarly, Chuang et al. [58] delved into enhancing the factuality of LLM's decoding process from a perspective of factual knowledge storage. They exploit the hierarchical encoding of factual knowledge within transformer LLMs, noting that lower-level information is captured in earlier layers and semantic information in the later ones. Drawing inspiration from [172], they introduce DoLa, a strategy that dynamically selects and contrasts logits from different layers to refine decoding factuality. By placing emphasis on knowledge from higher layers and downplaying that from the lower layers, DoLa showcases its potential to make LLMs more factual, thus reducing hallucinations.

Post-editing Decoding. Unlike methods that directly modify the probability distribution to prevent hallucinations during the initial decoding, post-editing decoding seeks to harness the self-correction capabilities of LLMs [234] to refine the originally generated content without relying on an external

knowledge base. Dhuliawala et al. [73] introduced the CoVE), which operates under the assumption that, when appropriately prompted, LLMs can self-correct their mistakes and provide more accurate facts. Starting with an initial draft, it first formulates verification questions and then systematically answers those questions in order to finally produce an improved revised response. Similarly, Ji et al. [135] focused on the medical domain and introduced an iterative self-reflection process. This process leverages the inherent ability of LLMs to first generate factual knowledge and then refine the response until it aligns consistently with the provided background knowledge.

Discussion. Factuality decoding methods, which typically assess the factuality at each decoding step, can offer substantial improvements. Furthermore, due to their plug-and-play nature, they allow for application without the need for computation-intensive training. Nevertheless, one of the primary limitations of these methods lies in balancing factual accuracy with maintaining the diversity and informativeness of the generated content, which can sometimes lead to compromises in either aspect. On the other hand, post-editing decoding strategies, despite their effectiveness, heavily rely on the self-correction capabilities of LLMs, which may be unreliable. Furthermore, applying self-reflection can be time-consuming, limiting their practicality for real-time applications. Hence, it is crucial to achieve an optimal balance between factuality and computational efficiency.

5.3.2 Faithfulness Enhanced Decoding. On the other hand, Faithfulness Enhanced Decoding prioritizes alignment with the provided context and also emphasizes enhancing the consistency within the generated content. Thus, in this section, we summarize existing work into two categories, including *Context Consistency* and *Logical Consistency*.

Context Consistency. In the era of LLMs, the issue of faithfulness hallucination typically lies in insufficient attention to the given context, which inspired numerous research to design inference-time strategies to enhance context consistency. Shi et al. [271] proposed **context-aware decoding (CAD)**, which modifies the model's original output distribution in a contrastive formulation [172]. By amplifying the difference between output probabilities with and without context, CAD encourages the LLM to focus more on contextual information rather than over-rely on prior knowledge. However, due to the inherent tradeoff between diversity and context attribution [102, 359], overemphasizing contextual information can reduce diversity. To address this, Chang et al. [36] introduced a dynamic decoding algorithm to bolster faithfulness while preserving diversity. Specifically, the algorithm involves two parallel decoding steps, one with the context and one without. During the decoding, the KL divergence between two token distributions serves as a guiding signal, indicating the relevance of the source context. This signal is utilized to dynamically adjust the sampling temperature to improve source attribution when the source is relevant. In a parallel line of work, Choi et al. [53] introduced knowledge-constrained decoding, which employed a token-level hallucination detection discriminator to identify contextual hallucinations and then guides the faithful generation process by reweighing the token distribution. In addition to modifying output distribution in place to enhance contextual attention, another line of work has explored a generic post-edit approach to enhance faithfulness. Gao et al. [93] adopted a *research-and-revise* workflow, where the research stage raises questions about various aspects of the model's initial response and gathers evidence for each query, while the revision stage detects and revises any disagreements between the model's response and the evidence. Similarly, Lei et al. [158] first detected contextual hallucinations at both the sentence and entity levels and then incorporated the judgments to refine the generated response. Moreover, several studies have explored methods to overcome the softmax bottleneck, which constrains the expression of diversity and faithful representations. These approaches include employing a mixture of Softmax, which uses multiple hidden states to compute softmax multiple times and merge the resulting distributions [339] and incorporating pointer networks, which enables LLMs to copy the context words [37], thereby reducing context hallucinations.

Logical Consistency. Inspired by the human thinking process, chain-of-thought [322] has been introduced to encourage LLMs to decompose complex problems into explicitly intermediate steps, thereby enhancing the reliability of the reasoning process. Despite effective, recent research [154, 298] demonstrated that the intermediate rationales generated by LLMs do not faithfully capture their underlying behavior. A branch of research has been inspired to improve the consistency of intermediate rationales generated by LLMs, particularly in multi-step reasoning [60] and logical reasoning [18]. To enhance the self-consistency in chain-of-thought, Wang et al. [314] employed a knowledge distillation framework. They first generate a consistent rationale using contrastive decoding [172] and then fine-tune the student model with a counterfactual reasoning objective, which effectively eliminates reasoning shortcuts [27] that derive answers without considering the rationale. Furthermore, by employing contrastive decoding directly, LLMs can reduce surface-level copying and prevent missed reasoning steps [226]. In addition, Li et al. [164] conducted a deep analysis of the causal relevance among the context, CoT, and answer during unfaithful reasoning. Analysis revealed that the unfaithfulness issue lies in the inconsistencies in the context information obtained by the CoT and the answer. To address this, they proposed inferential bridging, which takes the attribution method to recall contextual information as hints to enhance CoT reasoning and filter out noisy CoTs that have low semantic consistency and attribution scores to the context. Paul et al. [238] decomposed the reasoning process into two modules: an inference module, which employs Direct Preference Optimization [250] to align the LLM towards preferring correct reasoning chains over counterfactual chains, and a reasoning module, which encourages the LLM to reason faithfully over the reasoning steps using a counterfactual and causal preference objective. Compared to natural language reasoning, logical reasoning demands rigorous logical calculation, whereas plain text often lacks precise logical structure, leading to unfaithful reasoning. To address this, Xu et al. [334] introduced **Symbolic CoT (SymbCoT)**, which incorporates symbolic expressions within CoT to describe intermediate reasoning steps. Specifically, SymbCoT translates the natural language context into a symbolic representation and then formulates a step-by-step plan to address the logical reasoning problem, followed by a verifier to check the translation and reasoning chain, thereby ensuring faithful logical reasoning.

Discussion. Faithfulness Enhanced Decoding significantly advances the alignment of LLM outputs with provided contexts and enhances the internal consistency of the generated content. However, strategies such as CAD often lack adaptive mechanisms, limiting their effectiveness in scenarios that demand dynamic attention to context. Furthermore, many decoding strategies require the integration of additional models that do not focus on context, introducing significant computational overhead and reducing efficiency.

6 Hallucinations in Retrieval Augmented Generation

RAG has emerged as a promising strategy to mitigate hallucinations and improve the factuality of LLM outputs [129, 162, 252, 273]. By incorporating large-scale external knowledge bases during inference, RAG equips LLMs with up-to-date knowledge, thus reducing the potential risk of hallucination due to the inherent knowledge boundaries of LLMs [257]. Despite being designed to mitigate LLM hallucinations, retrieval-augmented LLMs can still produce hallucinations [16]. Hallucinations in RAG present considerable complexities, manifesting as outputs that are either factually inaccurate or misleading. These hallucinations occur when the content generated by the LLM does not align with real-world facts, fails to accurately reflect the user's query, or is not supported by the retrieved information. Such hallucinations can stem from two primary factors: *retrieval failure* (Section 6.1) and *generation bottleneck* (Section 6.2). Through a comprehensive analysis of the limitations present in current RAG systems, we aim to shed light on potential improvements for retrieval-augmented LLMs, paving the way for more reliable IR systems.

Table 5. Examples of Each Category of User Queries That May Lead to LLM Hallucinations

Type	Example Dialog	Explanation
Retrieval Intent	*Retrieved context*: At high altitudes, water boils at lower temperatures than at sea level, such as 95℃ at elevations above 2,000 meters. This altitude-induced reduction can affect cooking times and methods. *Query*: What is the boiling point of water at sea level? *Response*: The boiling point of water at sea level is 95℃.	*This query involves a basic scientific fact that should be available within the model's parametric knowledge. However, the system inappropriately retrieves* and uses altitude-specific information about boiling temperatures, leading to an incorrect response for sea-level conditions.
Ambiguous Queries	*Retrieved context*: The Hennessey Venom GT famously set a speed record by completing a mile on a racetrack in 32 seconds, highlighting its performance in motorsport events. *Query*: What's the record for the fastest mile run on track? *Response*: The fastest mile on a track was completed in 32 seconds by the Hennessey Venom GT.	*The query is ambiguous because it does not specify whether it refers to a human runner or a vehicle.* The retrieval system erroneously pulls information related to car racing records instead of athletics.
Complex Queries	*Retrieved context*: Alexander Fleming, known for his groundbreaking work in discovering penicillin, significantly influenced modern antibiotics. Fleming's discovery saved countless lives. *Query*: Which university did the scientist who discovered penicillin graduate from? *Response*: Alexander Fleming graduated from Oxford University.	*The retrieval system only manages to fetch information about Fleming's professional achievements in the discovery of penicillin.* However, the document does not provide information about his educational background, thus the model generates a hallucinatory answer.

Content marked in Red represents the hallucinatory output.

6.1 Retrieval Failure

The retrieval process is a crucial initial step in the RAG framework, tasked with retrieving the most relevant information for information-seeking queries. Consequently, failures in the retrieval stage can have serious downstream effects on the RAG pipeline, leading to hallucinations. These failures typically stem from three primary parts: the formulation of user queries, the reliability and scope of retrieval sources, and the effectiveness of the retriever.

6.1.1 User Queries. User queries play a fundamental role in guiding the retrieval process with RAG systems. The specificity and clarity of these queries critically influence the effectiveness of retrieval outcomes. In this section, we discuss factors that may contribute to hallucinations from three perspectives: blind retrieval, misinterpretation of ambiguous queries, and the challenges in accurate retrieval of complex queries. Some examples are presented in Table 5 for a better understanding.

Retrieval Intent Decisions. Not all queries necessitate retrieval. Blind retrieval for queries that do not require external knowledge can counterproductively lead to misleading responses. As shown in Table 5, the query about "*the boiling point of water at sea level*" pertains to a basic scientific fact that the model could address without external retrieval. However, the retrieval system was inappropriately activated, blindly retrieving inaccurate information and consequently leading to an undesirable response. Consequently, several studies [74, 201, 225, 374] have proposed to make a shift from passive retrieval to adaptive retrieval. In general, these strategies can be divided into two categories: *heuristic-based* and *self-aware judgment. Heuristic-based* methods employ heuristic rules to determine the necessity of retrieval. For instance, Mallen et al. [201] observed a positive correlation between LLMs' memorization capabilities and entity popularity and suggested triggering retrieval only when the entity popularity in the user query falls below a certain threshold. Similarly, Jeong et al. [133] determined the timing of retrieval based on

the query complexity, whereas Asai et al. [11] considered whether the query is factual relevant. *Self-aware judgment* leverages the models' intrinsic judgment to decide the necessity for IR. Feng et al. [85], Ren et al. [257] and Wang et al. [317] directly prompted LLMs for retrieval decisions, recognizing that LLMs possess a certain level of awareness regarding their knowledge boundaries [141, 346]. Moreover, Jiang et al. [138] introduced an active retrieval strategy that triggers retrieval only when the LLM generates low-probability tokens. Similarly, Su et al. [284] not only considered the uncertainty of each token but also its semantic contribution and impact on the subsequent context. More recently, Cheng et al. [48] proposed four orthogonal criteria for determining the retrieval timing, which include intent-aware, knowledge-aware, time-sensitive-aware, and self-aware.

Ambiguous Queries. Ambiguous user queries, containing omission, coreference, and ambiguity, significantly complicate the retrieval system's ability to fetch precisely relevant information, thereby increasing the likelihood of generating undesirable responses. As shown in Table 5, due to the ambiguity of the query about "*the record for the fastest mile run on track*," the retrieval system erroneously retrieved information from automobile racing events, which led the model to generate a response suited for vehicles instead of athletes. A prevalent mitigation strategy is query rewriting, where queries are refined and decontextualized to better match relevant documents. Wang et al. [313] and Jagerman et al. [130] have explored prompting approaches where the LLM is prompted to generate a pseudo-document or rationale based on the original query, which is then used for further retrieval. Additionally, Ma et al. [197] introduced a trainable rewriter which is trained using the feedback from the LLM via reinforcement learning. Mao et al. [204] employed the feedback signals from the reranker to train the rewrite model, thus eliminating the reliance on annotated data. However, the challenges deepen in conversational search, which encounters a more complex issue of context-dependent query understanding with the lengthy conversational history. Addressing this, Yoon et al. [347] proposed a similar framework for optimizing the LLM to generate retriever-preferred query rewrites. This operated by generating a variety of queries and then using the preference of the rank of retrieved passage to optimize the query rewriting model.

Complex Queries. Complex user queries, characterized by requiring intensive reasoning [282] or encompassing multiple aspects [268, 315], pose significant challenges to the retrieval system. Such queries require advanced understanding and decomposition capabilities, which may exceed the current capabilities of the current retrieval methods based on keyword or semantic matching, often leading to partial or incorrect retrievals. For example, as shown in Table 5, due to the multi-step nature of the query about "*Which university did the scientist who discovered penicillin graduate from?*," direct retrieval often leads to incomplete results, thereby resulting in hallucinatory responses. A common approach involves query decomposition, where the complex query is decomposed into sub-queries to facilitate more accurate IR. For instance, Wang et al. [315] implemented a sub-aspect explorer that utilizes the extensive world knowledge embedded LLMs to identify potential sub-aspects of user queries, thereby providing explicit insights into the user's underlying intents. Similarly, Shao et al. [268] concentrated on the demanding task of expository writing, aiming at retrieving comprehensive information to compose Wikipedia-like articles from scratch on a specific topic. This approach involves decomposing the topic into various perspectives and simulating multi-turn conversations with LLMs, each personified with different perspectives for question asking. Additionally, Cao et al. [32] and Chu et al. [56] explored knowledge-intensive complex reasoning and employed a divide-and-conquer strategy. This strategy begins with decomposing complex questions into question trees, where at each node, the LLM retrieves and aggregates answers from diverse knowledge sources.

6.1.2 Retrieval Sources. The reliability and scope of retrieval sources are crucial determinants of the efficacy of RAG systems. Effective retrieval depends not only on the clarity of the user queries but also on the quality and comprehensiveness of the sources from which information is retrieved. When these sources contain factually incorrect or outdated information, the risk of retrieval failures increases significantly, potentially leading to the generation of incorrect or misleading information.

As the landscape of content creation evolves with the rapid advancement of Artificial Intelligence Generated Content [33], an increasing volume of LLM-generated content is permeating the internet, subsequently becoming integrated into retrieval sources [39]. This integration is reshaping the dynamics of IR, as evidenced by recent empirical studies [67, 335] suggesting that modern retrieval models tend to favor LLM-generated content over human-authored content. Recent research [44] has explored the implications of progressively integrating LLM-generated content into RAG systems. The findings indicate that, without appropriate intervention, human-generated content may progressively lose its influence within RAG systems. Additionally, Tan et al. [289] investigated the performance of RAG systems when incorporating LLM-generated into retrieved contexts, revealing a significant bias favoring generated contexts. This bias stems from the high similarity between generated context and questions, as well as the semantic incompleteness of retrieved contexts. More seriously, the propensity of LLMs to produce factually inaccurate hallucinations exacerbates the reliability issues of retrieval sources. As LLM-generated content often contains factual errors, its integration into retrieval sources can mislead retrieval systems, further diminishing the accuracy and reliability of the information retrieved.

To combat these biases, several approaches have been explored. Inspired by common practice in pre-training data processing [23], Asai et al. [12] proposed a scenario that incorporates a quality filter designed to ensure the high quality of the retrieval datastore. Additionally, Pan et al. [235] proposed Credibility-aware Generation, which equips LLMs with the ability to discern and handle information based on its credibility. This approach assigns different credibility levels to information, considering its relevance, temporal context, and the trustworthiness of its source, thus effectively reducing the impact of flawed information in RAG systems.

6.1.3 Retriever. When the user query is explicit and the retrieval source is reliable, the effectiveness of the retrieval process depends crucially on the performance of the retriever. In such scenarios, the retriever's effectiveness is significantly compromised by improper chunking and embedding practices.

Chunking. Given the extensive nature of retrieval sources, which often encompass lengthy documents like web pages, it poses significant challenges for LLMs with limited context length. Thus, chunking emerges as an indispensable step in RAG, which involves segmenting these voluminous documents into smaller, more manageable chunks to provide precise and relevant evidence for LLMs. According to actual needs, the chunking granularity ranges from documents to paragraphs, even sentences. However, inappropriate retrieval granularity can compromise the semantic integrity and affect the relevance of retrieved information [221], thereby affecting the performance of LLMs. Fixed-size chunking, which typically breaks down the documents into chunks of a specified length such as 100-word paragraphs, serves as the most crude and prevalent strategy of chunking, which is widely used in RAG systems [24, 108, 162]. Considering fixed-size chunking falls short in capture structure and dependency of lengthy documents, Sarthi et al. [263] proposed RAPTOR, an indexing and retrieval system. By recursively embedding, clustering, and summarizing chunks of text, RAPTOR constructs a tree to capture both high-level and low-level details. When retrieval, RAPTOR enables LLMs to integrate information from different levels of abstraction, providing a more comprehensive context for user queries. Instead of chunking text with a fixed chunk

size, semantic chunking adaptively identifies breakpoints between sentences through embedding similarity, thereby preserving semantic continuity [142]. Furthermore, Chen et al. [43] pointed out the limitations of the existing retrieval granularity. On the one hand, while a coarser retrieval with a longer context can theoretically provide a more comprehensive context, it often includes extraneous details that could potentially distract LLMs. On the other hand, a fine-grain level can provide more precise and relevant information, it has limitations such as not being self-contained and lacking necessary contextual information. To address these shortcomings, Chen et al. [43] introduced a novel retrieval granularity, proposition, which is defined as atomic expressions within the text, each encapsulating a distinct factoid and presented in a concise self-contained natural language format.

Embedding. Once the retrieval text is chunked, text chunks are subsequently transformed into vector representation via an embedding model. Such a representation scheme is supported by the well-known data structure of *vector database* [140], which systematically organizes data as key-value pairs for efficient text retrieval. In this manner, the relevance score can be computed according to the similarity function between the text representation and query representation. However, a sub-optimal embedding model may compromise performance, which affects the similarity and matching of chunks to user queries, potentially misleading LLMs. Typically, a standard embedding model [95, 128, 147, 378] learns the query and text representations with encoder-based architecture (e.g., BERT [72], RoBERTa [188]) via contrastive learning [300], where the loss is constructed by contrasting a positive pair of query-document against a set of random negative pairs. However, these embeddings showcase their limitations when applied to new domains, such as medical and financial applications [219, 291]. In these cases, recent studies [70, 231, 273, 328] propose to fine-tune the embedding models on domain-specific data to enhance retrieval relevance. For example, REPLUG [273] utilizes language modeling scores of the answers as a proxy signal to train the dense retriever. More recently, Muennighoff et al. [218] have introduced generative representational instruction tuning where a single LLM is trained to handle both generative and embedding tasks, which largely reduces inference latency in RAG by caching representations. Despite these advancements, the field faces challenges, particularly with the fine-tuning of high-performing yet inaccessible embedding models, such as OpenAI's text-embedding-ada-002. Addressing this gap, Zhang et al. [363] introduced a novel approach for fine-tuning a black-box embedding model by augmenting it with a trainable embedding model which significantly enhances the performance of the black-box embeddings.

6.2 Generation Bottleneck

After the retrieval process, the generation stage emerges as a pivotal point, responsible for generating content that faithfully reflects the retrieved information. However, this stage can encounter significant bottlenecks that may lead to hallucinations. We summarize two key capabilities of LLMs that are closely related to these bottlenecks: contextual awareness and contextual alignment. Each plays an important role in ensuring the reliability and credibility of the RAG system.

6.2.1 Contextual Awareness. Contextual awareness involves understanding and effectively utilizing contextual information retrieved. This section discusses the key factors that impact the LLM's ability to maintain contextual awareness, which can be categorized into three main parts: (1) the presence of noisy retrieval in context, (2) context conflicts, and (3) insufficient utilization of context information.

Noisy Context. As emphasized in Section 6.1, the failure in the retrieval process may inevitably introduce irrelevant information, which will propagate into the generation stage. When the

generator is not robust enough to these irrelevant retrievals, it will mislead the generator and even introduce hallucinations [64].

Yoran et al. [348] conducted a comprehensive analysis on the robustness of current retrieval-augmented LLMs, revealing a significant decrease in performance with random retrieval. While using an NLI model to filter out irrelevant passages is effective, this method comes with the tradeoff of inadvertently discarding some relevant passages. A more effective solution is to train LLMs to ignore irrelevant contexts by incorporating irrelevant contexts in training data. Similarly, Yu et al. [353] introduced Chain-of-Note, which enables LLMs to first generate reading notes for retrieved contexts and subsequently formulate the final answer. In this way, LLMs can not only filter irrelevant retrieval to improve noise robustness but also respond with unknown when retrieval is insufficient to answer user queries. In addition to improving LLM robustness by learning to ignore irrelevant content in the context, several studies [137, 173, 319, 332] propose to compress the context to filter out irrelevant information. Specifically, Li [173] and Jiang et al. [137] made use of small language models to compute self-information and perplexity for prompt compression, finding the most informative content. Similarly, Wang et al. [319] proposed to filter out irrelevant content and leave precisely supporting content based on lexical and information-theoretic approaches. Besides, efforts have been also made to employ summarization models as compressors. Xu et al. [332] presented both extractive and abstractive compressors, which are trained to improve LLMs' performance while keeping the prompt concise. Liu et al. [185] involved summarization compression and semantic compression, where the former achieves compression by summarizing while the latter removes tokens with a lower impact on the semantic.

Context Conflict. Retrieval-augmented LLMs generate answers through the combined effect of parametric knowledge and contextual knowledge. As discussed in Section 3.3.2, LLMs may sometimes exhibit over-confidence, which can bring new challenges to the faithfulness of RAG systems when facing knowledge conflicts. Knowledge conflicts in RAG are situations where contextual knowledge contradicts LLMs' parametric knowledge. Longpre et al. [191] first investigated knowledge conflicts in open-domain question answering, where conflicts are automatically created by replacing all spans of the gold answer in the retrieval context with a substituted entity. Findings demonstrate that generative QA reader models (e.g., T5) tend to trust parametric memory over contextual information. By further training the retriever to learn to trust the contextual evidence with augmented training examples by entity substitution, the issue of over-reliance on parametric knowledge is mitigated. Similar findings are also reported by Li et al. [163] who demonstrated that fine-tuning LLMs on counterfactual contexts can effectively improve the controllability of LLMs when dealing with contradicts contexts. Also building upon counterfactual data augmentation, Neeman et al. [224] trained models to predict two disentangled answers, one based on contextual knowledge and the other leveraging parametric knowledge to address knowledge conflicts. Besides, Zhou et al. [386] introduced two effective prompting-based strategies, namely opinion-based prompts and counterfactual demonstrations. Opinion-based prompts transform the context to narrators' statements, soliciting the narrators' opinions, whereas counterfactual demonstrations employ counterfactual instances to improve faithfulness in situations of knowledge conflict. While Longpre et al. [191] and Li et al. [163] concentrated their research on the context of a limited single evidence setting, Chen et al. [40] further expanded this study to consider a more realistic scenario in which models consider multiple evidence passages and find models rely almost exclusively on contextual evidence.

Considering previous studies [163, 191] mostly focused on smaller models, Xie et al. [330] raised doubts about the applicability of their conclusions in the era of LLMs. Such heuristic entity-level substitution may lead to incoherent counter-memory, thereby making it trivial for LLMs to overlook the construct knowledge conflicts. By directly eliciting LLMs to generate a coherent

counter-memory that factually conflicts with the parametric memory, LLMs exhibit their high receptivity to external evidence.

Context Utilization. Despite successfully retrieving evidence relevant to factoid queries, LLMs can encounter a significant performance degradation due to insufficient utilization of the context, especially for information located in the middle of the long context window, a notable issue known as the *lost-in-the-middle* phenomenon [186]. Beyond factoid QA, recent studies have further demonstrated such a *middle-curse* also holds in abstractive summarization [254], long-form QA [41] and passage ranking [290]. One potential explanation lies in the use of **rotary positional embedding (RoPE)** [283], which is widely used in open-source LLMs, due to its excellent performance in length extrapolation [376]. As a representative relative position embedding, RoPE features a long-term decay property, which inherently biases the LLM to give precedence to current or proximate tokens, thereby diminishing its attention on those that are more distant. Another contributing factor is that the most salient information often resides at the beginning or the end of pre-training data, a characteristic commonly observed in news reports [254]. Such an issue brings forth challenges in retrieval-augmented LLMs, as retrieval-augmented LLMs are typically designed with extensive lengths to accommodate more retrieval documents.

To mitigate this crucial issue, He et al. [113] introduced several tasks specially designed for information seeking to enhance the capability of information utilization by explicitly repeating the question and extracting the index of supporting documents before generating answers. Furthermore, Zhang et al. [373] introduced **Multi-scale Positional Encoding (Ms-PoE)**, which mitigates the long-term decay effect characteristic of RoPE by rescaling position indices. Ms-PoE provides a plug-and-play solution to enhance the ability of LLMs to effectively capture information in the middle of the context without the need for additional fine-tuning. Besides, Ravaut et al. [254] proposed hierarchical and incremental summarization, which effectively preserves the salient information and compresses the length of context to avoid the *middle-curse*.

6.2.2 Contextual Alignment. Contextual alignment ensures that LLM outputs faithfully align with relevant context. This section outlines the primary components of contextual alignment, which include: (1) source attribution and (2) faithful decoding.

Source Attribution. Source attribution [120] in retrieval-augmented LLMs refers to the process by which the model identifies and utilizes the origins of information within its generation process. This component is crucial for ensuring that the outputs of RAG systems are not only relevant but also verifiable and grounded in credible sources.

To achieve source attribution in RAG systems, recent studies have been explored, which can be categorized into three lines based on the type of attribution. (1) *Plan-then-Generate*: Fierro et al. [87] introduced the blueprint model for attribution, which conceptualizes text plans as a series of questions that serve as blueprints for generation process, dictating both the content and the sequence of the output. Compared with abstractive questions, Huang et al. [119] enabled the model to first ground to extractive evidence spans, which guides the subsequent generation process. Leveraging either abstract questions or extractive spans as planning facilitates a built-in attribution mechanism, as they provide a natural link between retrieved information and the subsequent generation. Similarly, Slobodkin et al. [278] broke down the conventional end-to-end generation process into three intuitive stages: content selection, sentence planning, and sentence fusion. By initially identifying relevant source segments and subsequently conditioning the generation process on them, the selected segments naturally serve as attributions. (2) *Generate-then-Reflect*: Asai et al. [11] proposed training the LLM to generate text with reflection tokens. These reflection tokens empower the LLM to decide whether to retrieve, assess the relevance of the retrieved document, and critique its own generation to ensure attributability. By critiquing its generation. Furthermore,

Ye et al. [344] introduced AGREE, designed to facilitate self-grounding in LLMs. AGREE trains LLMs to generate well-grounded claims with citations and identify claims that lack verification. An iterative retrieval process is then employed to actively seek additional information for these unsupported statements. (3) *Self-Attribution*: In addition to leveraging external supervised signals for attribution, Qi et al. [245] proposed a self-attribution mechanism that utilizes model-internal signals. It operates by first identifying context-sensitive answer tokens, which are then paired with retrieved documents that contributed to the model generation via saliency methods.

Faithful Decoding. Despite significant optimizations in the RAG pipeline that facilitate the incorporation of highly relevant content into the model's context, current LLMs still cannot guarantee faithful generation. The unfaithful utilization of relevant context by LLMs undermines the reliability of their outputs, even when the sources of information are verifiably accurate. Wu et al. [327] analyzed the model's knowledge preference when internal knowledge conflicts with contextual information and observed the tug-of-war between the LLM's internal prior and external evidence. To tackle this issue, recent research [271, 326] has focused on faithful decoding within RAG systems, aiming to improve the models' ability to generate content that faithfully aligns with contextual information. Shi et al. [271] presented CAD, which modifies the model's original output probability distribution into the pointwise mutual information formulation. The strategy operates by amplifying the difference between the output probabilities when a model is used with and without context, thereby enhancing the faithfulness of LLMs to the provided context. Li et al. [170] adopted a semi-parametric language modeling approach [147] which facilitates the integration of contextual spans of arbitrary length into LM generations. The generation is then verified via speculative decoding, further ensuring model faithfulness. More recently, Wu et al. [326] proposed faithfulness-oriented decoding, which leverages a lightweight faithfulness detector to monitor the beam-search process. The detector leverages fine-grained decoding dynamics including sequence likelihood, uncertainty quantification, context influence, and semantic alignment to synchronously detect unfaithful sentences. When an unfaithful generation is detected, it triggers the backtrack operation and selects the beam with the more faithful score, thus ensuring greater faithfulness to the retrieval sources.

7 Future Discussion

As the field of research on hallucinations in LLMs continues to evolve, our focus shifts towards the next horizon of inquiry. We explore prospective areas of study, notably the phenomenon of hallucinations in vision-language models (Section 7.1) and the challenge of delineating and understanding knowledge boundaries within LLMs (Section 7.2).

7.1 Hallucination in LVLMs

Enabling the visual perception ability, along with exceptional language understanding and generation capabilities, LVLMs have exhibited remarkable vision-language capabilities [47, 122, 183, 198, 351, 352, 356, 388]. Unlike previous pre-trained multi-modal models that gain limited vision-language abilities from large-scale visual-language pre-training datasets [167, 194, 321, 383], LVLMs exploit advanced LLMs to unleash the power of interacting with humans and the environment. The consequent diverse applications of LVLMs also bring new challenges to maintaining the reliability of such systems. Recent studies have revealed that current LVLMs are suffering from multi-modal hallucinations, where models provide responses misaligned with the corresponding visual information [103, 184, 293]. Such multi-modal hallucinations could cause unexpected behaviors when applying LVLMs to real-world scenarios, which therefore had to be further investigated and mitigated.

Li et al. [175] and Lovenia et al. [192] took the first step towards evaluating the object hallucinations in the LVLMs. Evaluations and experiments reveal that current LVLMs are prone to generate

inconsistent responses with respect to the associated image, including non-existent objects, wrong object types, and attributes, incorrect semantic relationships, etc. [311, 357]. Furthermore, Liu et al. [182], Zong et al. [391] and Liu et al. [181] show that LVLMs can be easily fooled and experience a severe performance drop due to their over-reliance on the strong language prior, as well as its inferior ability to defend against inappropriate user inputs [111, 132]. Jiang et al. [136], Wang et al. [311] and Jing et al. [139] took a step forward to holistically evaluate multi-modal hallucination. What's more, when presented with multiple images, LVLMs sometimes mix or miss parts of the visual context, as well as fail to understand temporal or logical connections between them, which might hinder their usage in many scenarios, yet properly identifying the reason for such disorders and tackling them still requires continued efforts. Despite the witnessed perception errors, LVLMs can generate flawed logical reasoning results even when correctly recognizing all visual elements, which remains further investigation.

Efforts have been made towards building more robust LVLMs. Gunjal et al. [107], Lu et al. [193], Wang et al. [312], and Liu et al. [182] proposed to further finetune the model for producing more truthful and helpful responses. Another line of work chooses to post hoc rectify the generated inconsistent content, such as [387], and [345], which introduced expert models. To free from the external tools, Leng et al. [159], Huang et al. [121], and Zhao et al. [375] tried to fully utilize the LVLM itself to alleviate hallucinations. Though proved to be effective, those methods usually require additional data annotations, visual experts, training phases, and computational costs, which prevent LVLMs from effectively scaling and generalizing to various fields. Thus, more universal approaches are expected to build a more reliable system, such as faithful and large-scale visual-text pre-training and alignment methods.

7.2 Understanding Knowledge Boundary in LLMs

Despite the impressive capacity to capture factual knowledge from extensive data, LLMs still face challenges in recognizing their own knowledge boundaries. This shortfall leads to the occurrence of hallucinations, where LLMs confidently produce falsehoods without an awareness of their own knowledge limits [232, 261, 380]. Numerous studies delve into probing knowledge boundaries of LLMs, utilizing strategies such as evaluating the probability of a correct response in a multiple-choice setting [141], or quantifying the model's output uncertainty by evaluating the similarity among sets of sentences with uncertain meanings.

Furthermore, a line of work [13, 31, 169, 217] has revealed that LLMs contain latent structures within their activation space that relate to beliefs about truthfulness. Recent research [277] also found substantial evidence for LLMs' ability to encode the unanswerability of questions, despite the fact that these models exhibit overconfidence and produce hallucinations when presented with unanswerable questions. Nonetheless, Levinstein and Herrmann [160] have employed empirical and conceptual tools to probe whether or not LLMs have beliefs. Their empirical results suggest that current lie-detector methods for LLMs are not yet fully reliable, and the probing methods proposed by Burns et al. [31] and Azaria and Mitchell [13] do not adequately generalize. Consequently, whether we can effectively probe LLMs' internal beliefs is ongoing, requiring further research.

8 Conclusion

In this comprehensive survey, we have undertaken an in-depth examination of hallucinations within LLMs, delving into the intricacies of their underlying causes, pioneering detection methodologies as well as related benchmarks, and effective mitigation strategies. Although significant strides have been taken, the conundrum of hallucination in LLMs remains a compelling and ongoing concern that demands continuous investigation. Moreover, we envision this survey as a guiding beacon for researchers dedicated to advancing robust IR systems and trustworthy artificial intelligence.

By navigating the complex landscape of hallucinations, we hope to empower these dedicated individuals with invaluable insights that drive the evolution of AI technologies toward greater reliability and safety [66, 125, 156, 216, 305].

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. arXiv:2303.09540. Retrieved from https://arxiv.org/abs/2303.09540

[2] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. arXiv:2307.16877. Retrieved from https://arxiv.org/abs/2307.16877

[3] Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? arXiv:2305.18248. Retrieved from https://arxiv.org/abs/2305.18248

[4] Perplexity AI. 2023. Perplexity AI. https://www.perplexity.ai/

[5] Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yun-Hsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. arXiv:2302.05578. Retrieved from https://arxiv.org/abs/2302.05578

[6] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. arXiv:2204.06031. Retrieved from https://arxiv.org/abs/2204.06031

[7] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. arXiv:2312.11805. Retrieved from https://arxiv.org/abs/2312.11805

[8] Anthropic. 2023. Claude. Retrieved from https://claude.ai/

[9] Antropic. 2024. Claude 3 Haiku: Our Fastest Model Yet. 2024. Retrieved from https://www.anthropic.com/news/claude-3-haiku

[10] ArXiv. 2023. arxiv dataset. Retrieved from https://www.kaggle.com/datasets/Cornell-University/arxiv/versions/134

[11] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv:2310.11511. Retrieved from https://arxiv.org/abs/2310.11511

[12] Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. arXiv:2403.03187. Retrieved from https://arxiv.org/abs/2403.03187

[13] Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when its lying. arXiv: 2304.13734. Retrieved from https://arxiv.org/abs/2304.13734

[14] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv:2306.04136. Retrieved from https://arxiv.org/abs/2306.04136

[15] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv:2302.04023. Retrieved from https://arxiv.org/abs/2302.04023

[16] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. arXiv:2401.05856. Retrieved from https://arxiv.org/abs/2401.05856

[17] Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial nli for factual correctness in text summarisation models. arXiv:2005.11739. Retrieved from https://arxiv.org/abs/2005.11739

[18] Pierre Basso. 1993. Conditional causal logic: A formal theory of the meaning generating processes in a cognitive system. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Ruzena Bajcsy (Ed.), Morgan Kaufmann, 845–851. Retrieved from http://ijcai.org/Proceedings/93-2/Papers/002.pdf

[19] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150. Retrieved from https://arxiv.org/abs/2004.05150

[20] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.), ACM, New York, NY, 610–623. DOI: https://doi.org/10.1145/3442188.3445922

[21] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.), 1171–1179. Retrieved from https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html

[22] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". arXiv:2309.12288. Retrieved from https://arxiv.org/abs/2309.12288

[23] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. GPT-NeoX-20B: An open-source autoregressive language model. arXiv:2204.06745. Retrieved from https://arxiv.org/abs/2204.06745

[24] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the International Conference on Machine Learning (ICML '22)*. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.), Proceedings of Machine Learning Research, Vol. 162, PMLR, 2206–2240. Retrieved from https://proceedings.mlr.press/v162/borgeaud22a.html

[25] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. arXiv:2211.03540. Retrieved from https://arxiv.org/abs/2211.03540

[26] Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345. Retrieved from https://www.jstor.org/stable/2334029

[27] Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1504–1521. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.113

[28] Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 21–29.

[29] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.), 1877–1901. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[30] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712. Retrieved from https://arxiv.org/abs/2303.12712

[31] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. arXiv:2212.03827. Retrieved from https://arxiv.org/abs/2212.03827

[32] Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. 2023. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 12541–12560. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.835

[33] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. arXiv:2303.04226. Retrieved from https://arxiv.org/abs/2303.04226

[34] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. arXiv:2202.07646. Retrieved from https://arxiv.org/abs/2202.07646

[35] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *Proceedings of the International Conference on 30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[36] Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. 2023. KL-divergence guided temperature sampling. 2306.01286. Retrieved from https://arxiv.org/abs/2306.01286

[37] Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 12707–12730. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-ACL.805

[38] Haw-Shiuan Chang and Andrew McCallum. 2022. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8048–8073. DOI: https://doi.org/10.18653/v1/2022.acl-long.554

[39] Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of LLMs: Opportunities and challenges. arXiv:2311.05656. Retrieved from https://arxiv.org/abs/2311.05656

[40] Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*. Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.), Association for Computational Linguistics, 2292–2307. DOI: https://doi.org/10.18653/V1/2022.EMNLP-MAIN.146

[41] Hung-Ting Chen, Fangyuan Xu, Shane A. Arora, and Eunsol Choi. 2023. Understanding retrieval augmentation for long-form question answering. arXiv:2310.12150. Retrieved from https://arxiv.org/abs/2310.12150

[42] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking factuality evaluation of large language models. arXiv:2310.00741. Retrieved from https://arxiv.org/abs/2310.00741

[43] Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense X retrieval: What retrieval granularity should we use? /arXiv:2312.06648. Retrieved from https://arxiv.org/abs/2312.06648

[44] Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. Spiral of silence: How is large language model killing information retrieval? – A case study on open domain question answering. arXiv:2404.10496. Retrieved from https://arxiv.org/abs/2404.10496

[45] Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/9b6d7202750e8e32cd5270eb7fc131f7-Abstract-Conference.html

[46] Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions. arXiv:2308.12674. Retrieved from https://arxiv.org/abs/2308.12674

[47] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Measuring and improving chain-of-thought reasoning in vision-language models. arXiv:2309.04461. Retrieved from https://arxiv.org/abs/2309.04461

[48] Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. arXiv:2406.12534. Retrieved from https://arxiv.org/abs/2406.12534

[49] Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. arXiv:2310.03368. Retrieved from https://arxiv.org/abs/2310.03368

[50] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality detection in generative AI–A tool augmented framework for multi-task and multi-domain scenarios. arXiv:2307.13528. Retrieved from https://arxiv.org/abs/2307.13528

[51] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? arXiv:2305.01937. Retrieved from https://arxiv.org/abs/2305.01937

[52] David Chiang and Peter Cholak. 2022. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7654–7664. DOI: https://doi.org/10.18653/v1/2022.acl-long.527

[53] Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. arXiv:2310.09044. Retrieved from https://arxiv.org/abs/2310.09044

[54] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24 (2023), 240:1–240:113. Retrieved from http://jmlr.org/papers/v24/22-1144.html

[55] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 4299–4307. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[56] Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. BeamAggR: Beam aggregation reasoning over multi-source knowledge for multi-hop question answering. arXiv:2406.19820. Retrieved from https://arxiv.org/abs/2406.19820

[57] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: advances, frontiers and future. arXiv:2309.15402. Retrieved from https://arxiv.org/abs/2309.15402

[58] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. arXiv:2309.03883. Retrieved from https://arxiv.org/abs/2309.03883

[59] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv:2210.11416. Retrieved from https://arxiv.org/abs/2210.11416

[60] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. arXiv:2110.14168. Retrieved from https://arxiv.org/abs/2110.14168

[61] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. arXiv:2305.13281. Retrieved from https://arxiv.org/abs/2305.13281

[62] Together Computer. 2023. RedPajama: An Open Dataset for Training Large Language Models. Retrieved from https://github.com/togethercomputer/RedPajama-Data

[63] Ajeya Cotra. 2021. Why AI Alignment Could Be Hard with Modern Deep Learning. Cold Takes. Retrieved from https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/

[64] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. arXiv:2401.14887. Retrieved from https://arxiv.org/abs/2401.14887

[65] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8493–8502. DOI: https://doi.org/10.18653/v1/2022.acl-long.581

[66] Damai Dai, Wenbin Jiang, Qingxiu Dong, Yajuan Lyu, Qiaoqiao She, and Zhifang Sui. 2022. Neural knowledge bank for pretrained transformers. arXiv:2208.00399. Retrieved from https://arxiv.org/abs/2208.00399

[67] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023. LLMs may dominate information access: Neural retrievers are biased towards LLM-Generated texts. arXiv:2310.20501. Retrieved from https://arxiv.org/abs/2310.20501

[68] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=H1edEyBKDS

[69] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6491–6506. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.522

[70] Maria Angels de Luis Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, et al. 2024. RAG vs Fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv:2401.08406. Retrieved from https://arxiv.org/abs/2401.08406

[71] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language modeling is compression. arXiv:2309.10668. Retrieved from https://arxiv.org/abs/2309.10668

[72] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '19)*. Jill Burstein, Christy Doran, and Thamar Solorio (Eds.), Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/V1/N19-1423

[73] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv preprint* abs/2309.11495 (2023). Retrieved from https://arxiv.org/abs/2309.11495

[74] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve Only When It Needs: Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models. arXiv:2402.10612. Retrieved from https://arxiv.org/abs/2402.10612

[75] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. arXiv:2309.13345. Retrieved from https://arxiv.org/abs/2309.13345

[76] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5055–5070. DOI: https://doi.org/10.18653/v1/2020.acl-main.454

[77] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2197–2214. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.168

[78] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. arXiv:2105.00071. Retrieved from https://arxiv.org/abs/2105.00071

[79] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2587–2601. DOI: https://doi.org/10.18653/v1/2022.naacl-main.187

[80] Alexander R. Fabbri, Wojciech KryŚciński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409. DOI: https://doi.org/10.1162/tacl_a_00373

[81] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2214–2220. DOI: https://doi.org/10.18653/v1/P19-1213

[82] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3558–3567. DOI: https://doi.org/10.18653/v1/P19-1346

[83] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 889–898. DOI: https://doi.org/10.18653/v1/P18-1082

[84] Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2023. Improving factual consistency of text summarization by adversarially decoupling comprehension and embellishment abilities of LLMs. arXiv:2310.19347. Retrieved from https://arxiv.org/abs/2310.19347

[85] Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Cook: Empowering general-purpose language models with modular and collaborative knowledge. arXiv:2305.09955. Retrieved from https://arxiv.org/abs/2305.09955

[86] Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023. Retrieval-generation synergy augmented large language models. arXiv:2310.05149. Retrieved from https://arxiv.org/abs/2310.05149

[87] Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. arXiv:2404.03381. Retrieved from https://arxiv.org/abs/2404.03381

[88] Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 864–870. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.76

[89] Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for LLM hallucination detection. arXiv:2310.18344. Retrieved from https://arxiv.org/abs/2310.18344

[90] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as you desire. arXiv:2302.04166. Retrieved from https://arxiv.org/abs/2302.04166

[91] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning (ICML '16)*. Maria-Florina Balcan and Kilian Q. Weinberger (Eds.), JMLR Workshop and Conference Proceedings, Vol. 48, JMLR.org, 1050–1059. Retrieved from http://proceedings.mlr.press/v48/gal16.html

[92] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2021. The pile: An 800gb dataset of diverse text for language modeling. arXiv:2101.00027. Retrieved from https://arxiv.org/abs/2101.00027

[93] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 16477–16508. Retrieved from https://aclanthology.org/2023.acl-long.910

[94] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. arXiv:2304.02554. Retrieved from https://arxiv.org/abs/2304.02554

[95] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. arXiv:2104.08821. Retrieved from https://arxiv.org/abs/2104.08821

[96] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards interactive and explainable LLMs-augmented recommender system. arXiv:2303.14524. Retrieved from https://arxiv.org/abs/2303.14524

[97] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv:2405.05904. Retrieved from https://arxiv.org/abs/

[98] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.), ACM, New York, NY, 166–175. DOI: https://doi.org/10.1145/3292500.3330955

[99] Google. 2023. Bard. Retrieved from https://bard.google.com/

[100] Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 3592–3603. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.322

[101] Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1449–1462. DOI: https://doi.org/10.18653/v1/2021.naacl-main.114

[102] Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. 2022. Improving controllable text generation with position-aware weighted decoding. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (ACL '22)*. Association for Computational Linguistics, 3449–3467. DOI: https://doi.org/10.18653/v1/2022.findings-acl.272

[103] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. arXiv:2310.14566. Retrieved from https://arxiv.org/abs/2310.14566

[104] Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. arXiv:2303.16104. Retrieved from https://arxiv.org/abs/2303.16104

[105] Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1059–1075. Retrieved from https://aclanthology.org/2023.eacl-main.75

[106] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. arXiv:2306.11644. Retrieved from https://arxiv.org/abs/2306.11644

[107] Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. arXiv:2308.06394. Retrieved from https://arxiv.org/abs/2308.06394

[108] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*, Proceedings of Machine Learning Research, Vol. 119, PMLR, 3929–3938. Retrieved from http://proceedings.mlr.press/v119/guu20a.html

[109] Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. 2020. Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 901–910. Retrieved from https://aclanthology.org/2020.lrec-1.113

[110] Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics* 8 (2020), 156–171. DOI: https://doi.org/10.1162/tacl_a_00306

[111] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in MLLMs. arXiv:2402.03757. Retrieved from https://arxiv.org/abs/2402.03757

[112] Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. arXiv:2301.00303. Retrieved from https://arxiv.org/abs/2301.00303

[113] Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Yuxin Liang, Hao Wang, Qianguo Sun, Songxin Zhang, Zejian Xie, and Jiaxing Zhang. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. arXiv:2311.09198. Retrieved from https://arxiv.org/abs/2311.09198

[114] Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Sanmi Koyejo, S. Mohamed,

A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/bc218a0c656e49d4b086975a9c785f47-Abstract-Datasets_and_Benchmarks.html

[115] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR '21)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=d7KBjmI3GmQ

[116] Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. arXiv:2304.00740. Retrieved from https://arxiv.org/abs/2304.00740

[117] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=rygGQyrFvH

[118] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7856–7870. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.619

[119] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, et al. 2024. Learning fine-grained grounded citations for attributed large language models. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (ACL '24)*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), Association for Computational Linguistics, 14095–14113. DOI: https://doi.org/10.18653/V1/2024.FINDINGS-ACL.838

[120] Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. Advancing large language model attribution through self-improving. arXiv:2410.13298. Retrieved from https://arxiv.org/abs/2410.13298

[121] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. arXiv:2311.17911. Retrieved from https://arxiv.org/abs/2311.17911

[122] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. arXiv:2302.14045. Retrieved from https://arxiv.org/abs/2302.14045

[123] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023a. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. arXiv:2305.08322. Retrieved from https://arxiv.org/abs/2305.08322

[124] Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. arXiv:2104.14839. Retrieved from https://arxiv.org/abs/2104.14839

[125] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023d. Transformer-Patcher: One mistake worth one neuron. In *Proceedings of the 11th International Conference on Learning Representations (ICLR '23)*. OpenReview.net. Retrieved from https://openreview.net/pdf?id=4oYUGeGBPm

[126] Siqing Huo, Negar Arabzadeh, and Charles L. A. Clarke. 2023. Retrieving supporting evidence for LLMs generated answers. arXiv:2306.13781. Retrieved from https://arxiv.org/abs/2306.13781

[127] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. arXiv:2212.12017. Retrieved from https://arxiv.org/abs/2212.12017

[128] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research* 2022 (2022). Retrieved from https://openreview.net/forum?id=jKN1pXi7b0

[129] Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24 (2023), 251:1–251:43. Retrieved from http://jmlr.org/papers/v24/23-0037.html

[130] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. arXiv:2305.03653. Retrieved from https://arxiv.org/abs/2305.03653

[131] Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. arXiv:2306.01200. Retrieved from https://arxiv.org/abs/2306.01200

[132] Joonhyun Jeong. 2023. Hijacking context in large multi-modal models. arXiv:2312.07553. Retrieved from https://arxiv.org/abs/2312.07553

[133] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. arXiv:2403.14403. Retrieved from https://arxiv.org/abs/2403.14403

[134] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys* 55, 12 (2023), 248:1–248:38. DOI: https://doi.org/10.1145/3571730

[135] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. arXiv:2310.06271. Retrieved from https://arxiv.org/abs/2310.06271

[136] Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-Eval: A universal and fine-grained hallucination evaluation framework for large vision language models. arXiv:2402.15721. Retrieved from https://arxiv.org/abs/2402.15721

[137] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. arXiv:2310.05736. Retrieved from https://arxiv.org/abs/2310.05736

[138] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. arXiv:2305.06983. Retrieved from https://arxiv.org/abs/2305.06983

[139] Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. arXiv:2311.01477. Retrieved from https://arxiv.org/abs/2311.01477

[140] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. 2024. When large language models meet vector databases: A survey. arXiv:2402.01763. Retrieved from https://arxiv.org/abs/2402.01763

[141] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. arXiv:2207.05221. Retrieved from https://arxiv.org/abs/2207.05221

[142] Greg Kamradt. 2024. The 5 Levels of Text Splitting for Retrieval. Youtube. Retrieved from https://www.youtube.com/watch?v=8OJC21T2SL4

[143] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the International Conference on Machine Learning(ICML '23)*. Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), Proceedings of Machine Learning Research, Vol. 202, PMLR, 15696–15707. Retrieved from https://proceedings.mlr.press/v202/kandpal23a.html

[144] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.), Association for Computational Linguistics, 6769–6781. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.550

[145] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. RealTime QA: What's the answer right now? arXiv:2207.13332. Retrieved from https://arxiv.org/abs/2207.13332

[146] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 Passes the Bar Exam. Retrieved from https://www.datascienceassn.org/sites/default/files/GPT-4%20Passes%20the%20Bar%20Exam.pdf

[147] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=HklBjCEKvH

[148] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 22199–22213.

[149] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 9332–9346. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.750

[150] Philippe Laban, Wojciech KryŚciński, Divyansh Agarwal, Alexander R. Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. LLMs as factual reasoners: Insights from existing benchmarks and beyond. arXiv:2305.14540. Retrieved from https://arxiv.org/abs/2305.14540

[151] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177. DOI: https://doi.org/10.1162/tacl_a_00453

[152] Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori 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*. Association for Computational Linguistics, 3206–3219. Retrieved from https://aclanthology.org/2023.eacl-main.234

[153] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 6402–6413. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[154] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702. Retrieved from https://arxiv.org/abs/2307.13702

[155] Barrett Martin Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. arXiv:2310.13189. Retrieved from https://arxiv.org/abs/2310.13189

[156] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8424–8445. DOI: https://doi.org/10.18653/v1/2022.acl-long.577

[157] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 34586–34599.

[158] Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. arXiv:2310.03951. Retrieved from https://arxiv.org/abs/2310.03951

[159] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. arXiv:2311.16922. Retrieved from https://arxiv.org/abs/2311.16922.

[160] BA Levinstein and Daniel A. Herrmann. 2023. Still no lie detector for language models: Probing empirical and conceptual roadblocks. arXiv:2307.00175. Retrieved from https://arxiv.org/abs/2307.00175

[161] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7871–7880. DOI: https://doi.org/10.18653/v1/2020.acl-main.703

[162] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.), Retrieved from https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[163] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 1774–1793. DOI: https://doi.org/10.18653/v1/2023.findings-acl.112

[164] Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Towards faithful chain-of-thought: large language models are bridging reasoners. arXiv:2405.18915. Retrieved from https://arxiv.org/abs/2405.18915

[165] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv:2401.03205. Retrieved from https://arxiv.org/abs/2401.03205

[166] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv:2305.11747. Retrieved from https://arxiv.org/abs/2305.11747

[167] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597. Retrieved from https://arxiv.org/abs/2301.12597

[168] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. In *Proceedings of the SIGIR Workshop on eCommerce Co-located with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 23)*. Surya Kallumadi, Yubin Kim, Tracy Holloway King, Shervin Malmasi, Maarten de Rijke, and Jacopo Tagliabue (Eds.), CEUR Workshop Proceedings, Vol. 3589, CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-3589/paper_2.pdf

[169] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. arXiv:2306.03341. Retrieved from https://arxiv.org/abs/2306.03341

[170] Minghan Li, Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Wen-tau Yih, and Xi Victoria Lin. 2024. Nearest neighbor speculative decoding for LLM generation and attribution. arXiv:2405.19325. Retrieved from https://arxiv.org/abs/2405.19325

[171] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. arXiv:2203.05227. Retrieved from https://arxiv.org/abs/2203.05227

[172] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. arXiv:2210.15097. Retrieved from https://arxiv.org/abs/2210.15097

[173] Yucheng Li. 2023. Unlocking context constraints of LLMs: Enhancing context efficiency of LLMs with self-information-based content filtering. arXiv:2304.12102. Retrieved from https://arxiv.org/abs/2304.12102

[174] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need II: phi-1.5 technical report. arXiv:2309.05463. Retrieved from https://arxiv.org/abs/2309.05463

[175] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. Retrieved from https://arxiv.org/abs/2305.10355

[176] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. arXiv:2303.14070. Retrieved from https://arxiv.org/abs/2303.14070

[177] Zuchao Li, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. 2023i. BatGPT: A bidirectional autoregessive talker from generative pre-trained transformer. arXiv:2307.00360 (2023). Retrieved from https://arxiv.org/abs/2307.00360

[178] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81. Retrieved from https://aclanthology.org/W04-1013

[179] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 3214–3252. DOI: https://doi.org/10.18653/v1/2022.acl-long.229

[180] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. 2023. Exposing attention glitches with flip-flop language modeling. arXiv:2306.00946. Retrieved from https://arxiv.org/abs/2306.00946

[181] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V(Ision), LLaVA-1.5, and Other Multi-Modality Models. Retrieved from https://arxiv.org/abs/2310.14566

[182] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv:2306.14565. Retrieved from https://arxiv.org/abs/2306.14565

[183] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. arXiv:2304.08485. Retrieved from https://arxiv.org/abs/2304.08485

[184] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. arXiv:2402.00253. Retrieved from https://arxiv.org/abs/2402.00253

[185] Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. TCRA-LLM: Token compression retrieval augmented large language model for inference cost reduction. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 9796–9810. Retrieved from https://aclanthology.org/2023.findings-emnlp.655

[186] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. arXiv:2307.03172. Retrieved from https://arxiv.org/abs/2307.03172

[187] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv:2303.16634. Retrieved from https://arxiv.org/abs/2303.16634

[188] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from http://arxiv.org/abs/1907.11692

[189] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. arXiv:2308.05374. Retrieved from https://arxiv.org/abs/2308.05374

[190] Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. 2023. Instruction position matters in sequence generation with large language models. arXiv:2308.12097. Retrieved from https://arxiv.org/abs/2308.12097

[191] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*. Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.), Association for Computational Linguistics, 7052–7063. DOI: https://doi.org/10.18653/V1/2021.EMNLP-MAIN.565

[192] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. arXiv:2310.05338. Retrieved from https://arxiv.org/abs/2310.05338

[193] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023. Evaluation and Mitigation of Agnosia in Multimodal Large Language Models. arXiv:2309.04041. Retrieved from https://arxiv.org/abs/2309.04041

[194] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv:2002.06353. Retrieved from https://arxiv.org/abs/2002.06353

[195] Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. arXiv:2309.02654. Retrieved from https://arxiv.org/abs/2309.02654

[196] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. arXiv:2303.15621. Retrieved from https://arxiv.org/abs/2303.15621

[197] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. arXiv:2305.14283. Retrieved from https://arxiv.org/abs/2305.14283

[198] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. arXiv:2306.05424. Retrieved from https://arxiv.org/abs/2306.05424

[199] Fiona Macpherson and Dimitris Platchias. 2013. *Hallucination: Philosophy and Psychology*. MIT Press. Retrieved from https://books.google.com/books?hl=zh-CN{&}lr={&}id=_bwtAAAAQBAJ{&}oi=fnd{&}pg=PR5{&}dq=Hallucination:+Philosophy+and+psychology{&}ots=2E62kf7_yC{&}sig=rH9HGXYacNkxOJNMVbw514aChZo

[200] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. ExpertQA: Expert-curated questions and attributed answers. arXiv:2309.07852. Retrieved from https://arxiv.org/abs/2309.07852

[201] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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 (Volume 1: Long Papers) (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 9802–9822. DOI: https://doi.org/10.18653/v1/2023.acl-long.546

[202] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv:2303.08896. Retrieved from https://arxiv.org/abs/2303.08896

[203] Udi Manber and Gene Myers. 1993. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing* 22, 5 (1993), 935–948.

[204] Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking feedback improves query rewriting for RAG. arXiv:2405.14431. Retrieved from https://arxiv.org/abs/2405.14431

[205] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1906–1919. DOI: https://doi.org/10.18653/v1/2020.acl-main.173

[206] Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2173–2185. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.170

[207] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html

[208] Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of the 11th International Conference on Learning Representations (ICLR '23)*. OpenReview.net. Retrieved from https://openreview.net/pdf?id=MkbcAHIYgyS

[209] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 3456–3468. DOI: https://doi.org/10.18653/v1/2021.acl-long.268

[210] Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. arXiv:2308.00436. Retrieved from https://arxiv.org/abs/2308.00436

[211] Microsoft. 2023. New Bing. Retrieved from https://www.bing.com/new

[212] Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2023. SILO language models: Isolating legal risk in a nonparametric datastore. arXiv:2308.04430. Retrieved from https://arxiv.org/abs/2308.04430

[213] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251. Retrieved from https://arxiv.org/abs/2305.14251

[214] Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1322–1336. DOI: https://doi.org/10.18653/v1/2021.naacl-main.104

[215] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *Proceedings of the 10th International Conference on Learning Representations (ICLR '22)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=0DcZxeWfOPt

[216] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *Proceedings of the International Conference on Machine Learning(ICML '22)*. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.), Proceedings of Machine Learning Research, Vol. 162, PMLR, 15817–15831. Retrieved from https://proceedings.mlr.press/v162/mitchell22a.html

[217] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodola. 2022. Relative representations enable zero-shot latent space communication. arXiv:2209.15430. Retrieved from https://arxiv.org/abs/2209.15430

[218] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. arXiv:2402.09906. Retrieved from https://arxiv.org/abs/2402.09906

[219] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL '23)*. Andreas Vlachos and Isabelle Augenstein (Eds.), Association for Computational Linguistics, 2006–2029. DOI: https://doi.org/10.18653/V1/2023.EACL-MAIN.148

[220] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. arXiv:2307.06908. Retrieved from https://arxiv.org/abs/2307.06908

[221] Inderjeet Nair, Aparna Garimella, Balaji Vasan Srinivasan, Natwar Modani, Niyati Chhaya, Srikrishna Karanam, and Sumit Shekhar. 2023. A neural CRF-based hierarchical approach for linear text segmentation. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (EACL '23)*. Andreas Vlachos and Isabelle Augenstein (Eds.), Association for Computational Linguistics, 853–863. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-EACL.65

[222] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2727–2733. DOI: https://doi.org/10.18653/v1/2021.eacl-main.235

[223] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 116–122. Retrieved from https://aclanthology.org/2023.eacl-main.9

[224] Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 10056–10070. DOI: https://doi.org/10.18653/V1/2023.ACL-LONG.559

[225] Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do LLMs need retrieval augmentation? Mitigating LLMs' overconfidence helps retrieval augmentation. arXiv:2402.11457. Retrieved from https://arxiv.org/abs/2402.11457

[226] Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. arXiv:2309.09117. Retrieved from https://arxiv.org/abs/2309.09117

[227] Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. In *Proceedings of the International Conference on Findings of the Association for Computational*

Linguistics (NAACL '22). Association for Computational Linguistics, 693–702. DOI: https://doi.org/10.18653/v1/2022.findings-naacl.52

[228] OpenAI. 2022. Introducing chatgpt. Retrieved from https://openai.com/blog/chatgpt

[229] OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[230] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[231] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? Comparing knowledge injection in LLMs. arXiv:2312.05934. Retrieved from https://arxiv.org/abs/2312.05934

[232] Lorenzo Pacchiardi, Alex J Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y Pan, Yarin Gal, Owain Evans, and Jan Brauner. 2023. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. arXiv:2309.15840. Retrieved from https://arxiv.org/abs/2309.15840

[233] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4812–4829. DOI: https://doi.org/10.18653/v1/2021.naacl-main.383

[234] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. arXiv:2308.03188. Retrieved from https://arxiv.org/abs/2308.03188

[235] Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching LLMs credibility-aware generation. arXiv:2404.06809. Retrieved from https://arxiv.org/abs/2404.06809

[236] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318. DOI: https://doi.org/10.3115/1073083.1073135

[237] Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.), Association for Computational Linguistics, 1173–1186. DOI: https://doi.org/10.18653/V1/2020.EMNLP-MAIN.89

[238] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. arXiv:2402.13950. Retrieved from https://arxiv.org/abs/2402.13950

[239] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. DOI: https://doi.org/10.1016/J.PATTER.2021.100336

[240] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. arXiv:2306.01116. Retrieved from https://arxiv.org/abs/2306.01116

[241] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. arXiv:2304.03277. Retrieved from https://arxiv.org/abs/2304.03277

[242] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 13387–13434. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-ACL.847

[243] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2463–2473. DOI: https://doi.org/10.18653/v1/D19-1250

[244] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. arXiv:2210.03350. Retrieved from https://arxiv.org/abs/2210.03350

[245] Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. arXiv:2406.13663. Retrieved from https://arxiv.org/abs/2406.13663

[246] Zhixiao Qi, Yijiong Yu, Meiqi Tu, Junyi Tan, and Yongfeng Huang. 2023. FoodGPT: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. arXiv:2308.10173. Retrieved from https://arxiv.org/abs/2308.10173

[247] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. arXiv:2212.09597. Retrieved from https://arxiv.org/abs/2212.09597

[248] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[249] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[250] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. arXiv:2305.18290. Retrieved from https://arxiv.org/abs/2305.18290

[251] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21 (2020), 140:1–140:67. Retrieved from http://jmlr.org/papers/v21/20-074.html

[252] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. arXiv:2302.00083. Retrieved from https://arxiv.org/abs/2302.00083

[253] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR '16)*. Yoshua Bengio and Yann LeCun (Eds.), Retrieved from http://arxiv.org/abs/1511.06732

[254] Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. arXiv:2310.10570. Retrieved from https://arxiv.org/abs/2310.10570

[255] Vipula Rawte, Amit P. Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. arXiv:2309.05922. Retrieved from https://arxiv.org/abs/2309.05922

[256] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 Retrieved from https://arxiv.org/abs/2403.05530

[257] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. arXiv:2307.11019. Retrieved from https://arxiv.org/abs/2307.11019

[258] Reuters. 2023. U.S. Copyright Office Says Some AI-Assisted Works May Be Copyrighted. Retrieved from https://www.reuters.com/world/us/us-copyright-office-says-some-ai-assisted-works-may-be-copyrighted-2023-03-15/

[259] Nina Rimsky. 2023. Modulating Sycophancy in an RLHF Model via Activation Steering. Retrieved from https://www.alignmentforum.org/posts/zt6hRsDE84HeBKh7E/reducing-sycophancy-and-improving-honesty-via-activation

[260] Nina Rimsky. 2023. Reducing Sycophancy and Improving Honesty via Activation Steering. Retrieved from https://www.alignmentforum.org/posts/zt6hRsDE84HeBKh7E/reducing-sycophancy-and-improving-honesty-via-activation

[261] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? arXiv:2303.11156. Retrieved from https://arxiv.org/abs/2303.11156

[262] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. arXiv:2110.05456. Retrieved from https://arxiv.org/abs/2110.05456

[263] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. arXiv:2401.18059. Retrieved from https://arxiv.org/abs/2401.18059

[264] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. arXiv:2206.05802. Retrieved from https://arxiv.org/abs/2206.05802

[265] John Schulman. 2023. Reinforcement Learning from Human Feedback: Progress and Challenges. Berkeley EECS. Retrieved from https://www.youtube.com/watch?v=hhiLw5Q_UFg

[266] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv:1707.06347. Retrieved from https://arxiv.org/abs/1707.06347

[267] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6594–6604. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.529

[268] Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. arXiv:2402.14207. Retrieved from https://arxiv.org/abs/2402.14207

[269] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. arXiv:2305.15294. Retrieved from https://arxiv.org/abs/2305.15294

[270] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. 2023. Towards understanding sycophancy in language models. arXiv:2310.13548. Retrieved from https://arxiv.org/abs/2310.13548

[271] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. arXiv:2305.14739. Retrieved from https://arxiv.org/abs/2305.14739

[272] Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2023. In-context pretraining: language modeling beyond document boundaries. arXiv:2310.10638. Retrieved from https://arxiv.org/abs/2310.10638

[273] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-augmented black-box language models. arXiv:2301.12652. Retrieved from https://arxiv.org/abs/2301.12652

[274] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (EMNLP '21)*. Association for Computational Linguistics, 3784–3803. DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.320

[275] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. arXiv:2305.09617. Retrieved from https://arxiv.org/abs/2305.09617

[276] Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=HJedXaEtvS

[277] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory unanswerablity: Finding truths in the hidden states of over-confident large language models. arXiv:2310.11877. Retrieved from https://arxiv.org/abs/2310.11877

[278] Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. arXiv:2403.17104. Retrieved from https://arxiv.org/abs/2403.17104

[279] Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3356–3362. DOI: https://doi.org/10.18653/v1/D19-1331

[280] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 8273–8288. Retrieved from https://aclanthology.org/2022.emnlp-main.566

[281] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS '20)*. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.), Retrieved from https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

[282] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2024. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. arXiv:2407.12883. Retrieved from https://arxiv.org/abs/2407.12883

[283] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced transformer with rotary position embedding. arXiv:2104.09864. Retrieved from https://arxiv.org/abs/2104.09864

[284] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. arXiv:2403.10081. Retrieved from https://arxiv.org/abs/2403.10081

[285] Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics (ACL '22)*. Association for Computational Linguistics, 566–581. DOI: https://doi.org/10.18653/v1/2022.findings-acl.48

[286] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (LLM)? A.K.A. will LLMs replace knowledge graphs? arXiv:2308.10168. Retrieved from https://arxiv.org/abs/2308.10168

[287] Ilya Sutskever. 2023. An Observation on Generalization. Youtube. Retrieved from https://www.youtube.com/watch?v=AKMuA_TVz3A{&}t=5s

[288] Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language model via meta learning. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=L6L1CJQ2PE

[289] Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain QA? arXiv:2401.11911. Retrieved from https://arxiv.org/abs/2401.11911

[290] Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. arXiv:2310.07712. Retrieved from https://arxiv.org/abs/2310.07712

[291] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv:2104.08663. Retrieved from https://arxiv.org/abs/2104.08663

[292] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *ArXiv preprint* abs/1910.08684 (2019). Retrieved from https://arxiv.org/abs/1910.08684

[293] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. arXiv:2401.06209. Retrieved from https://arxiv.org/abs/2401.06209

[294] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv:2401.01313. Retrieved from https://arxiv.org/abs/2401.01313

[295] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971. Retrieved from https://arxiv.org/abs/2302.13971

[296] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. Retrieved from https://arxiv.org/abs/2307.09288

[297] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 10014–10037. Retrieved from https://aclanthology.org/2023.acl-long.557

[298] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388. Retrieved from https://arxiv.org/abs/2305.04388

[299] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. arXiv:2307.15343. Retrieved from https://arxiv.org/abs/2307.15343

[300] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748. Retrieved from http://arxiv.org/abs/1807.03748

[301] Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5956–5965. Retrieved from https://aclanthology.org/2022.emnlp-main.399

[302] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. arXiv:2307.03987. Retrieved from https://arxiv.org/abs/2307.03987

[303] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5797–5808. DOI: https://doi.org/10.18653/v1/P19-1580

[304] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. FreshLLMs: Refreshing large language models with search engine augmentation. arXiv:2310.03214. Retrieved from https://arxiv.org/abs/2310.03214

[305] David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2864–2880. Retrieved from https://aclanthology.org/2023.eacl-main.210

[306] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5008–5020. DOI: https://doi.org/10.18653/v1/2020.acl-main.450

[307] Binjie Wang, Ethan Chern, and Pengfei Liu. 2023. ChineseFactEval: A Factuality Benchmark for Chinese LLMs.

[308] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv:2310.07521. Retrieved from https://arxiv.org/abs/2310.07521

[309] Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3544–3552. DOI: https://doi.org/10.18653/v1/2020.acl-main.326

[310] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. arXiv:2303.04048. Retrieved from https://arxiv.org/abs/2303.04048

[311] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An LLM-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv:2311.07397. Retrieved from https://arxiv.org/abs/2311.07397

[312] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 32–45.

[313] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 9414–9423. DOI: https://doi.org/10.18653/V1/2023.EMNLP-MAIN.585

[314] Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. arXiv:2305.01879. Retrieved from https://arxiv.org/abs/2305.01879

[315] Shuting Wang, Xin Yu, Mang Wang, Weipeng Chen, Yutao Zhu, and Zhicheng Dou. 2024. RichRAG: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. arXiv:2406.12566. Retrieved from https://arxiv.org/abs/2406.12566

[316] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023. Knowledge editing for large language models: A survey. arXiv:2310.16218. Retrieved from https://arxiv.org/abs/2310.16218

[317] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 10303–10315. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.691

[318] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. arXiv:2307.12966. Retrieved from https://arxiv.org/abs/2307.12966

[319] Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. arXiv:2311.08377. Retrieved from https://arxiv.org/abs/2311.08377

[320] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1072–1086. DOI: https://doi.org/10.18653/v1/2020.acl-main.101

[321] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple visual language model pretraining with weak supervision. In *Proceedings of the 10th International Conference on Learning Representations (ICLR '22)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=GUrhfTuf_3

[322] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.

[323] Jerry W. Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models. arXiv:2308.03958. Retrieved from https://arxiv.org/abs/2308.03958

[324] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. arXiv:2112.04359. Retrieved from https://arxiv.org/abs/2112.04359

[325] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv:2308.09729. Retrieved from https://arxiv.org/abs/2308.09729

[326] Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. arXiv:2406.13692. Retrieved from https://arxiv.org/abs/2406.13692

[327] Kevin Wu, Eric Wu, and James Zou. 2024. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. arXiv:2404.10198. Retrieved from https://arxiv.org/abs/2404.10198

[328] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-Pack: Packaged resources to advance general chinese embedding. arXiv:2309.07597. Retrieved from https://arxiv.org/abs/2309.07597

[329] Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2734–2744. DOI: https://doi.org/10.18653/v1/2021.eacl-main.236

[330] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes. arXiv:2305.13300. Retrieved from https://arxiv.org/abs/2305.13300

[331] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv:2306.13063. Retrieved from https://arxiv.org/abs/2306.13063

[332] Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. arXiv:2310.04408. Retrieved from https://arxiv.org/abs/2310.04408

[333] Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 6275–6281. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.508

[334] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. arXiv:2405.18357. Retrieved from https://arxiv.org/abs/2405.18357

[335] Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2023. AI-generated images introduce invisible relevance bias to text-image retrieval. arXiv:2311.14084. Retrieved from https://arxiv.org/abs/2311.14084

[336] Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. arXiv:2310.06498. Retrieved from https://arxiv.org/abs/2310.06498

[337] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. arXiv:2312.07000. Retrieved from https://arxiv.org/abs/2312.07000

[338] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the softmax bottleneck: A high-rank RNN language model. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=HkwZSG-CZ

[339] Zhilin Yang, Thang Luong, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Mixtape: Breaking the softmax bottleneck efficiently. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS '19)*. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.), 15922–15930. Retrieved from https://proceedings.neurips.cc/paper/2019/hash/512fc3c5227f637e41437c999a2d3169-Abstract.html

[340] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2369–2380. DOI: https://doi.org/10.18653/v1/D18-1259

[341] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. LLM Lies: Hallucinations are not bugs, but features as adversarial examples. arXiv:2310.01469. Retrieved from https://arxiv.org/abs/2310.01469

[342] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv:2210.03629. Retrieved from https://arxiv.org/abs/2210.03629

[343] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. arXiv:2305.13172. Retrieved from https://arxiv.org/abs/2305.13172

[344] Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2023. Effective large language model adaptation for improved grounding. arXiv:2311.09533. Retrieved from https://arxiv.org/abs/2311.09533

[345] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. arXiv:2310.16045. Retrieved from https://arxiv.org/abs/2310.16045

[346] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 8653–8665. DOI: https://doi.org/10.18653/V1/2023.FINDINGS-ACL.551

[347] Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2024. Ask optimal questions: Aligning large language models with retriever's preference in conversational search. arXiv:2402.11827. Retrieved from https://arxiv.org/abs/2402.11827

[348] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. arXiv:2310.01558. Retrieved from https://arxiv.org/abs/2310.01558

[349] Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. arXiv:2212.01326. Retrieved from https://arxiv.org/abs/2212.01326

[350] Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. arXiv:2303.14725. Retrieved from https://arxiv.org/abs/2303.14725

[351] Weijiang Yu, Jian Liang, Lei Ji, Lu Li, Yuejian Fang, Nong Xiao, and Nan Duan. 2021. Hybrid reasoning network for video-based commonsense captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5213–5221.

[352] Weijiang Yu, Haofan Wang, Guohao Li, Nong Xiao, and Bernard Ghanem. 2023. Knowledge-aware global reasoning for situation recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2023), 8621–8633.

[353] Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. arXiv:2311.09210. Retrieved from https://arxiv.org/abs/2311.09210

[354] Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. arXiv:2305.14002. Retrieved from https://arxiv.org/abs/2305.14002

[355] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS '21)*. Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.), 27263–27277. Retrieved from https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html

[356] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '19)*. Computer Vision Foundation/IEEE, 6720–6731. DOI: https://doi.org/10.1109/CVPR.2019.00688

[357] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. arXiv:2310.01779. Retrieved from https://arxiv.org/abs/2310.01779

[358] Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-Tuning: Teaching large language models to refuse unknown questions. arXiv:2311.09677. Retrieved from https://arxiv.org/abs/2311.09677

[359] Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics, 25–33. Retrieved from https://aclanthology.org/2021.humeval-1.3

[360] Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. 2023. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. arXiv:2311.01740. Retrieved from https://arxiv.org/abs/2311.01740

[361] Jiajun Zhang, Yang Zhao, Haoran Li, and Chengqing Zong. 2018. Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 3 (2018), 507–518.

[362] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*, Proceedings of Machine Learning Research, Vol. 119, PMLR, 11328–11339. Retrieved from http://proceedings.mlr.press/v119/zhang20ae.html

[363] Mingtian Zhang, Shawn Lan, Peter Hayes, and David Barber. 2024. Mafin: Enhancing black-box embeddings with model augmented fine-tuning. arXiv:2402.12177. Retrieved from https://arxiv.org/abs/2402.12177

[364] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. arXiv:2305.13534. Retrieved from https://arxiv.org/abs/2305.13534

[365] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. arXiv:2401.01286. Retrieved from https://arxiv.org/abs/2401.01286

[366] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. arXiv:2308.10792. Retrieved from https://arxiv.org/abs/2308.10792

[367] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating language model hallucination with interactive question-knowledge alignment. arXiv:2305.13669. Retrieved from https://arxiv.org/abs/2305.13669

[368] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. arXiv:2205.01068. Retrieved from https://arxiv.org/abs/2205.01068

[369] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. arXiv:2301.13848. Retrieved from https://arxiv.org/abs/2301.13848

[370] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 915–932. Retrieved from https://aclanthology.org/2023.emnlp-main.58

[371] Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. arXiv:2402.09267. Retrieved from https://arxiv.org/abs/2402.09267

[372] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv:2309.01219. Retrieved from https://arxiv.org/abs/2309.01219

[373] Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. arXiv:2403.04797.

[374] Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. arXiv:2402.16457. Retrieved from https://arxiv.org/abs/2402.16457

[375] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. arXiv:2402.08680. Retrieved from https://arxiv.org/abs/2402.08680

[376] Liang Zhao, Xiaocheng Feng, Xiachong Feng, Bing Qin, and Ting Liu. 2023. Length extrapolation of transformers: A survey from the perspective of position encoding. arXiv:2312.17044. Retrieved from https://arxiv.org/abs/2312.17044

[377] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*. Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, 5823–5840. DOI: https://doi.org/10.18653/v1/2023.acl-long.320

[378] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems* 42, 4 (2024), 1–60.

[379] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv:2303.18223. Retrieved from https://arxiv.org/abs/2303.18223

[380] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023. Knowing what LLMs do not know: A simple yet effective self-detection method. arXiv:2310.17918. Retrieved from https://arxiv.org/abs/2310.17918

[381] Danna Zheng, Mirella Lapata, and Jeff Z. Pan. 2024. Large language models as reliable knowledge bases? arXiv:2407.13578. Retrieved from https://arxiv.org/abs/2407.13578

[382] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does ChatGPT fall short in answering questions faithfully? arXiv:2304.10513. Retrieved from https://arxiv.org/abs/2304.10513

[383] Weihong Zhong, Mao Zheng, Duyu Tang, Xuan Luo, Heng Gong, Xiaocheng Feng, and Bing Qin. 2023. STOA-VLP: Spatial-temporal modeling of object and action for video-language pre-training. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence and 35th Conference on Innovative Applications of Artificial Intelligence and 30th Symposium on Educational Advances in Artificial Intelligence*, Vol. 37, 3715–3723. DOI: https://doi.org/10.1609/aaai.v37i3.25483

[384] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. arXiv:2305.11206. Retrieved from https://arxiv.org/abs/2305.11206

[385] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (ACL-IJCNLP '21)*. Association for Computational Linguistics, 1393–1404. DOI: https://doi.org/10.18653/v1/2021.findings-acl.120

[386] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Proceedings of the International Conference on Findings of the Association for Computational Linguistics (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 14544–14556. Retrieved from https://aclanthology.org/2023.findings-emnlp.968

[387] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. arXiv:2310.00754. Retrieved from https://arxiv.org/abs/2310.00754

[388] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592. Retrieved from https://arxiv.org/abs/2304.10592

[389] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. arXiv:2304.04675. Retrieved from https://arxiv.org/abs/2304.04675

[390] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. arXiv:2308.07107. Retrieved from https://arxiv.org/abs/2308.07107

[391] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2023. Fool your (vision and) language model with embarrassingly simple permutations. arXiv:2310.01651. Retrieved from https://arxiv.org/abs/2310.01651

Received 8 December 2023; revised 2 August 2024; accepted 24 September 2024