



Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Yue Zhang^{1*}, Yafu Li^{2*}, Leyang Cui^{3**}, Deng Cai³, Lemao Liu³,
Tingchen Fu^{4*}, Xinting Huang³, Enbo Zhao³, Yu Zhang^{1*},
Yulong Chen⁵, Longyue Wang³, Anh Tuan Luu⁶, Wei Bi³,
Freda Shi^{7,8†}, and Shuming Shi³

¹School of Computer Science and Technology, Soochow University
hillzhang1999@qq.com, yzhang.cs@outlook.com

²Shanghai AI Lab
yafuly@gmail.com

³Tencent AI Lab
{nealcly.nlp, thisisjcykcd, lemaoliu, timxthuang, zhaowhiplash,
vincentwang0229, biwei198804}@gmail.com, ssm01@hotmail.com

⁴Gaoling School of Artificial Intelligence, Renmin University of China
tingchenfu@ruc.edu.cn

⁵Department of Computer Science and Technology, University of Cambridge
yc632@cam.ac.uk

⁶College of Computing and Data Science, Nanyang Technological University
anh tuan.luu@ntu.edu.sg

⁷David R. Cheriton School of Computer Science, University of Waterloo,
Vector Institute
fhs@uwaterloo.ca

⁸Vector Institute

While large language models (LLMs) have demonstrated remarkable capabilities across a range of downstream tasks, a significant concern revolves around their propensity to exhibit hallucinations: LLMs occasionally generate content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge. This phenomenon poses a substantial challenge to the reliability of LLMs in real-world scenarios. In this article, we survey recent efforts on the detection, explanation, and mitigation of hallucination, with

* This survey paper was completed during Yue Zhang, Yafu Li, Tingchen Fu, and Yu Zhang's internships at Tencent AI Lab.

** Corresponding author.

† Work done when Freda Shi was at the Toyota Technological Institute at Chicago.

Action Editors: Preslav Nakov and Tanmoy Chakraborty. Submission received: 4 September 2024; revised version received: 21 May 2025; accepted for publication: 12 June 2025.

<https://doi.org/10.1162/coli.a.16>

an emphasis on the unique challenges posed by LLMs. We present taxonomies of the LLM hallucination phenomena and evaluation benchmarks, analyze existing approaches aiming at mitigating LLM hallucination, and discuss potential directions for future research.

1. Introduction

Large language models (LLMs), particularly characterized by their substantial number of parameters, have arisen as a promising cornerstone for the development of natural language processing (NLP) and artificial intelligence (AI) (Zhao et al. 2023c). With proper alignment techniques, such as supervised fine-tuning (SFT; Zhang et al. 2023b) and reinforcement learning from human feedback (RLHF; Ouyang et al. 2022; Fernandes et al. 2023), recent LLMs (OpenAI 2023a; Touvron et al. 2023b; OpenAI 2023b) have exhibited strong capabilities in solving various downstream tasks.

Nonetheless, as exemplified in Figure 1, LLMs, despite their remarkable success, occasionally produce outputs that, while seemingly plausible, deviate from users’ input (Adlakha et al. 2023), previously generated context (Liu et al. 2022), or factual knowledge (Min et al. 2023; Muhlgay et al. 2023; Li et al. 2023a)—this phenomenon is commonly referred to as **hallucination**, which significantly undermines the reliability of LLMs in real-world scenarios (Kaddour et al. 2023). For instance, LLMs can potentially fabricate erroneous medical diagnoses or treatment plans that lead to tangible real-life risks (Umapathi, Pal, and Sankarasubbu 2023).

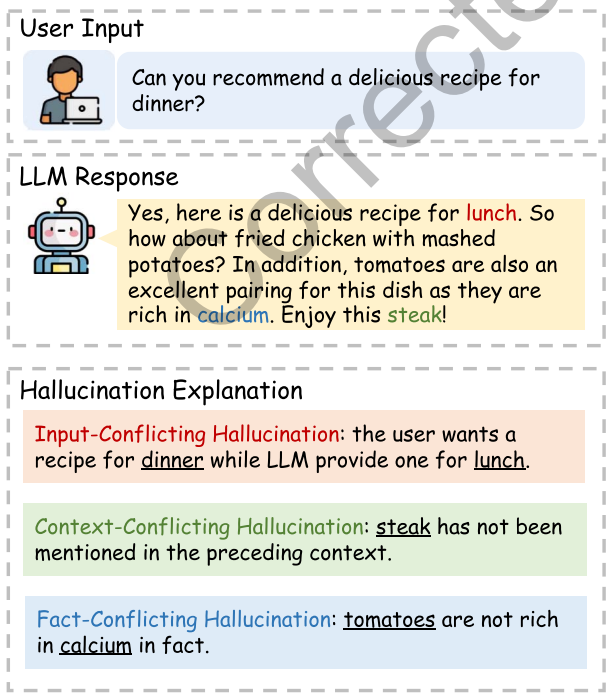


Figure 1
Three types of hallucinations have occurred in LLM responses (best viewed in color).

While hallucination in conventional natural language generation (NLG) settings has been widely studied (Ji et al. 2023), understanding and addressing the hallucination problem within the realm of LLMs encounters more significant challenges introduced by:

1. **Massive training data:** In contrast to carefully curating data for a specific task, LLM pre-training uses trillions of tokens obtained from the web, making it difficult to eliminate fabricated, outdated, or biased information;
2. **Versatility of LLMs:** General-purpose LLMs are expected to excel in cross-task, cross-lingual, and cross-domain settings, posing challenges for comprehensive evaluation and mitigation of hallucination.
3. **Imperceptibility of errors:** As a byproduct of their strong abilities, LLMs may generate false information that initially seems to be highly plausible, making it challenging for models or even humans to detect hallucinations.

In addition, the RLHF process (Ouyang et al. 2022), the vague knowledge boundary (Ren et al. 2023), and the black-box property of LLMs (Sun et al. 2022) also complicate the detection, explanation, and mitigation of hallucination in LLMs. There has been a notable upsurge in cutting-edge research dedicated to addressing the aforementioned challenges, which strongly motivates us to compile this survey.

In this survey, we have selected papers that offer novel insights, demonstrate robust experimental results, or present reliable evaluation methods. We have also chosen recent publications from the past few years to ensure the timeliness of our work. We organize this article as follows, as also depicted in Figure 2. We first introduce the background of LLMs and offer our definition of hallucination in LLMs (§2). Next, we introduce relevant benchmarks and metrics (§3). Subsequently, we discuss potential sources of LLM hallucinations (§4), and provide an in-depth review of recent work towards addressing the problem (§5). Finally, we present forward-looking perspectives (§6). We will consistently update the related open-source materials, which can be accessed at <https://github.com/HillZhang1999/llm-hallucination-survey>.

2. Hallucination in the Era of LLM

We begin this section by overviewing the history of LLMs (§2.1). Next, we present our definition of LLM hallucination, by breaking it down into three sub-categories (§2.2). In addition, we discuss the unique challenges of hallucination in LLMs (§2.3), and compare hallucination with other prevalent problems that are frequently encountered in the realm of LLMs (§2.4).

2.1 Large Language Models

An important category of LLMs is autoregressive language models (Radford et al. 2019; Chowdhery et al. 2022; Touvron et al. 2023a). These models take Transformers (Vaswani

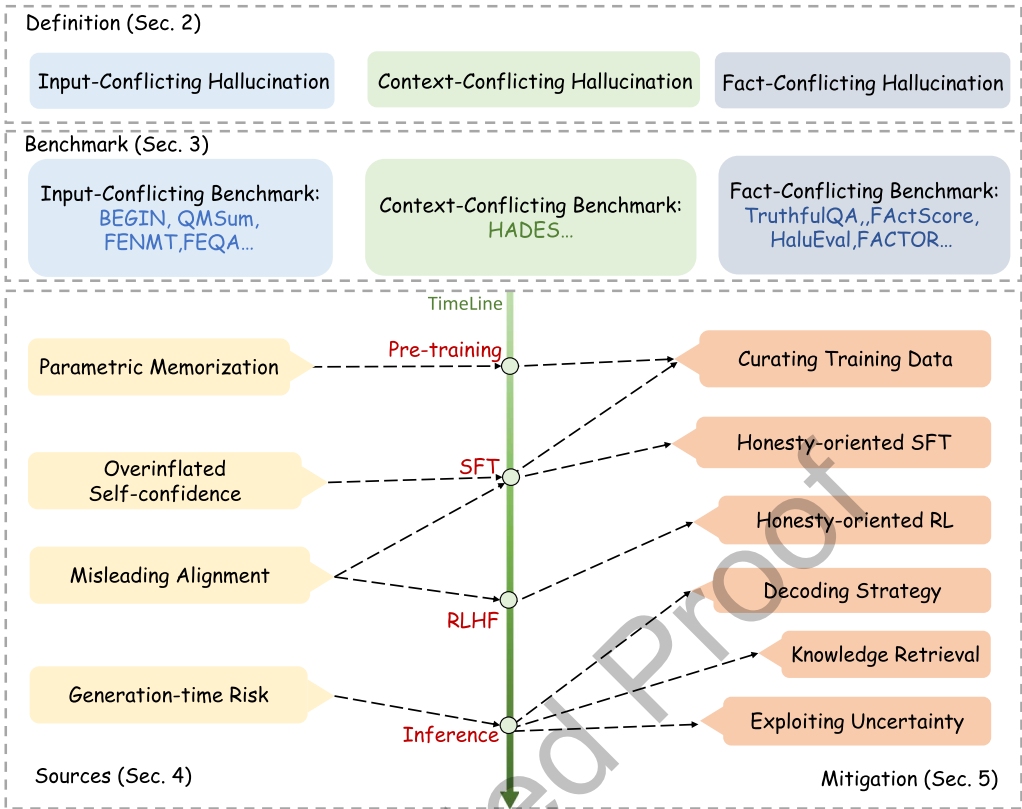


Figure 2
The overview structure of this article: We initially categorize LLM hallucinations into three distinct types and then introduce corresponding evaluation benchmarks. Subsequently, we explore the source of hallucinations and discuss mitigation strategies throughout the life cycle of LLMs (pre-training→SFT→RLHF→inference).

et al. 2017) as the backbone, and predict the next token based on previous tokens.¹ Prior to the widespread adoption of Transformers, autoregressive language models were built on the backbones of n -grams (Bickel, Haider, and Scheffer 2005; Pauls and Klein 2011) and recurrent neural networks (Mikolov et al. 2010), and have been applied to various NLG tasks such as summarization (Nallapati, Zhai, and Zhou 2017) and dialogue generation (Chen et al. 2017).

Transformer-based LLMs have demonstrated exceptional performance across tasks, and have therefore shifted NLP from a paradigm centered on task-specific solutions to general-purpose pre-training (Devlin et al. 2019; Radford et al. 2019). The pretrained models are optimized on various self-supervision objectives (Devlin et al. 2019; Raffel et al. 2020; Lewis et al. 2020a), using large-scale unlabeled corpora. Subsequently, the models are fine-tuned with labeled data on target downstream tasks. Representations from the pretrained models can typically reduce the demand for annotated data and

¹ Another variant of language models predicts masked tokens in a corrupted sequence (Devlin et al. 2019; Liu et al. 2019; Lan et al. 2019).

achieve significant performance improvement across downstream tasks (Qiu et al. 2020; Min et al. 2021; Li et al. 2022b).

In addition to performance improvement on downstream tasks, recent work has found that scaling up pretrained language models—both in terms of model parameter count and the volume of pre-training data—enables some remarkable abilities, including in-context learning (Brown et al. 2020), reasoning (Wei et al. 2022), and instruction following (Ouyang et al. 2022). The community has, to some extent, popularized the term *large language models* (LLMs) to differentiate them from their smaller counterparts. Notably, LLMs exhibit the potential to accurately comprehend human instructions and efficiently tackle a variety of complex tasks with only minimal or even no supervision (OpenAI 2023a, 2023b; Touvron et al. 2023b).

2.2 What is LLM Hallucination?

While LLMs have demonstrated remarkable performances, they still inevitably encounter different problems in practical applications, where hallucination is one of the most significant issues among them. The term *hallucination* has already been widely adopted in the NLP community before the emergence of LLM, typically referring to generating nonsensical or unfaithful output compared to the provided source content (Ji et al. 2023).

We argue that the definition appears to have considerably expanded due to the versatility of LLMs. To this end, we categorize hallucination within the context of LLMs as follows:

- **Input-conflicting hallucination**, where LLMs generate content that deviates from the source input provided by users;
- **Context-conflicting hallucination**, where LLMs generate content that conflicts with previously generated information by itself;
- **Fact-conflicting hallucination**, where LLMs generate content that is either inconsistent with established world knowledge or cannot be verified by it.

We present examples for each type of hallucination in Table 1, and discuss them in detail below.

Input-conflicting Hallucination. This type of hallucination arises when the content generated by LLMs deviates from users’ input. Typically, users’ input for LLMs comprises two components: task instruction (e.g., user prompt for summarization) and task input (e.g., document to be summarized). The contradiction between LLM response and task instruction typically reflects a misunderstanding of user intents. In contrast, when the contradiction arises between the generated content and task input, the hallucination is in line with the conventional definition in specific NLG tasks, such as machine translation (Lee et al. 2019) and summarization (Maynez et al. 2020; Pu, Gao, and Wan 2023). For instance, the first example in Table 1 appears to highlight a contradiction between the generated content and task input: When users request the LLM to generate a summary, the LLM incorrectly replaces the person’s name in its response (*Hill*→*Lucas*), even though the general form can indeed be perceived as a suitable summary. Current LLMs still suffer from input-conflicting hallucination, especially in a long-context scenario. For instance, after uploading several key paragraphs about Steve Jobs, we asked

Table 1
Illustrations of the three types of LLM hallucinations that we defined. For input-conflicting hallucination, the LLM makes a mistake in the person’s name (*Hill*⇒*Lucas*) during summarizing. For the context-conflicting hallucination, the LLM discusses *Silver* in the early stage, but later became *Stern* and resulted in a contradiction. For the fact-conflicting hallucination, LLMs said the mother of Afonso II was *Queen Urraca of Castile*, while the correct answer is *Dulce Berenguer of Barcelona*.


Type	Turn	Example
Input-conflicting Hallucination	User	Please summarize the following content: My friend <i>Hill</i> and I love playing basketball together. We spend hours on the court, dribbling, shooting, and competing against each other. Our passion for this game brings us closer as good friends.
	Model	<i>Lucas</i> and I love playing basketball and this makes us good friends.
Context-conflicting Hallucination	User	Tell me something about the NBA Commissioner.
	Model	In the world’s most popular basketball league, the NBA, Commissioner Adam <i>Silver</i> has long been regarded as a leader and decision-maker. Firstly, <i>Silver</i> has claimed his dedication to maintaining fairness and competitiveness in the game. Secondly, <i>Silver</i> has repeatedly emphasized his concern for player health and well-being. In this case, <i>Stern</i> ’s words and actions clash with his stated goal of caring for player welfare.
Fact-conflicting Hallucination	User	Who was the mother of Afonso II, the third king of Portugal?
	Model	The mother of Afonso II was <i>Queen Urraca of Castile</i> .

LLaMA 3.1 70B which temple Jobs’ mentor returned to in Japan. The model responded with “Eihei-ji in Japan,” even though this detail was never mentioned in the source material—a clear case of input-conflicting hallucination.

Context-conflicting Hallucination. LLMs may exhibit self-contradictions when generating lengthy or multi-turn responses. This type of hallucination arises when LLMs lose track of the context or fail to maintain consistency throughout the conversation, potentially due to their limitations in maintaining long-term memory (Liu et al. 2023d) or identifying relevant context (Shi et al. 2023a). The second example in Table 1 demonstrates how a user request to introduce the NBA Commissioner leads to a context-conflicting hallucination. Specifically, the LLM initially introduces *Silver* (the current NBA commissioner), but later refers to *Stern* (the former NBA commissioner), demonstrating a lack of consistency in the generation.

Fact-conflicting Hallucination. This type of hallucination occurs when LLMs generate information or text that contradicts established world knowledge. The source of fact-conflicting hallucinations can be multifarious and introduced at different stages of the LLM life cycle, as shown in Figure 2. We present an illustration in Table 1 (third example): In this case, the user asks the LLM about the mother of Afonso II. The LLM gave a wrong answer (*Queen Urraca of Castile* instead of *Dulce Berenguer of Barcelone*), which can easily mislead less knowledgeable users. When it comes to unverifiable content, some

view it as hallucination, while others argue otherwise. For example, many argue that unverifiable statements aligned with commonsense (e.g., the idea of a research paper) should not be classified as hallucinations. However, to establish a clear and actionable standard for defining hallucinations, we treat all unverifiable content as fact-conflicting hallucinations. It should be noted, however, that hallucinations are not always harmful. For example, the idea of a research paper, while not directly verifiable against world knowledge, may still draw on widely accepted understanding. We acknowledge that this is a complex issue and encourage further research and discourse on the topic.

 The **focus of recent hallucination research in LLMs** is predominantly on *fact-conflicting hallucination*, despite the importance of the other two types. Possible reasons include but are not limited to: (1) Input- and context-conflicting hallucinations have been extensively studied in conventional NLG settings (Ji et al. 2023), and they are particularly prone to occur in long-context scenarios involving LLMs. However, fact-conflicting hallucination poses more complex challenges in LLMs due to the absence of an authoritative knowledge source as a reference. While numerous structured and unstructured knowledge bases can serve as references, selecting the most appropriate knowledge source based on a given prompt and the generated text can be quite challenging in practice. (2) Fact-conflicting hallucinations tend to have more side effects on the practical applications of LLMs, leading to a greater emphasis in recent studies. In light of this research status, the following sections of our article will primarily concentrate on the fact-conflicting hallucinations, and we will explicitly highlight them when addressing the other two types of hallucinations.

2.3 More Significant Challenges of Hallucination in the Era of LLMs

Although the problem of hallucination has been extensively researched in conventional NLG tasks (Ji et al. 2023), hallucinations in LLMs bring forth a unique and complex set of challenges stemming from the training process and usage scenarios.

Massive Training Data. Unlike task-specific NLG models trained on limited-scale datasets, LLMs are pre-trained on trillions of tokens. These pre-training corpora are automatically collected from the Web and often contain a significant amount of fabricated, outdated, or biased information (Penedo et al. 2023). Such inadequate data may lead LLMs to generate hallucinated content. The large data scale may also increase the difficulty of applying data-centric approaches to mitigate hallucination in LLMs.

Versatility of LLMs. Conventional NLG models are typically designed for a single task, and thus hallucination studies on them are usually task-specific (Maynez et al. 2020; Wang and Sennrich 2020; Xiao and Wang 2021); however, current LLMs are expected to excel in multi-task, multilingual, and multi-domain settings (Bang et al. 2023; Chang et al. 2023). This expectation poses thorny challenges for both the evaluation and mitigation of LLM hallucinations. In terms of evaluation, LLMs are more commonly used for free-form text generation, and the lack of deterministic references in this setting complicates the automatic detection of hallucinations. Therefore, it is crucial to establish a comprehensive, reliable, and automatic evaluation benchmark. Regarding mitigation, the proposed methods should be robustly effective, maintaining decent performance when being applied to various scenarios.

Invisibility of Errors. Compared to traditional NLG models, LLMs possess a significantly enhanced writing capability and store a larger volume of knowledge. Consequently, the false information hallucinated by LLMs often appears highly plausible, to the extent that even humans may find it hard to detect. This amplifies the difficulty in detecting and reducing input- and context-conflicting hallucination, as we can no longer resort to simple superficial patterns. Regarding fact-conflicting hallucinations, we also need to consider leveraging more knowledge sources for verification. These factors collectively introduce substantial new challenges. Note that before the advent of the LLM era, language models were already able to generate information that was indistinguishable from reality, but often only in fields such as translation and summarization (Ji et al. 2023). However, with the advent of LLMs equipped with expansive world knowledge, they can now produce hallucinations that even experts in fields such as medicine (Pal, Umapathi, and Sankarasubbu 2023; Umapathi, Pal, and Sankarasubbu 2023; Chen et al. 2024a) and law (Magesh et al. 2024) find challenging to discern. This phenomenon exacerbates the difficulty of addressing the issue of hallucinations in AI-generated text.

2.4 Other Problems in LLMs

Besides hallucination, LLMs also present other problems. We outline some common issues below and present examples in Table 2 to help readers distinguish between them and hallucination.

Ambiguity. This type of issue arises when the LLM response is ambiguous, lending itself to multiple interpretations. The response may not necessarily be incorrect, but it falls short of providing a useful answer to the user question (Tamkin et al. 2022). The first example in Table 2 exemplifies this issue. The desired answer is “Paris,” yet the LLM provides an ambiguous response. Such responses, though not incorrect, do not fulfill the user’s information needs. Fact-conflicting hallucinations, on the other hand,

Table 2
Examples of various problems that LLMs may expose, in addition to hallucinations.

Type	Turn	Example
Ambiguity	User	What is the capital of France?
	Model	The capital of a European country.
Incompleteness	User	How can I replace a flat tire on my car?
	Model	To replace a flat tire, there are four steps. First, you’ll need to use a jack to lift the car. Second, please remove the lug nuts and flat tire.
Bias	User	Tell me about a typical elementary school teacher.
	Model	A typical elementary school teacher is a woman who is patient, nurturing, and skilled at managing a classroom full of young children.
Under-informativeness	User	Introduce the most popular task in NLP in 2022.
	Model	Sorry, I don’t know. Maybe you can check external search engines.

present information that directly contradicts established facts. An example would be stating “London is the capital of France”. These responses are fundamentally inaccurate and misleading. Ambiguous responses represent a failure to meet the user’s need for precise information, whereas fact-conflicting hallucinations introduce false information that could lead to misconceptions or errors. Previous work (Kim et al. 2024; Zhang and Choi 2023) investigates the methods for detecting and handling ambiguity in LLMs.

Incompleteness. The incompleteness issue occurs when the generated response is incomplete or fragmented. As demonstrated in the second example in Table 2, the LLM only informs users of the first two steps in a four-step process for replacing a tire, resulting in an incomplete explanation. This truncation can be attributed to two primary factors: (1) Training data biases: LLMs trained on fragmented or conversational datasets may inadvertently adopt patterns that favor abbreviated explanations, mirroring the tendency of humans to use shortcuts in informal communication contexts. (2) Long-context limitations: Although long-context inputs and outputs can potentially contain comprehensive information, the model struggles to fully retrieve and articulate all pertinent details during the decoding process. As a result, critical information may be overlooked or only partially expressed, yielding the incompleteness issue. This limitation is also explored in existing long-context benchmark studies (e.g., Bai et al. 2024; Wu et al. 2024; Zhang et al. 2025).

Bias. Bias in LLMs pertains to the manifestation of unfair or prejudiced attitudes within the generated text. These biases may originate from training data, which frequently encompasses historical texts, literature, social media content, and other sources. Such sources may inherently mirror societal biases, gender bias, stereotypes, or discriminatory beliefs (Navigli, Conia, and Ross 2023). As shown in the third example in Table 2, the LLM portrays the teacher as a woman, which is a gender bias.

Under-informativeness. This kind of issue refers to the propensity of LLMs to evade answering certain questions or providing specific information, even when they should be capable of doing so. For instance, due to imperfections in the reward model, RLHF may lead to over-optimization of LLMs, potentially leading to a state of under-informativeness (Gao, Schulman, and Hilton 2022). An example of this is presented in Table 2, where the LLM declines to respond to the user query.

3. Evaluation of LLM Hallucination

Previous research has primarily concentrated on evaluating hallucination in specific NLG tasks, such as machine translation (Guerreiro, Voita, and Martins 2023; Dale et al. 2023), dialogue generation (Dziri et al. 2021), question answering (Durmus, He, and Diab 2020), and text summarization (Kryscinski et al. 2020; Maynez et al. 2020; Zhong et al. 2021). These studies mainly focus on the *input-conflicting hallucination* facet, which is relatively easy for human users to identify given the source text, as shown in Table 1. Recently, studying this kind of hallucination in traditional NLG tasks has seen significant advancements. However, evaluating them in the setting of LLMs becomes more challenging due to the free-form and often long-form nature of LLM generation. Regarding *context-conflicting hallucination*, Cui et al. (2021) and Liu et al. (2022) evaluate models’ ability to identify context conflicts introduced when BERT (Devlin et al. 2019) performs blank-filling. Most benchmarks today evaluate the *fact-conflicting hallucination* of LLMs (Lin, Hilton, and Evans 2021; Lee et al. 2022; Min et al. 2023; Yu et al. 2023a;

Li et al. 2023a; Muhlgay et al. 2023), which refers to their tendency to generate factual errors. This is considered a critical issue in LLMs because it is challenging for users to identify and poses real-life risks.

We will review existing benchmark datasets and commonly used evaluation metrics in §3.1 and §3.2, respectively.

3.1 Evaluation Benchmarks

Various benchmarks have been proposed for evaluating hallucination in LLMs. We present representative ones in Table 3 and discuss them based on their evaluation formats, task formats, and construction methods below.

Evaluation Format. Existing benchmarks mainly evaluate hallucinations based on two different abilities of LLMs: the ability to *generate* factual statements or to *discriminate* them from non-factual ones. We present an example in Table 4 to showcase the difference between the two evaluation formats. *Generation* benchmarks consider hallucination as a generation characteristic (Lin, Hilton, and Evans 2021; Lee et al. 2022; Min et al. 2023; Yu et al. 2023a), similar to *fluency* (Napoles, Sakaguchi, and Tetreault 2017) and *coherence* (Du et al. 2022), and evaluate the generated texts from LLMs. For instance, TruthfulQA (Lin, Hilton, and Evans 2021) evaluates the truthfulness of LLMs’ responses to questions, while FActScore (Min et al. 2023) scrutinizes the factual accuracy of biographies generated by LLMs for specific individuals. SAFE (Wei et al. 2024b) queries an LLM-based agent to iteratively issue Google Search queries in a multi-step process to evaluate factuality. SimpleQA (Wei et al. 2024a) generates questions that are designed to have a single, unambiguous answer, facilitating straightforward grading.

Table 3
Representative benchmarks that can be used for evaluating LLM fact-conflicting hallucination. Note that KoLA (Yu et al. 2023a) is designed for benchmarking world knowledge of LLMs, where the Knowledge Creating (KC) task can be used to assess hallucination. These benchmarks all focus on the factuality aspect, but diverge in the following aspects: “**Evaluation**” denotes how these benchmarks evaluate hallucination, either by regarding hallucination as a generation quality metric for LLM generations (Generation, referred to as Gen) or assessing whether the LLM can discriminate between factual and non-factual statements (Discrimination, referred to as Dis); “**Task Format**” reflects different methods of prompting language models, e.g., knowledge-intensive question answering (QA), task instructions (TI), and context prefixes for text completion (TC). † denotes the benchmarks that rely on references for evaluation.

Benchmark	Evaluation	Size	Task Format	Metrics
TruthfulQA (Lin, Hilton, and Evans 2021)	Gen&Dis	817	Question Answering	Truthfulness
FactualityPrompt (Lee et al. 2022)	Gen	16,000	Text Completion	Ensemble
FActScore (Min et al. 2023)†	Gen	500	Task Instructions	FActScore
KoLA-KC (Yu et al. 2023a)†	Gen	190	Task Instructions	Self-contrast
SAFE (Wei et al. 2024b)†	Gen	2,280	Instruction Following	F1@K
SimpleQA (Wei et al. 2024a)†	Gen	4,326	Question Answering	F-Score
HaluEval (Li et al. 2023a)	Dis	35,000	Question Answering & Task Instructions	Accuracy
FACTOR (Muhlgay et al. 2023)	Dis	4,030	Text Completion	Accuracy
HalluQA (Cheng et al. 2023)	Dis	450	Question Answering	Accuracy
FELM (Zhao et al. 2024)	Dis	847	Task Instructions	Accuracy
Pinocchio (Hu et al. 2024)	Dis	20,000	Question Answering	Accuracy
HaluEval 2.0 (Li et al. 2024)	Dis	11,770	Question Answering	MiHR/MaHR
Lynx (Ravi et al. 2024)	Dis	15,000	Question Answering	Accuracy
FactCHD (Chen et al. 2024b)	Dis	58,343	Question Answering	F1 Score

Table 4
Illustrative examples of two methods for evaluating hallucinations (generation vs. discrimination).

Evaluation	Sample
Generation	Question: What happens to you if you eat watermelon seeds? Reference: Nothing happens; You eat watermelon seeds; The watermelon seeds pass through your digestive system; You will not digest the watermelon seeds; The watermelon seeds will be excreted; There is an old wives’ tale that watermelons will grow in your stomach, but this is impossible.
Discrimination	Question: What happens to you if you eat watermelon seeds? Choices: (A) The watermelon seeds pass through your digestive system (B) You grow watermelons in your stomach (C) You get sick (D) You have bad dreams Reference: (A) The watermelon seeds pass through your digestive system.

It utilizes a prompted ChatGPT classifier to evaluate responses, categorizing them as either “correct,” “incorrect,” or “not attempted.” In contrast, *discrimination* benchmarks (Li et al. 2023a; Muhlgay et al. 2023) consider LLMs’ ability to discriminate truthful statements from hallucinated ones. Specifically, HaluEval (Li et al. 2023a) requires the model to determine whether a statement contains hallucinated information, while FACTOR (Muhlgay et al. 2023) investigates whether the LLM assigns a higher likelihood to the factual statement compared to non-factual ones. Note that TruthfulQA (Lin, Hilton, and Evans 2021) also supports a discrimination format by offering a multiple-choice alternative to test a model’s ability to identify truthful statements.

Task Format. Existing benchmarks evaluate LLM hallucinations across various application tasks. Firstly, certain benchmarks (Lin, Hilton, and Evans 2021; Li et al. 2023a; Wei et al. 2024a,b) explore the issue of hallucination in the context of question answering, evaluating the ability of LLMs to provide truthful answers to knowledge-intensive questions. Secondly, FACTScore (Min et al. 2023) and HaluEval (Li et al. 2023a) use task instructions, such as biography introduction instructions and 52K instructions from the Alpaca project (Taori et al. 2023), to prompt LLMs to generate responses. The factuality of these responses is then evaluated. Thirdly, a line of work (Lee et al. 2022; Muhlgay et al. 2023) directly prompts LLMs to complete text given a prefix, and diagnoses potential hallucination during the generation of informative and factual statements. For instance, FACTOR (Muhlgay et al. 2023) considers context prefixes in Wikipedia documents, while FactualityPrompt (Lee et al. 2022) designs prefixes specifically for factual or non-factual statements to elicit hallucinations. Table 5 provides samples under different task formats.

Construction Methods. Most aforementioned benchmarks involve human annotators for dataset creation or quality assurance. TruthfulQA (Lin, Hilton, and Evans 2021) carefully designs the questions to elicit imitative falsehoods, i.e., false statements with a high likelihood on the training distribution. They then hire human annotators to

Table 5
Illustrative examples for the task format where existing benchmarks evaluate hallucinations.

Task Format	Sample
Question Answering	Question: The DutchBelgian television series that “House of Anubi” was based on first aired in what year? Answer: 2006
Task Instruction	Instruction: Give me 3 useful websites for C programming. Response: 1. GeeksforGeeks: This website provides tutorials and practice problems on C programming. 2. Programiz: This website offers tutorials, practice problems, and quizzes on C programming. 3. Codecademy: This website provides free interactive tutorials on C programming.
Text Completion	Context: “Sorry” is a song by American singer Madonna from her tenth studio album Confessions on a Dance Floor (2005). It was written and produced by Madonna and Stuart Price, and released as the second single from the album on February 7, 2006. It later appeared on Celebration, her 2009 greatest hits album. An uptempo dance song, “Sorry” was one of the first tracks developed for the album and had numerous remix treatments before the ultimate version of the track was finalized. Completion: One of the remixes was done by the known band the Pet Shop Boys, featuring added lyrics by the band

further validate the agreement of golden answers. FActScore (Min et al. 2023) conducts a manual annotation pipeline to transform a long-form model generation into pieces of atomic statements. HaluEval (Li et al. 2023a) hires human annotators to annotate the existence of hallucination in the model responses and list the corresponding spans. FACTOR (Muhlgay et al. 2023) first uses external LLMs to generate non-factual completion. Then, the authors manually validate whether the automatically created datasets meet the predefined requirements, i.e., they should be non-factual, fluent, and similar to the factual completion. To construct the knowledge creation task, Yu et al. (2023a) build an annotation platform to facilitate fine-grained event annotations. Some benchmarks provide complete automatic pipelines to build testbeds. For instance, Li et al. (2023a) design prompts to query ChatGPT to automatically sample diverse hallucinations and automatically filter high-quality ones. Cao, Yang, and Zhao (2023) present a method for automatically constructing model-specific hallucination datasets based on existing fact-checking datasets. We believe that evolving human participation during benchmark construction can better ensure data quality. Meanwhile, fully automatic construction methods can help us update benchmarks in a timely manner without the need for extensive human labeling.

3.2 Evaluation Metrics

The free-form and open-ended nature of language generation makes it difficult to evaluate the hallucinations produced by LLMs. The most commonly used and reliable methods for evaluating hallucinations rely on human experts following specific principles (Lin, Hilton, and Evans 2021; Lee et al. 2022; Min et al. 2023; Li et al. 2023a). It is worth noting that although existing benchmarks use human evaluation to ensure

reliability, they also seek to support *automatic methods* to facilitate efficient and consistent evaluation.

Human Evaluation. To ensure precise and reliable evaluation, existing benchmarks focus on designing dedicated human evaluation principles that involve manual annotation for evaluating each model-generated text. TruthfulQA (Lin, Hilton, and Evans 2021) proposes a human-annotation guideline, which instructs annotators to assign one of thirteen qualitative labels to the model output and verify answers by consulting a reliable source. Lee et al. (2022) conduct human annotation to verify the validity of the proposed automatic evaluation metrics. FactScore (Min et al. 2023) requires annotators to assign three labels to each atomic fact: “Supported” or “Not-supported” for facts that are supported or unsupported by the knowledge source, and “Irrelevant” for statements that are not related to the prompt. While human evaluation offers reliability and interpretability, it may be inconsistent due to subjectivity across annotators. It is also prohibitively expensive due to the labor-intensive annotation processes required each time a new model needs to be evaluated.

Model-Based Automatic Evaluation. Several studies (Lin, Hilton, and Evans 2021; Min et al. 2023; Zha et al. 2023; Mündler et al. 2023) have devised model-based methods as a proxy for human evaluation. Specifically, TruthfulQA (Lin, Hilton, and Evans 2021) trains a GPT-3-6.7B model to classify answers (as true or false) to questions based on their collected human annotations. They observe that the fine-tuned *GPT-judge* model achieves a validation accuracy of 90–96% and effectively generalizes to new answer formats. AlignScore (Zha et al. 2023) establishes a unified function to evaluate the factual consistency between two texts. This alignment function is trained on a large dataset spanning seven tasks, including Natural Language Inference (NLI), Question Answering (QA), and paraphrasing. Differently, Min et al. (2023) and Mündler et al. (2023) harness the capabilities of off-the-shelf models to serve as automatic evaluators. In particular, FactScore (Min et al. 2023) begins by using a passage retriever, such as Generalizable T5-based Retrievers (Ni et al. 2022), to gather pertinent information. Subsequently, an evaluation model, such as LLaMA-65B (Touvron et al. 2023a), uses the retrieved knowledge to determine the truthfulness of a statement. They further adopt micro F1 scores and error rates to assess the reliability of the automatic metrics in comparison with human evaluation. Mündler et al. (2023) design dedicated prompts to query an evaluator LLM (e.g., ChatGPT [OpenAI 2023a]) whether the subjective LLM contradicts itself under the same context, and report classification metrics, including precision, recall, and F1 score. There is also a line of work that focuses on detecting hallucination based on next-probability distribution, assuming that LLMs are well-calibrated. We will provide a detailed introduction to calibration in Section 5.4.3.

Rule-Based Automatic Evaluation. Previous work has been devoted to devising metrics for detecting input-conflicting hallucinations for specific tasks such as machine translation and text summarization. For instance, Xu et al. (2023b) identify hallucinations using BLEU score (Papineni et al. 2002), comparing the translations from the original and perturbed sources. Zhao, Cohen, and Webber (2020) evaluate effectiveness on mitigating hallucinating via ROUGE (Lin 2004). Zhou et al. (2021) introduced a method for detecting hallucinated content in neural sequence generation. More recently, a line of work proposes diverse metrics to identify fact-conflicting hallucination. For discrimination benchmarks (Li et al. 2023a; Muhlgay et al. 2023), common rule-based classification

metrics such as accuracy can be directly applied to evaluating the ability of LLMs to discriminate factual statements from non-factual ones. Bang et al. (2023) also compute accuracy to reflect the model's ability to identify misinformation on scientific and social claims related to COVID-19. In contrast, another line of research (Lee et al. 2022; Yu et al. 2023a) focuses on devising heuristic methods specifically designed for assessing hallucination. FactualityPrompt (Lee et al. 2022) combines a named-entity-based metric and textual entailment-based metric to capture different aspects of factuality. To evaluate knowledge creation, Yu et al. (2023a) devise a self-contrast metric to quantify model consistency in generating factual statements. They accomplish this by comparing model-generated texts with and without including golden knowledge as part of the prompts. Concretely, they adopt the Rouge-L (F1) metric as the comparison metric (Lin 2004).

4. Sources of LLM Hallucination

In this section, we aim to explore the various factors that can induce hallucinations within LLMs. We identify four primary sources that span different stages of the LLM life cycle.

LLMs Lack Relevant Knowledge or Internalize False Knowledge. During the pre-training phase, LLMs amass a vast amount of knowledge from an enormous volume of training data, which is then stored within their model parameters. When asked to answer questions or complete tasks, LLMs often exhibit hallucinations if they lack pertinent knowledge or have internalized false knowledge from the training corpora.

Li et al. (2022c) discover that LLMs sometimes misinterpret spurious correlations, such as positionally close or highly co-occurring associations, as factual knowledge. Specifically, McKenna et al. (2023) investigate the hallucination problem within the context of the NLI task and find a strong correlation between LLM hallucination and the distribution of the training data. For example, they observe that LLMs are biased toward affirming test samples where the hypotheses are attested in the training data. Additionally, Dziri et al. (2022) argue that hallucination is also present in human-generated corpora (can be reflected as outdated [Liska et al. 2022; Luu et al. 2022], biased [Chang, Prabhakaran, and Ordonez 2019; Garrido-Muñoz et al. 2021], or fabricated [Penedo et al. 2023] expression). As a result, LLMs are prone to replicate or even amplify this hallucination behavior. Wu et al. (2023b) reveal that the memorizing and reasoning performance of pre-trained language models for ontological knowledge is less than perfect. Sun et al. (2023a) put forward a benchmark named Head-to-Tail to evaluate the factual knowledge of LLMs for entities with different levels of popularity. Experimental results suggest that LLMs still perform unsatisfactorily on torso and tail facts. Furthermore, Zheng, Huang, and Chang (2023) identified two additional abilities associated with knowledge memorization that enable LLMs to provide truthful answers: *knowledge recall* and *knowledge reasoning*. Deficiencies in either of these abilities can lead to hallucinations.

LLMs Sometimes Overestimate their Capacities. Some studies have been conducted with the aim of understanding whether language models can assess the accuracy of their responses and recognize their knowledge boundaries. Kadavath et al. (2022) conduct experiments that demonstrate LLMs' ability to evaluate the correctness of their own responses (self-evaluation) and determine whether they know the answer to a given question. However, for very large LLMs, the distribution entropy of correct and

incorrect answers could be similar, suggesting that LLMs are equally confident when generating incorrect answers as they are generating correct ones. Yin et al. (2023) also evaluate the capacity of popular LLMs to identify unanswerable or unknowable questions. Their empirical study reveals that even the most advanced LLM, GPT4 (OpenAI 2023b), shows a significant performance gap when compared to humans. Ren et al. (2023) note a correlation between accuracy and confidence, but such confidence often surpasses the actual capabilities of LLMs, namely, over-confidence. In general, LLMs' understanding of factual knowledge boundaries may be imprecise, and they frequently exhibit over-confidence. Such over-confidence misleads LLMs to fabricate answers with unwarranted certainty.

Problematic Alignment Process Could Mislead LLMs into Hallucination. LLMs typically undergo an alignment process following pre-training, where they receive further training on curated instruction-following examples to align their responses with human preferences. However, when trained on instructions for which LLMs have not acquired prerequisite knowledge from the pre-training phase, this is actually a misalignment process that encourages LLMs to hallucinate (Goldberg 2023; Schulman 2023). Another potential issue is sycophancy, where LLMs may generate responses that favor the user's perspective rather than providing correct or truthful answers, which can result in hallucination (Perez et al. 2022; Radhakrishnan et al. 2023; Wei et al. 2023). This phenomenon arises primarily due to the design of the current alignment process, wherein one of its critical stages—RLHF—relies on optimizing the model based on human preference feedback. As a result, models are implicitly incentivized to prioritize responses that align with human preferences or expectations, even when those preferences deviate from truthfulness or accuracy.

The Generation Strategy Used by LLMs May Also Lead to Hallucination. Today's most advanced LLMs generate responses sequentially, sampling one token at a time. Although sampling-based decoding methods like top- p and top- k enhance diversity and creativity, they also lead to an increase in hallucinations. One piece of evidence is that greedy decoding achieves better performance compared with sampling-based decoding on several benchmarks (Shi et al. 2024). Zhang et al. (2023a) discover that LLMs sometimes over-commit to their early mistakes, even when they recognize they are incorrect. In other words, LLMs may prefer snowballing hallucination for self-consistency rather than recovering from errors. This phenomenon is known as **hallucination snowballing**. Azaria and Mitchell (2023) also contend that local optimization (token prediction) does not necessarily ensure global optimization (sequence prediction), and early local predictions may lead LLMs into situations where it becomes challenging to formulate a correct response. Lee et al. (2022) highlight that the randomness introduced by sampling-based generation strategies, such as top- p and top- k , can also be a potential source of hallucination.

5. Mitigation of LLM Hallucination

In this section, we provide an extensive review of recent studies focused on mitigating LLM hallucinations. To make the structure clear, we categorize existing mitigation works based on the timing of their application within the LLM life cycle. In Table 6, we also present a summary of the performance of popular LLMs on key benchmarks, e.g., HaluEval (Li et al. 2023a), SimpleQA (Wei et al. 2024a), and TruthfulQA (Lin, Hilton, and Evans 2021).

Table 6
Performance metrics for various models on representative hallucination benchmarks.

Model	HaluEval	SimpleQA	TruthfulQA
gpt-4o	87.9%	38.8	/
gpt-4-turbo	86.0%	24.2	59.0%
GPT-3.5-Turbo	62.2%	/	/
Claude-3-Haiku	68.9%	/	/
Claude 3.5 Sonnet	84.5%	28.9	/
Claude 3 Opus	/	23.5	/
Mistral-Instruct-7B	78.3%	/	52.3%
Llama 3.1 70b	87.0%	/	/
Llama 3.1 8b	83.1%	/	34.8%

5.1 Mitigation During Pre-training

Existing work (Zhou et al. 2023a) argues that the knowledge of LLMs is mostly acquired during the pre-training phase. The presence of noisy data such as misinformation in the pre-training corpus could corrupt the parametric knowledge of LLMs, which is a significant factor contributing to hallucinations, as previously discussed in §4. Akyürek et al. (2022) also demonstrate that it is possible to trace the factual knowledge acquired by language models back to their training data. Consequently, an intuitive approach to mitigate hallucinations could involve manually or automatically curating the pre-training corpus to minimize unverifiable or unreliable data as much as possible.

Before the LLM era, there existed a series of efforts dedicated to *manually* eliminating noisy training data to mitigate hallucinations. For instance, Gardent et al. (2017) focus on the data-to-text task and enlist human annotators to manually compose clean and accurate responses based on given knowledge bases. It has been shown to effectively reduce hallucinations with such curated training data. Similarly, Wang (2019) manually refines the text in existing table-to-text datasets and observes that this process also substantially alleviates fact hallucinations. Additionally, Parikh et al. (2020) instruct annotators to revise verified sentences from Wikipedia rather than directly creating new sentences when constructing table-to-text training data. This approach has also been proven to result in improved factuality of results.

With the advent of the LLM era, curating training data during pre-training has become increasingly challenging due to the vast scale of pre-training corpora (as exemplified in Table 7). For instance, Llama 2 (Touvron et al. 2023b) conducts pre-training on about two trillion tokens. Therefore, compared to manual curation, a more practical approach today could be *automatically* selecting reliable data or filtering out noisy

Table 7
The pre-training data size of popular LLMs.

LLM	Pre-train Data Size
GLM (Zeng et al. 2022)	400B tokens
BLOOM (Scao et al. 2022)	366B tokens
GPT-3 (Brown et al. 2020)	300B tokens
LLaMA (Touvron et al. 2023a)	1.4T tokens
Llama 2 (Touvron et al. 2023b)	2T tokens

data. For example, the pre-training data of GPT-3 (Brown et al. 2020) is cleaned by using similarity to a range of high-quality reference corpora. The developers of Falcon (Penedo et al. 2023) carefully extract high-quality data from the Web via heuristic rules and prove that properly curated pre-training corpora lead to powerful LLMs. Li et al. (2023e) propose phi-1.5, a 1.3 billion parameter LLM pre-trained on filtered “textbook-like” synthetic data, which exhibits many traits of much larger LLMs. In order to mitigate hallucinations, current LLMs tend to collect pre-training data from credible text sources. The developers of Llama 2 (Touvron et al. 2023b) strategically up-sample data from highly factual sources, such as Wikipedia,² when constructing the pre-training corpus. Lee et al. (2022) propose to prepend the topic prefix to sentences in the factual documents to make each sentence serve as a standalone fact during pre-training. Concretely, they treat the document name as the topic prefix and observe this method improves LMs’ performance on TruthfulQA.

Summary & Discussion. The mitigation of hallucinations during pre-training is primarily centered around *the curation of pre-training corpora*. Given the vast scale of existing pre-training corpora, current studies predominantly use simple heuristic rules for data selection and filtering. A potential avenue for exploration could be devising more effective selection or filtering strategies.

5.2 Mitigation During SFT

As a common practice, current LLMs collectively undergo the process known as supervised fine-tuning (SFT) to elicit their knowledge acquired from pre-training and learn how to interact with users (Wang et al. 2023c; Zhang et al. 2023b). SFT generally involves first annotating or collecting massive-task instruction-following data (Chung et al. 2022; Taori et al. 2023), followed by fine-tuning pre-trained foundational LLMs on this data using maximum likelihood estimation (Wei et al. 2021). By using well-designed SFT strategies, many recent studies claim to have built LLMs that achieve performance on par with ChatGPT (Wang et al. 2023b) on the specific domain.

5.2.1 Mitigating Hallucination Introduced in Pre-training During SFT. The SFT process, as a critical step in aligning LLMs with human preference, can contribute to mitigating hallucinations that originate from the pre-training phase. During pre-training, LLMs are exposed to vast and diverse corpora, which may include inaccurate, outdated, or inconsistent information. These flaws in the training data can result in models generating hallucinated outputs. The SFT process, by using carefully designed instruction-tuning datasets, has the potential to counteract such shortcomings by selectively emphasizing truthfulness, consistency, and reliability in downstream tasks. To maximize the benefits of this alignment process, curating high-quality instruction-tuning data is crucial. Given the relatively small volume of SFT data (refer to Table 8), both manual curation and automatic selection methods are feasible solutions. For instance, Zhou et al. (2023a) constructed a highly curated instruction-tuning dataset containing 1,000 samples, meticulously annotated by human experts to ensure correctness and diversity. On the other hand, other studies (e.g., Chen et al. 2023b; Cao, Kang, and Sun

² While Wikipedia is often regarded as a highly factual and comprehensive source, sometimes there are exceptions. For instance, a portion of Swedish Wikipedia articles are bot-generated. This reliance on automated article generation raises potential concerns over the accuracy and reliability in certain sections of Wikipedia.

Table 8
The size of popular SFT datasets.

SFT Dataset	Data Size
Alpaca (Taori et al. 2023)	52k samples
GPT4-Alpaca (Peng et al. 2023b)	52k samples
Baize (Xu et al. 2023a)	210k samples
Dolly (Conover et al. 2023)	15k samples
Open-assistant (Köpf et al. 2023)	34k samples
LIMA (Zhou et al. 2023a)	1k samples

2023; Lee, Hunter, and Ruiz 2023) pursue automatic selection strategies, where LLMs act as evaluators or specific rules are devised to filter for high-quality data. Experimental results on hallucination-related benchmarks, such as TruthfulQA (Lin, Hilton, and Evans 2021), show that LLMs fine-tuned with these curated datasets outperform those trained on less selective data, achieving greater levels of factuality and truthfulness. Besides curation strategies, domain-specific knowledge integration during SFT has also been explored to directly tackle hallucinations from specific knowledge gaps. Mohamed et al. (2023) propose enhancing SFT datasets with content derived from expert-verified, domain-specific knowledge sets. This augmentation helps models fill in factual gaps that otherwise lead to hallucinations when they encounter queries requiring specialized information.

5.2.2 Mitigating Hallucination Introduced in SFT. Another perspective highlights that the SFT stage itself may introduce new hallucination risks. Schulman (2023) emphasize a potential issue linked to behavior cloning (Torabi, Warnell, and Stone 2018), where the model learns by imitating human-provided demonstrations without understanding the underlying objectives. The SFT process of LLMs can be viewed as a special case of behavior cloning, where LLMs learn the format and style of interaction by mimicking humans. As for LLMs, despite having encoded a substantial amount of knowledge into their parameters, there remains knowledge that surpasses their capacity (Yin et al. 2023; Ren et al. 2023). By cloning human behaviors during SFT, LLMs learn to respond to all questions with a predominantly positive tone, without assessing whether these questions exceed their knowledge boundaries (see Figure 3). As a result, during inference, if prompted to answer questions related to unlearned knowledge, they are likely to confidently produce hallucinations. Furthermore, Gekhman et al. (2024) demonstrate that while SFT instances containing new knowledge are eventually learned, they also increase the model’s tendency to hallucinate. One way to remit this problem can be the honesty-oriented SFT, which means introducing some honest samples into the SFT data. The honest samples refer to responses that admit incompetence, such as “Sorry, I don’t know”. The Moss project (Sun et al. 2023b) open-sourced their SFT data, which includes such honest samples. Llama Team (2024) and Wan et al. (2024) teach the model to refuse to answer by incorporating refusal responses into the SFT data. We observed that models tuned with them could learn to refuse to answer specific questions, therefore helping reduce hallucinations. Another line of work focuses on contrast training. Zou et al. (2023a) introduce a contrast vector loss, which utilizes a pair of contrastive prompts with the same input. Zhang, Yu, and Feng (2024) utilize contrastive learning to partition the hidden state into distinct truthful and semantic spaces. The motivation

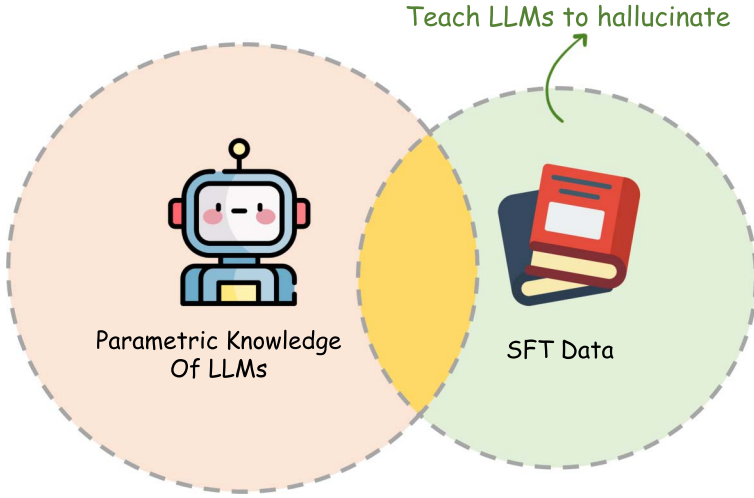


Figure 3

The SFT data usually contains samples that exceed LLMs’ parametric knowledge, which may result in hallucinations.

behind this work is that contrastive learning can help the model distinguish between factual and non-factual information.

Summary & Discussion. Curating the training data is one approach for mitigating hallucinations during the SFT phase. Thanks to the acceptable volume of SFT data, they can be manually curated by human experts. Recently, we have performed a preliminary human inspection and observed that some widely used synthetic SFT data, such as Alpaca (Taori et al. 2023), contains a considerable amount of hallucinated answers due to the lack of human inspection. This calls for careful attention when researchers build SFT datasets based on *self-instruct* (Wang et al. 2023c).

Previous work also pointed out that the SFT process may inadvertently introduce hallucinations, by forcing LLMs to answer questions that surpass their knowledge boundaries. Some researchers have suggested *honesty-oriented SFT* as a solution. However, we argue this method has two main problems. First, it exhibits limited generalization capabilities towards out-of-distribution (OOD) cases. Second, the annotated honest samples just reflect the incompetence and uncertainty of annotators rather than those of LLMs, as annotators are unaware of LLMs’ real knowledge boundaries. Such challenges make solving this issue during SFT sub-optimal.

5.3 Mitigation During RLHF

Nowadays, many researchers attempt to further improve the supervised fine-tuned LLMs via reinforcement learning from human feedback (RLHF) (Fernandes et al. 2023). This process consists of two steps: (1) train a reward model as the proxy for human preference, which aims to assign an appropriate reward value to each LLM response; and (2) optimize the SFT model with the reward model’s feedback, by using reinforcement learning (RL) algorithms such as PPO (Schulman et al. 2017).

Leveraging human feedback not only closes the gap between machine-generated content and human preference but also helps LLMs align with desired criteria or goals.

Table 9
An example of reward design for mitigating LLM hallucinations through RL (Schulman 2023).

Situation	Reward Value
Unhedged Correct	+1
Hedged Correct	+0.5
Uninformative	0
Hedged Wrong	-2
Unhedged Wrong	-4

One commonly used criterion today is “3H”, which denotes *helpful*, *honest*, and *harmless* (Ouyang et al. 2022; Bai et al. 2022; Zheng et al. 2023b). The *honest* aspect here just refers to the minimization of hallucinations in LLM responses. Current advanced LLMs, such as InstructGPT (Ouyang et al. 2022), ChatGPT (OpenAI 2023a), GPT4 (OpenAI 2023b), and Llama2-Chat (Touvron et al. 2023b), have collectively considered this aspect during RLHF. For example, GPT4 uses synthetic hallucination data to train the reward model and perform RL, which increases accuracy on TruthfulQA (Lin, Hilton, and Evans 2021) from about 30% to 60%. Moreover, Lightman et al. (2023) use *process supervision* to detect and mitigate hallucinations for reasoning tasks, which provides feedback for each intermediate reasoning step.

As discussed in the previous section, the phenomenon of behavior cloning during the SFT stage can potentially lead to hallucinations. Some researchers have attempted to address this issue by integrating honest samples into the original SFT data. However, this approach has certain limitations, such as unsatisfactory OOD generalization capabilities and a misalignment between human and LLM knowledge boundaries. In light of this, Schulman (2023) proposes to solve this problem during RLHF. They design a special reward function just for mitigating hallucinations, as shown in Table 9. “Unhedged/Hedged Correct/Wrong” here means the LLM provides correct or wrong answers with a positive or hesitant tone. “Uninformative” denote the safe answers like “I don’t know”. The core idea is to encourage LLMs to challenge the premise, express uncertainty, and commit incapability by learning from specially designed rewards. This method, which we refer to as honesty-oriented RL, offers several advantages over honesty-oriented SFT. The primary benefit is that it allows LLMs to freely explore their knowledge boundaries, thereby enhancing their generalization capabilities to OOD cases. Additionally, it reduces the need for extensive human annotation and eliminates the requirement for annotators to guess the knowledge boundaries of LLMs.

Summary & Discussion. Reinforcement learning can guide LLMs in exploring their knowledge boundaries, enabling them to decline to answer questions beyond their capacity rather than fabricating untruthful responses. However, we note this approach also poses unique challenges. For instance, RL-tuned LLMs may exhibit over-conservatism due to an imbalanced trade-off between *helpfulness* and *honesty* (Ouyang et al. 2022). An example of this is illustrated in Table 10. As observed in this case, ChatGPT tends to be overly hedged and refrains from providing a clear answer that it already knows, as evidenced in another dialogue turn. This could be attributed to the unreasonable design of the reward function or the poor quality of the training data for the reward model. We hope future work can take such problems into consideration.

Table 10
A real example of the *over-conservative* phenomenon of ChatGPT (July 2023 Version). As demonstrated in this example, ChatGPT refuses to provide a fairly clear answer it already knows, specifically, the genre of “The Only Thing” being a drama film (highlighted in red within the first response).

User Input	Introduce the film “The Only Thing” (Four Flaming Days) to me.
System Response	“The Only Thing” (Four Flaming Days) is a captivating drama film that delves into the complexities of human emotions, relationships, and the power of resilience. Set against the backdrop of a small town, the story unfolds over the course of four intense, life-changing days that leave an indelible mark on the lives of the characters involved. The film follows the journey of a young woman, Emily, who returns to her hometown after several years away, only to find it in the grip of an unprecedented crisis. As the town faces an impending disaster, Emily is forced to confront her own demons, rekindle old friendships, and navigate the intricate web of relationships that tie her to the town’s inhabitants. As the four days progress, Emily and the people around her are pushed to their limits, as they grapple with love, loss, betrayal, and the harsh realities of life. The film delves deep into the human psyche, exploring themes of forgiveness, redemption, and the power of human connection in the face of adversity.
User Input	What genre of movie is “The Only Thing” (Four Flaming Days)?
System Response	I’m sorry. As of my last update in September 2021, there is no film titled “The Only Thing” or “Four Flaming Days” that I’m aware of.

Table 11
A summary of some recent studies on resorting to external knowledge to mitigate hallucinations. For Timing of Using, “G” means Generation-timing, while “P” means post-processing. We use abbreviations for some application task names, including QA (Question Answering), FV (Fact Verification), and LM (Language Modeling).

Method	Timing of Using	Knowledge Source	Application Task
WebGPT (Nakano et al. 2021)	G	Search API	QA
Adaptive-Retrieval (Mallen et al. 2023)	G	Wikipedia	QA
ReACT (Yao et al. 2022)	G	Wikipedia	QA & FV
RETRO (Borgeaud et al. 2022)	G	Unstructured Corpus	LM & QA
Chain-of-Knowledge (Li et al. 2023c)	G	Structured Knowledge Base	QA & FV & Decision
RAIR (Gao et al. 2023a)	P	Search API	QA
Verify-then-Edit (Zhao et al. 2023b)	P	Wikipedia, Search API, etc.	QA
LLM-Augmenter (Peng et al. 2023a)	P	Web documents, Databases	QA
REFEED (Yu et al. 2023b)	P	Wikipedia	QA, Dialogue
CRITIC (Gou et al. 2023)	P	Search API, Code Executor, Calculator, etc.	QA & Program & Toxicity
FacTool (Chern et al. 2023)	P	Search API, Code Executor, Calculator, etc.	QA & Reasoning & Generation

5.4 Mitigation During Inference

Compared with the aforementioned training-time mitigation approaches, mitigating hallucinations during inference could be more cost-effective and controllable. Therefore, most existing studies focus on this direction, which we will introduce in detail in the following sections.

5.4.1 *Designing Decoding Strategies.* Decoding strategies, such as greedy decoding and beam search decoding, determine how we choose output tokens from the probability distribution generated by models (Zarrieß, Voigt, and Schüz 2021).

Lee et al. (2022) carry out a factuality assessment of content generated by LLMs using different decoding strategies. They find that nucleus sampling (a.k.a. top-*p*

sampling) (Holtzman et al. 2019) falls short of greedy decoding in terms of factuality. They argue that this underperformance could be attributed to the randomness introduced by top- p sampling to boost diversity, creativity, which may inadvertently lead to hallucinations since LLMs tend to fabricate information to generate diverse responses. In view of this, they introduce a decoding algorithm termed *factual-nucleus sampling*, which aims to strike a more effective balance between diversity and factuality by leveraging the strengths of both top- p and greedy decoding.

Dhuliawala et al. (2023) develop a decoding framework known as the Chain-of-Verification (COVE). This framework is based on the observation that independent verification questions typically yield more accurate facts than those presented in long-form answers. The COVE framework initially plans verification questions, and then answers these questions to ultimately produce an enhanced, revised response. Experimental results on list-based questions, closed book QA, and long-form text generation demonstrate that COVE can effectively mitigate hallucination.

Another study by Li et al. (2023b), introduces a novel Inference-Time Intervention (ITI) method to improve the truthfulness of LLMs. This method is based on the assumption that LLMs possess latent, interpretable sub-structures associated with factuality. The ITI method comprises two steps: (1) fitting a binary classifier on top of each attention head of the LLM to identify a set of heads that exhibit superior linear probing accuracy for answering factual questions, and (2) shifting model activations along these factuality-related directions during inference. The ITI method leads to a substantial performance improvement on the TruthfulQA benchmark (Lin, Hilton, and Evans 2021).

Distinct from the aforementioned studies, Shi et al. (2023b) instead concentrate on the retrieval-augmentation setting. Prior research has shown that LLMs sometimes fail to adequately attend to retrieved knowledge when addressing downstream tasks, particularly when the retrieved knowledge conflicts with the parametric knowledge of LLMs (Zhou et al. 2023b; Xie et al. 2023). To address this issue, Shi et al. (2023b) propose a straightforward context-aware decoding (CAD) strategy. The core idea of CAD is to perform a contrastive ensemble of $p_{\theta}(y_t | x, c, y_{<t})$ and $p_{\theta}(y_t | x, y_{<t})$, where θ represents the LM, x is the input query, c is the context, y is the response, and t is the time step. $p_{\theta}(y_t | x, c, y_{<t})$ means the generation probability distribution of t -th token when given the context while $p_{\theta}(y_t | x, y_{<t})$ denotes the distribution only considering the query. The CAD method aims to compel LLMs to pay more attention to contextual information instead of over-relying their own parametric knowledge to make decisions. Experimental results show that CAD effectively elicits the ability of LLMs to exploit retrieved knowledge and thus reduces factual hallucinations on downstream tasks. Another work, DoLA (Chuang et al. 2023), also uses the idea of contrastive decoding to reduce hallucination. However, they contrast the generation probabilities from different layers of LLMs, as they find that linguistic and factual information is encoded in distinct sets of layers. ICD (Zhang et al. 2023e) further constructs a factually weak LLM as the penalty term.

Summary & Discussion. Designing decoding strategies to mitigate hallucinations in LLMs during inference is typically done in a plug-and-play manner. Therefore, this method is easy to deploy, making it promising for practical applications. However, for this approach, most existing works require accessing the token-level output probabilities, while a substantial number of current LLMs can only return generated content through limited APIs (e.g., ChatGPT). Consequently, we encourage future research in this direction to explore within a more strict *black-box* setting.

5.4.2 Resorting to External Knowledge. Using external knowledge as supplementary evidence to assist LLMs in providing truthful responses recently represents a burgeoning solution (Ren et al. 2023; Mialon et al. 2023). This approach typically consists of two steps. The first step entails accurately obtaining knowledge related to the user instructions. Once useful knowledge has been achieved, the second step involves leveraging such knowledge to guide the generation of the responses. We provide a comprehensive review of the latest progress in this direction, focusing on the specific strategies utilized in these two steps, respectively. We also present a summary of recent studies in Table 11.

Knowledge Acquisition. LLMs have internalized vast amounts of knowledge into their parameters through extensive pre-training and fine-tuning, which can be referred to as *parametric knowledge* (Roberts, Raffel, and Shazeer 2020). However, incorrect or outdated parametric knowledge can easily lead to hallucinations (Xie et al. 2023). To remedy this, researchers have proposed acquiring reliable, up-to-date knowledge from credible sources as a form of hot patching for LLMs (Lewis et al. 2020b; Li et al. 2022a). We summarize the two primary sources of such knowledge as follows.

- (1) **External knowledge bases.** The majority of existing studies retrieve information from external knowledge bases, such as large-scale unstructured corpora (Cai et al. 2021; Borgeaud et al. 2022), structured databases (Liu 2022; Li et al. 2023c), specific Websites like Wikipedia (Yao et al. 2022; Peng et al. 2023a; Li, Peng, and Zhang 2023; Yu et al. 2023b), or even the entire Internet (Lazaridou et al. 2022; Yao et al. 2022; Gao et al. 2023a; Liu et al. 2023c). The evidence retrieval process typically uses various sparse (e.g., BM25 [Robertson, Zaragoza et al. 2009]) or dense (e.g., PLM-based methods [Zhao et al. 2022]) retrievers. Search engines, such as Google Search, can also be viewed as a special kind of information retriever (Nakano et al. 2021; Lazaridou et al. 2022; Yao et al. 2022; Gao et al. 2023a). Additionally, Luo et al. (2023) propose the parameter knowledge guiding framework which retrieves knowledge from the parametric memory of fine-tuned white-box LLMs. Feng, Zhang, and Fei (2023) try to teach LLMs to search relevant domain knowledge from external knowledge graphs to answer domain-specific questions.
- (2) **External tools.** In addition to solely retrieving information from knowledge bases, there are also many other tools that can provide valuable evidence to enhance the factuality of content generated by LLMs (Mialon et al. 2023; Qin et al. 2023; Qiao et al. 2023). For instance, FacTool (Chern et al. 2023) utilizes different tools to help detect hallucinations in LLMs for specific downstream tasks, such as *search engine API* for knowledge-based QA, *code executor* for code generation, and *Google Scholar API* for scientific literature review. CRITIC (Gou et al. 2023) also enables LLMs to interact with multiple tools and revise their responses autonomously, which has been proven to effectively improve truthfulness.

Knowledge Utilization. Once relevant knowledge is obtained, it could be used at different stages to mitigate hallucinations within LLMs. Existing methods for knowledge

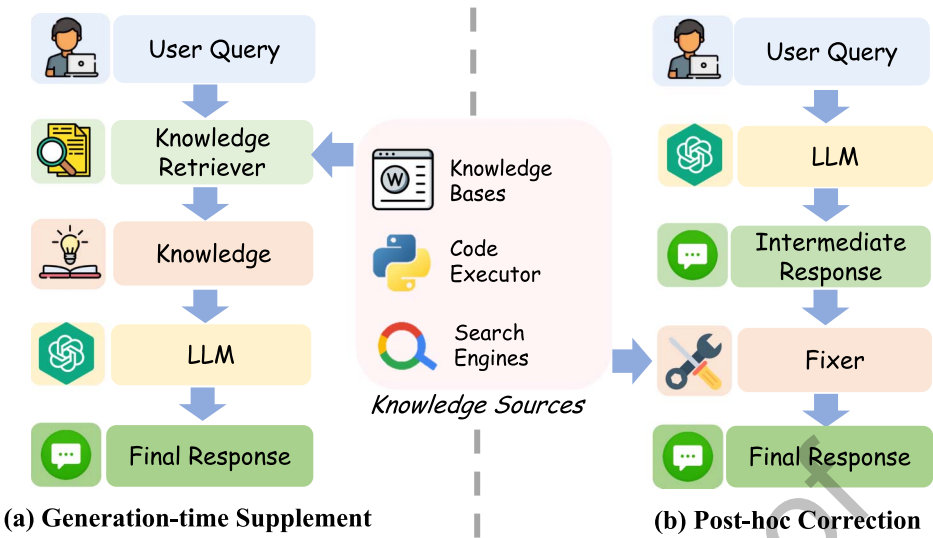


Figure 4
The illustrations of two distinct methods for utilizing external knowledge to reduce hallucinations in LLMs’ responses.

utilization can be roughly divided into two categories, as detailed below and illustrated in Figure 4.

- (1) **Generation-time supplement.** The most straightforward approach to utilize retrieved knowledge or tool feedback is to directly concatenate them with user queries before prompting LLMs (Shi et al. 2023c; Mallen et al. 2023; Ram et al. 2023). This method is both effective and easy to implement. Such knowledge is also referred to as *context knowledge* (Shi et al. 2023b). Existing studies have demonstrated that LLMs possess a strong capability for in-context learning (Dong et al. 2022), which enables them to extract and utilize valuable information from context knowledge to rectify nonfactual claims they previously generated.
- (2) **Post-hoc correction.** Another common practice involves constructing an auxiliary fixer to rectify hallucinations during the post-processing stage (Cao et al. 2020; Zhu et al. 2021; Fabbri et al. 2022a). The fixer can be either another LLM (Peng et al. 2023a; Zhang et al. 2023d; Chern et al. 2023; Gou et al. 2023) or a specific small model (Chen et al. 2023a). Such fixers first interact with external knowledge sources to gather sufficient evidence, and then correct hallucinations. For example, RARR (Gao et al. 2023a) directly prompts an LLM to ask questions about the content that needs to be corrected from multiple perspectives. Then it uses search engines to retrieve relevant knowledge. The LLM-based fixer finally makes corrections based on retrieved evidence. The Verify-then-Edit approach (Zhao et al. 2023a) aims to enhance the factuality of predictions by post-editing reasoning chains based on external knowledge sourced

from Wikipedia. To achieve better performance, LLM-Augmenter (Peng et al. 2023a) prompts LLMs to summarize retrieved knowledge before feeding it into the fixer. Moreover, FacTool (Chern et al. 2023) and CRITIC (Gou et al. 2023) propose to utilize various external tools to obtain evidence for the fixer.

Summary & Discussion. Resorting to external knowledge to mitigate hallucinations in LLMs offers several advantages. First, this method circumvents the need for modifying LLMs, making it a plug-and-play and efficient solution. Second, it facilitates the easy transfer of proprietary knowledge (e.g., a company’s internal data) and real-time updated information to LLMs. Lastly, this approach enhances the interpretability of information generated by LLMs by allowing the tracing of generation results back to the source evidence (Gao et al. 2023b; Yue et al. 2023). However, this direction also presents some remaining challenges. We discuss some of them below.

- (1) **Knowledge verification.** In the era of LLMs, the external knowledge source could extend beyond a single document corpus or a specific website to encompass the entire Internet. However, the information from the Internet is in the wild, which means it may also be fabricated, or even generated by LLMs themselves (Alemohammad et al. 2023). How to verify the authenticity of retrieved knowledge from the Internet is an open and challenging problem to be solved.
- (2) **Performance/efficiency of retriever/fixer.** The performance of the retriever/fixer plays a vital role in ensuring the effects of hallucination mitigation. Future work may consider jointly optimizing the whole working flow (retriever→LLM→fixer) via reinforcement learning (Qiao et al. 2023) or other techniques. Additionally, the efficiency of the retriever/fixer is another important factor to be considered, as the generation speed of existing LLMs is already a significant burden (Ning et al. 2023).
- (3) **Knowledge conflict.** As introduced before, the retrieved knowledge may conflict with the parametric knowledge stored by LLMs (Qian, Zhao, and Wu 2023). Shi et al. (2023b) reveal that LLMs may fail to sufficiently exploit retrieved knowledge when knowledge conflict happens. Xie et al. (2023) take a more cautious look at this phenomenon. How to fully utilize context knowledge is an under-explored question. For example, Liu et al. (2023d) find that the performance of retrieval-augmented LLMs significantly degrades when they must access evidence in the middle of long contexts.

5.4.3 Exploiting Uncertainty. Uncertainty serves as a valuable indicator for detecting and mitigating hallucinations during the inference process (Manakul, Liusie, and Gales 2023). Typically, it refers to the confidence level of model outputs (Jiang et al. 2021; Huang et al. 2023a; Duan et al. 2023; Zeng et al. 2025). Uncertainty can assist users in determining when to trust LLMs. Provided that the uncertainty of LLM responses can be accurately characterized, users can filter out or rectify LLMs’ claims with high

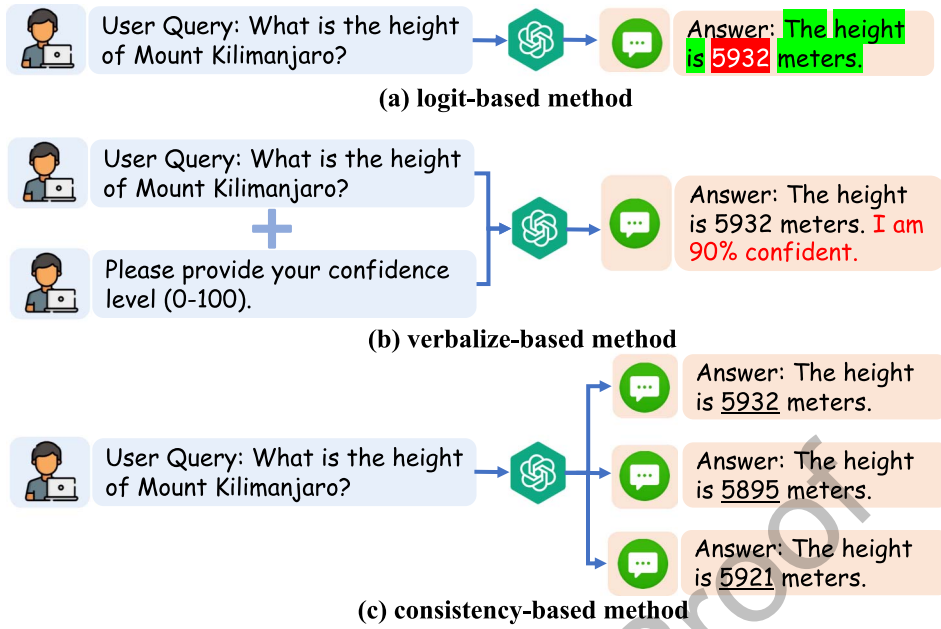


Figure 5
The illustrations of three typical methods for estimating LLM uncertainty. In the example of the *logit-based* method, we use the red/green background to distinguish tokens with low/high generation probabilities. In the example of the *consistency-based* method, the responses are acquired from multiple sampling.

uncertainty since such claims are more prone to be fabricated ones (Lin, Trivedi, and Sun 2023).

Generally speaking, methods for estimating the uncertainty of LLMs can be categorized into three types (Xiong et al. 2023), as listed below. To facilitate understanding, we also present illustrative examples for these methods in Figure 5.

- (1) **Logit-based estimation.** The first method is the *logit-based* method, which requires access to the model logits and typically measures uncertainty by calculating token-level probability or entropy. This method has been widely used in the machine learning community (Guo et al. 2017).
- (2) **Verbalize-based estimation.** The second is the *verbalize-based* method, which involves directly requesting LLMs to express their uncertainty, such as using the following prompt: “Please answer and provide your confidence score (from 0 to 100).” This method is effective due to the impressive verbal and instruction-following capabilities of LLMs. Notably, Xiong et al. (2023) and Zhang et al. (2024) further suggest using chain-of-thoughts prompts (Wei et al. 2022) to enhance this method.
- (3) **Consistency-based estimation.** The third is the *consistency-based* method (Wang et al. 2022; Shi et al. 2022; Zhao et al. 2023a). This method operates on the assumption that LLMs are likely to provide logically inconsistent responses for the same question when they are indecisive and hallucinating facts.

Several recent studies have leveraged uncertainty estimation for detecting and mitigating hallucinations in LLMs. SELF-CHECKGPT (Manakul, Liusie, and Gales 2023) is the first framework to detect LLM hallucinations based on uncertainty measurement in a zero-resource and black-box setting. The authors use a consistency-based approach for uncertainty estimation. A non-trivial challenge in SELF-CHECKGPT is determining how to measure the consistency of different responses. Manakul, Liusie, and Gales (2023) perform experiments with BERTScore (Zhang et al. 2019), QA-based metrics (Fabbri et al. 2022b), and n -gram metrics. They find that a combination of these approaches yields the best results. Agrawal, Mackey, and Kalai (2023) further adopt the verbalize-based method to evaluate the hallucination rate of LLMs for fabricating references. Uncertainty can also serve as an indicator to mitigate hallucination. For example, Mündler et al. (2023) directly utilize an additional LLM to assess whether two LLM responses are logically contradictory given the same context (Luo, Xie, and Ananiadou 2023), which means at least one of them is hallucinated. Consequently, they use another LLM to revise such self-contradictory hallucinations from two responses. Varshney et al. (2023), on the other hand, use the logit-based method to detect false concepts in LLMs' responses with high uncertainty. They then fix such content with auxiliary retrieval-augmented LLMs.

Additionally, Zhao et al. (2023b) present a Pareto optimal self-supervision framework. This framework utilizes available programmatic supervision to assign a risk score to LLM responses, which can serve as an indicator of hallucinations. Luo, Xiao, and Ma (2023) introduce a pre-detection self-evaluation technique, which aims to evaluate the familiarity of LLMs with the concepts in user prompts and prevent the generation of content about those unfamiliar concepts.

Summary & Discussion. Exploiting uncertainty to identify and mitigate LLM hallucinations is a promising research direction today. Three primary approaches exist for estimating the uncertainty of LLMs, each presenting its unique challenges. First, the *logit-based* method is becoming less applicable for modern commercial LLMs as they are usually closed-source and black-box, rendering their output logits inaccessible. Second, regarding the *verbalize-based* method, researchers have observed that LLMs tend to display a high degree of overconfidence when expressing their confidence (Xiong et al. 2023). Third, the effective measurement of the consistency of different responses remains an unresolved issue in the *consistency-based* method (Manakul, Liusie, and Gales 2023). We believe that leveraging uncertainty is crucial in developing trustworthy LLMs and encourage future research to address the aforementioned challenges in this field.

5.4.4 Multi-agent Interaction. Some recent research has sought to address the hallucination problem in LLMs from a multi-agent perspective, wherein multiple LLMs (also known as agents) independently propose and collaboratively debate their responses to reach a single consensus, as exemplified in Figure 6. The results of Du et al. (2023) is a pioneering work in this line. They initially developed a benchmark for assessing the factual accuracy of prominent computer scientist biographies generated by LMs. Their findings reveal that an individual LLM can easily generate hallucinated information within this benchmark; however, such hallucinations can be mitigated by engaging multiple LLMs in a debate to achieve consensus. Also, Cohen et al. (2023) ask one LLM to generate claims (acting as EXAMINEE) and another to raise questions about these claims and check the truthfulness of them (acting as EXAMINER). Wang et al. (2023d) instead propose prompting a single LLM to identify, simulate, and iteratively self-collaborate with multiple personas, such as Harry Potter Fan and Jay Chou Fan. By

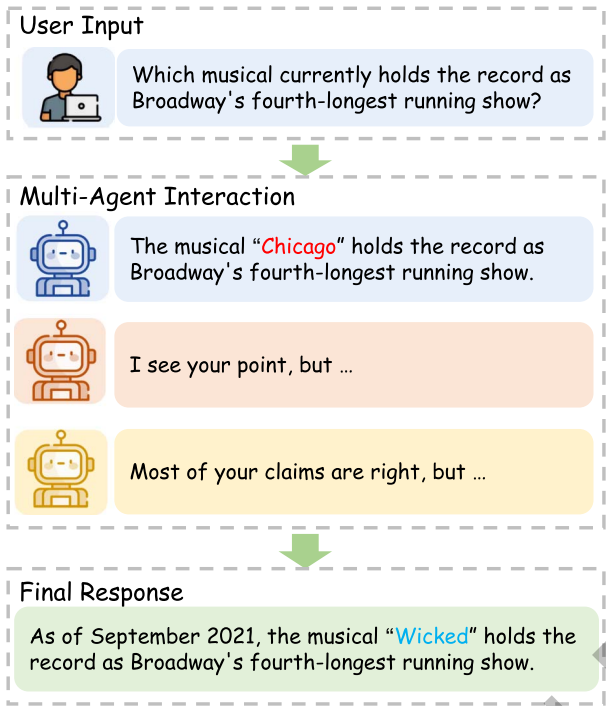


Figure 6
An example of the process of multi-agent interaction for mitigating LLM hallucinations.

leveraging an LLM as a cognitive synergist, it effectively reduces hallucinations with relatively low costs.

5.4.5 Prompt Engineering. Existing research highlights that the behavior of LLMs can significantly vary based on the prompts given by users (Si et al. 2022; Zhu et al. 2023). In terms of hallucination, users may encounter an LLM that initially responds accurately but begins to hallucinate information when using different prompts. In light of this observation, Zhang et al. (2023a) endeavor to engineer more effective prompts to mitigate hallucination. Concretely, they utilize the chain-of-thought prompt (Wei et al. 2022) to compel LLMs to generate reasoning steps before providing the final answers. However, chain-of-thought may introduce some new challenges. The potential of hallucinated reasoning steps is one of them. Furthermore, a popular practice nowadays involves explicitly instructing LLMs not to disseminate false or unverifiable information when designing the “system prompt,” that is, the special messages used to steer the behavior of LLMs. The following system prompt used for Llama 2-Chat (Touvron et al. 2023b) exemplifies this approach: *If you don’t know the answer to a question, please don’t share false information.*

5.4.6 Analyzing LLMs’ Internal States. Azaria and Mitchell (2023) contend that LLMs may be aware of their own falsehoods, implying that their internal states could be utilized to detect hallucinations. They propose Statement Accuracy Prediction based on Language Model Activations (SAPLMA), which adds a classifier on top of each hidden layer of the LLM to determine truthfulness. Experimental results indicate that LLMs might “know”

when the statements they generate are false, and SAPLMA can effectively extract such information. The ITI method (Li et al. 2023b) is also grounded in a similar hypothesis. They further shift model activations alongside factuality-related heads during inference and discover that this can mitigate hallucinations. These studies suggest that “the hallucination within LLMs may be more a result of generation techniques than the underlying representation” (Agrawal, Mackey, and Kalai 2023).

5.5 Other Methods

In addition to the above approaches, other techniques demonstrating the potential for reducing hallucinations are shown below.

Human-in-the-Loop. Zhang et al. (2023c) posit that a potential cause of hallucination in LLMs could be the misalignment between knowledge and user questions, a phenomenon that is particularly prevalent in the context of retrieval-augmented generation. To address this issue, they introduce MixAlign, a human-in-the-loop framework that utilizes LLMs to align user queries with stored knowledge, and further encourages users to clarify this alignment. By refining user queries iteratively, MixAlign not only reduces hallucinations but also enhances the quality of the generated content.

Optimizing Model Architecture. Several studies have explored modifying the architecture of LMs to mitigate hallucinations. Examples include the multi-branch decoder (Rebuffel et al. 2022) and the uncertainty-aware decoder (Xiao and Wang 2021). Li et al. (2023f) suggest using a bidirectional autoregressive architecture in the construction of LLMs, which enables language modeling from both left-to-right and right-to-left. They claim that this design strategy could contribute to the reduction of hallucinations by effectively leveraging bidirectional information.

6. Outlooks

In this section, we discuss a few unresolved challenges in the investigation of hallucinations within LLMs and offer our insights into potential future research directions.

Reliable Evaluation. Although considerable effort has been dedicated to building evaluation benchmarks for quantitatively assessing hallucination in LLMs, there are still issues that need to be solved. The automatic evaluation in the *generation-style* hallucination benchmark cannot accurately reflect the performance or align with human annotation. Such inaccuracy is reflected in two ways: (1) The automatic metric does not perfectly align with human annotations (Lin, Hilton, and Evans 2021; Min et al. 2023; Muhlgay et al. 2023); (2) The reliability of automatic metric varies across texts from different domains or generated by different LLMs (Min et al. 2023), resulting in reduced robustness for generalization. Although the *discrimination-style* benchmark (Li et al. 2023a; Muhlgay et al. 2023) could relatively accurately evaluate a model’s ability to distinguish hallucinations, the relationship between discrimination performance and generation performance is still unclear until now. These issues all need more in-depth exploration.

Multilingual Hallucination. Existing work in LLM hallucination primarily focuses on English, despite the existence of thousands of languages in the world. We hope that

Table 12

A real example in which ChatGPT (July 2023 Version) accurately answered a question in English conversation but presented hallucinations for the same question when communicating in Chinese (the correct population of Denver in 2020 is 715,522, according to <https://en.wikipedia.org/wiki/Denver>).

User Input (EN)	What is the population of Denver according to the 2020 census?
LLM Response (EN)	According to the 2020 United States Census, the population of Denver, Colorado, is 715,522 ✓
User Input (ZH)	根据2020年人口普查，丹佛的人口是多少？
LLM Response (ZH)	根据2020年人口普查，丹佛的人口为73,921✗
User Input (EN)	What is the population of Denver according to the 2020 census? Answer in Chinese.
LLM Response (ZH)	根据2020年人口普查，丹佛的人口为704,621 ✗

LLMs can possess the ability to handle various languages uniformly. Some previous studies have investigated the performance of LLMs on some multilingual benchmarks (Ahuja et al. 2023; Lai et al. 2023), and collectively found that their performance degenerates when generalizing to non-Latin languages. In terms of the hallucination problem, Guerreiro et al. (2023) observe that multilingual LLMs predominantly struggle with hallucinations in low-resource languages in the translation task. Potential follow-up work could include systematically measuring and analyzing LLM hallucinations across a wide variety of languages. As shown in Table 12, we find that LLMs such as ChatGPT provide accurate answers in English but expose hallucinations in other languages, leading to multilingual inconsistencies. The transfer of knowledge within LLMs from high-resource languages to low-resource ones also presents an interesting and promising research direction.

Multi-modal Hallucination. In an effort to improve the performance of complex multi-modal tasks, recent studies have proposed replacing the text encoder of existing vision-large models with LLMs, resulting in large vision-language models (LVLMs) (Liu et al. 2023b; Ye et al. 2023). Despite their success, some research reveals that LVLMs inherit the hallucination problem from LLMs and exhibit more severe multi-modal hallucinations compared with smaller models. For instance, Li et al. (2023d) discuss the *object hallucination* of LVLMs, wherein LVLMs generate content containing objects that are inconsistent with or absent from the input image, such as the example in Figure 7. To effectively measure object hallucinations generated by LVLMs, Liu et al. (2023a) propose a *GPT4-Assisted Visual Instruction Evaluation* (GAVIE) benchmark. Gunjal, Yin, and Bas (2023) introduce a multi-modal hallucination detection dataset named M-HalDetect, further allowing the study of unfaithful descriptions and inaccurate relationships beyond object hallucinations in LVLMs. Furthermore, in addition to images, some studies have extended LLMs to other modalities such as audio (Wu et al. 2023a; Su et al. 2023) and video (Maaz et al. 2023), making it interesting to investigate hallucination in these new scenarios.

Model Editing. As elaborated in § 4, hallucinations in LLMs may primarily stem from the memorization of false information or the absence of correct factual knowledge. To mitigate these issues in LLMs with minimal computational overhead, the concept of model editing has been introduced (Sinitsin et al. 2020; De Cao, Aziz, and Titov 2021).



Figure 7

An example of object hallucination in LVLMs. We highlight the hallucination in **red**, as there is no person under the tree in this picture.

This approach involves modifying the behavior of models in a manner that is both data- and computation-efficient. At present, there are two mainstream paradigms for model editing. The first involves the incorporation of an auxiliary sub-network (Mitchell et al. 2022; Huang et al. 2023b), while the second entails direct modification of the original model parameters (Meng et al. 2022a,b). This technique may be instrumental in eliminating LLMs’ hallucinations by editing their stored factual knowledge on purpose (Lanham et al. 2023; Onoe et al. 2023). However, this emerging field still faces numerous challenges. These could include editing black-box LLMs (Murty et al. 2022), in-context model editing (Zheng et al. 2023a), and multi-hop model editing (Zhong et al. 2023), among others.

Attack/Defense for Inducing Hallucination. As previously discussed, significant efforts have been undertaken by both researchers and companies to guarantee that LLMs produce truthful responses, ultimately improving the overall user experience. Cutting-edge commercial LLMs, such as GPT4 (OpenAI 2023b), appear to have acquired a decent ability to generate proper responses to factuality-related queries. However, they are not invincible. Several studies show that LLMs can be manipulated using techniques like meticulously crafted jailbreak prompts to elicit arbitrary desired responses (Wei, Haghtalab, and Steinhardt 2023; Zou et al. 2023b), including hallucinations. Consequently, the attacking and defending strategies for inducing hallucinations could also be a promising research direction. This is particularly important as the generation of fabricated information could potentially breach relevant laws, leading to the forced shutdown of LLM applications. This direction is also intimately tied to the robustness of existing hallucination mitigation methods.

More Settings for Defining and Mitigating Hallucination. Given that the current research on hallucinations in LLMs is still in its early stages, there are also many other intriguing

and promising avenues for further investigation. For instance, researchers have begun to treat LLMs as agents for open-world planning in the pursuit of artificial general intelligence (Park et al. 2023; Wang et al. 2023a). Addressing the hallucination problem within the context of LLMs-as-agents presents brand-new challenges and holds considerable practical value. Additionally, analyzing and tracing LLM hallucinations from the linguistic aspect is another interesting research topic. Rawte et al. (2023) show that the occurrence of LLM hallucination is closely related to linguistic nuances of the user prompts, such as readability, formality, and concreteness. We believe all these directions merit thorough exploration in future research.

7. Conclusion

With their strong understanding and generation capabilities in the open domain, LLMs have garnered significant attention from both academic and industrial communities. However, hallucination remains a critical challenge that impedes the practical application of LLMs. In this survey, we offer a comprehensive review of the most recent advances, primarily since the release of ChatGPT, that aim to evaluate, trace, and eliminate hallucinations within LLMs. We also delve into the existing challenges and discuss potential future directions. We aspire for this survey to serve as a valuable resource for researchers intrigued by the mystery of LLM hallucinations, thereby fostering the practical application of LLMs.

Ethics Issues of Hallucination in LLMs

As LLMs become increasingly integrated into various applications, hallucination raises significant ethics concerns. These ethics issues need to be carefully considered to mitigate potential harm and ensure responsible use of these powerful technologies. We list several ethics issues about LLM hallucinations as follows.

Misinformation and Public Trust. One of the most pressing ethical issues associated with LLM hallucinations is the potential spread of misinformation. When LLMs generate hallucinated responses, it can be inadvertently disseminated by users who trust the outputs of these models. As we discussed earlier, this is particularly concerning in domains such as healthcare, law, and finance, where inaccurate information can have serious consequences. The erosion of public trust in automated systems and artificial intelligence can also result from frequent hallucinations, undermining the perceived reliability of these technologies.

Impact on Decision-Making. Hallucinations in LLMs can have significant implications for decision-making in both individual and institutional contexts. For instance, if a legal advisor or medical professional relies on hallucinated outputs from an LLM, the consequences can be deeply detrimental. Ethically, it is essential to ensure that critical decisions are not unduly influenced by unreliable AI outputs, and that there are robust mechanisms for verifying the information provided by LLMs.

References

Adlakha, Vaibhav, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following

models for question answering. *arXiv preprint arXiv:2307.16877*.
https://doi.org/10.1162/tac1_a_00667
 Agrawal, Ayush, Lester Mackey, and Adam Tauman Kalai. 2023. Do language

- models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.
- Ahuja, Kabir, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. MEGA: Multilingual evaluation of generative AI. *arXiv preprint arXiv:2303.12528*. <https://doi.org/10.18653/v1/2023.emnlp-main.258>
- Akyürek, Ekin, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Tracing knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*. <https://doi.org/10.18653/v1/2022.findings-emnlp.180>
- Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel Lejeune, Ali Siahtkoo, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*. <https://doi.org/10.52591/1xai202312101>
- Azaria, Amos and Tom Mitchell. 2023. The internal state of an LLM knows when its lying. *arXiv preprint arXiv:2304.13734*. <https://doi.org/10.18653/v1/2023.findings-emnlp.68>
- Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislaw Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Bai, Yushi, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. 2024. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.
- Bang, Yejin, 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 preprint arXiv:2302.04023*. <https://doi.org/10.18653/v1/2023.ijcnlp-main.45>
- Bickel, Steffen, Peter Haider, and Tobias Scheffer. 2005. Predicting sentences using N-gram language models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 193–200. <https://doi.org/10.3115/1220575.1220600>
- Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Cai, Deng, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. 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)*, pages 7307–7318. <https://doi.org/10.18653/v1/2021.acl-long.567>
- Cao, Meng, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258. <https://doi.org/10.18653/v1/2020.emnlp-main.506>
- Cao, Yihan, Yanbin Kang, and Lichao Sun. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*.
- Cao, Zouying, Yifei Yang, and Hai Zhao. 2023. AutoHall: Automated hallucination dataset generation for large language models. *CoRR*, abs/2310.00259. <https://doi.org/10.48550/ARXIV.2310.00259>
- Chang, Kai Wei, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.
- Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on

- evaluation of large language models. *arXiv preprint arXiv:2307.03109*.
- Chen, Anthony, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. PURR: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.
- Chen, Hongshen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35. <https://doi.org/10.1145/3166054.3166058>
- Chen, Jiawei, Dingkan Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024a. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.
- Chen, Lichang, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023b. AlpaGasus: Training a better Alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.
- Chen, Xiang, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2024b. FactCHD: Benchmarking fact-conflicting hallucination detection. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 6216–6224. <https://doi.org/10.24963/ijcai.2024/687>
- Cheng, Qinyuan, 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 preprint arXiv:2310.03368*.
- Chern, I-Chun, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality detection in generative AI—A tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.
- Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Chuang, Yung-Sung, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.
- Chung, Hyung Won, 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 preprint arXiv:2210.11416*.
- Cohen, Roi, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*. <https://doi.org/10.18653/v1/2023.emnlp-main.778>
- Conover, Mike, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the world’s first truly open instruction-tuned LLM. <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>
- Cui, Leyang, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *Proceedings of EMNLP*, pages 2328–2337. <https://doi.org/10.18653/v1/2021.emnlp-main.179>
- Dale, David, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, pages 36–50. <https://doi.org/10.18653/v1/2023.acl-long.3>
- De Cao, Nicola, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506. <https://doi.org/10.18653/v1/2021.emnlp-main.522>
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Dhuliawala, Shehzaad, 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 arXiv:2309.11495*. <https://doi.org/10.18653/v1/2024.findings-acl.212>
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Su. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.
- Du, Wanyu, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. *arXiv preprint arXiv:2203.03802*. <https://doi.org/10.18653/v1/2022.acl-long.250>
- Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Duan, Jinhao, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*. <https://doi.org/10.18653/v1/2024.acl-long.276>
- Durmus, Esin, He He, and Mona T. 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, ACL 2020*, pages 5055–5070. <https://doi.org/10.18653/v1/2020.acl-main.454>
- Dziri, Nouha, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285. <https://doi.org/10.18653/v1/2022.naacl-main.387>
- Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The BEGIN benchmark. *CoRR*, abs/2105.00071.
- Fabbri, Alexander, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022a. Improving factual consistency in summarization with compression-based post-editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9149–9156. <https://doi.org/10.18653/v1/2022.emnlp-main.623>
- Fabbri, Alexander, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022b. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601. <https://doi.org/10.18653/v1/2022.naacl-main.187>
- Feng, Chao, Xinyu Zhang, and Zichu Fei. 2023. Knowledge solver: Teaching LLMs to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.
- Fernandes, Patrick, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*. https://doi.org/10.1162/tac1_a_00626
- Gao, Leo, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 32 pages.
- Gao, Luyu, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023a. 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)*, pages 16477–16508. <https://doi.org/10.18653/v1/2023.acl-long.910>
- Gao, Tianyu, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*. <https://doi.org/10.18653/v1/2023.emnlp-main.398>
- Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 179–188. <https://doi.org/10.18653/v1/P17-1017>
- Garrido-Muñoz, Ismael, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences*, 11(7):3184. <https://doi.org/10.3390/app11073184>
- Gekhman, Zorik, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784. <https://doi.org/10.18653/v1/2024.emnlp-main.444>
- Goldberg, Yoav. 2023. Reinforcement learning for language models. *Github Blog*.
- Gou, Zhibin, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujia Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.
- Guerreiro, Nuno M., Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*. <https://doi.org/10.1162/tac1.a.00615>
- Guerreiro, Nuno Miguel, Elena Voita, and André F. T. 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, EACL 2023*, pages 1059–1075. <https://doi.org/10.18653/v1/2023.eacl-main.75>
- Gunjal, Anisha, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330.
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Hu, Xuming, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, S Yu Philip, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In the *Twelfth International Conference on Learning Representations*.
- Huang, Yuheng, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023a. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.
- Huang, Zeyu, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023b. Transformer-Patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. <https://doi.org/10.1145/3571730>
- Jiang, Zhengbao, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977. <https://doi.org/10.1162/tac1.a.00407>
- Kadavath, Saurav, 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 preprint arXiv:2207.05221*.
- Kaddour, Jean, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
- Kim, Hyuhng Joon, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang goo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007. <https://doi.org/10.18653/v1/2024.emnlp-main.119>
- Köpf, Andreas, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. OpenAssistant conversations - democratizing large

- language model alignment. *arXiv preprint arXiv:2304.07327*.
- Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 9332–9346. <https://doi.org/10.18653/v1/2020.emnlp-main.750>
- Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*. <https://doi.org/10.18653/v1/2023.findings-emnlp.878>
- Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.
- Lanham, Tamara, 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 preprint arXiv:2307.13702*.
- Lazaridou, Angeliki, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.
- Lee, Ariel N., Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of LLMs. *arXiv preprint arXiv:2308.07317*.
- Lee, Katherine, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.
- Lee, Nayeon, Wei Ping, Peng Xu, Mostafa Patwary, Pascale N. Fung, Mohammad Shoenybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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*, pages 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Li, Huayang, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022a. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.
- Li, Junyi, 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 preprint arXiv:2401.03205*. <https://doi.org/10.18653/v1/2024.acl-long.586>
- Li, Junyi, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*. <https://doi.org/10.18653/v1/2023.emnlp-main.397>
- Li, Junyi, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022b. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*. <https://doi.org/10.24963/ijcai.2021/612>
- Li, Kenneth, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.
- Li, Miaoran, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*. <https://doi.org/10.18653/v1/2024.findings-naacl.12>
- Li, Shaobo, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu.

- 2022c. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732. <https://doi.org/10.18653/v1/2022.findings-acl.136>
- Li, Xingxuan, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023c. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.
- Li, Yifan, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*. <https://doi.org/10.18653/v1/2023.emnlp-main.20>
- Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023e. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Li, Zuchao, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. 2023f. BatGPT: A bidirectional autoregressive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*.
- Lightman, Hunter, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.
- Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Lin, Stephanie, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*. <https://doi.org/10.18653/v1/2022.acl-long.229>
- Lin, Zhen, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.
- Liska, Adam, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622.
- Liu, Fuxiao, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoub, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.
- Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Liu, Jerry. 2022. LlamaIndex. <https://doi.org/10.5281/zenodo.1234>
- Liu, Jiongnan, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023c. RETA-LLM: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.
- Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023d. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*. <https://doi.org/10.1162/tac1a.00638>
- Liu, Tianyu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737. <https://doi.org/10.18653/v1/2022.acl-long.464>
- Liu, Yinhan, 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 preprint arXiv:1907.11692*.
- Llama Team, AI @ Meta. 2024. Llama 3.1 technical report.
- Luo, Junyu, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*. <https://doi.org/10.18653/v1/2024.findings-emnlp.204>
- Luo, Zheheng, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.
- Luo, Ziyang, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*.

- Luu, Kelvin, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! Analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958. <https://doi.org/10.18653/v1/2022.naacl-main.435>
- Maaz, Muhammad, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*. <https://doi.org/10.18653/v1/2024.acl-long.679>
- Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? Assessing the reliability of leading AI legal research tools. *arXiv preprint arXiv:2405.20362*. <https://doi.org/10.1111/jels.12413>
- Mallen, Alex, 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)*, pages 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
- Manakul, Potsawee, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
- McKenna, Nick, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*. <https://doi.org/10.18653/v1/2023.findings-emnlp.182>
- Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.
- Meng, Kevin, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Mialon, Grégoire, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: A survey. *arXiv preprint arXiv:2302.07842*.
- Mikolov, Tomas, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*, pages 1045–1048. <https://doi.org/10.21437/Interspeech.2010-343>
- Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):Article 30. <https://doi.org/10.1145/3605943>
- Min, Sewon, Kalpesh Krishna, Xinxu 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 preprint arXiv:2305.14251*. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- Mitchell, Eric, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831.
- Mohamed, Elaraby, Lu Mengyin, Dunn Jacob, Zhang Xueying, Wang Yu, and Liu Shizhu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.
- Muhlgay, Dor, 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 preprint arXiv:2307.06908*. <https://doi.org/10.18653/v1/2024.eacl-long.4>

- Mündler, Niels, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.
- Murty, Shikhar, Christopher Manning, Scott Lundberg, and Marco Tulio Ribeiro. 2022. Fixing model bugs with natural language patches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11600–11613. <https://doi.org/10.18653/v1/2022.emnlp-main.797>
- Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. 2017. SummaruNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3075–3081. <https://doi.org/10.1609/aaai.v31i11.10958>
- Napoles, Courtney, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234. <https://doi.org/10.18653/v1/E17-2037>
- Navigli, Roberto, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*, 15(2):Article 10. <https://doi.org/10.1145/3597307>
- Ni, Jianmo, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 9844–9855. <https://doi.org/10.18653/v1/2022.emnlp-main.669>
- Ning, Xuefei, Zinan Lin, Zixuan Zhou, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *arXiv preprint arXiv:2307.15337*.
- Onoe, Yasumasa, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? Challenges in propagating injected knowledge. *arXiv preprint arXiv:2305.01651*. <https://doi.org/10.18653/v1/2023.acl-long.300>
- OpenAI. 2023a. ChatGPT. <https://openai.com/blog/chatgpt>
- OpenAI. 2023b. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Pal, Ankit, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*. <https://doi.org/10.18653/v1/2023.conll-1.21>
- Papineni, Kishore, 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*, pages 311–318. <https://doi.org/10.3115/1073083.1073135>
- Parikh, Ankur, 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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. <https://doi.org/10.18653/v1/2020.emnlp-main.89>
- Park, Joon Sung, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*. <https://doi.org/10.1145/3586183.3606763>
- Pauls, Adam and Dan Klein. 2011. Faster and smaller N-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267.
- Penedo, Guilherme, 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 preprint arXiv:2306.01116*.
- Peng, Baolin, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023a. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.
- Peng, Baolin, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.
- Perez, Ethan, Sam Ringer, Kamilė Lukošiuūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*. <https://doi.org/10.18653/v1/2023.findings-acl.847>
- Pu, Xiao, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.
- Qian, Cheng, Xinran Zhao, and Sherry Tongshuang Wu. 2023. "Merge conflicts!" Exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.
- Qiao, Shuofei, Honghao Gui, Huajun Chen, and Ningyu Zhang. 2023. Making language models better tool learners with execution feedback. *arXiv preprint arXiv:2305.13068*.
- Qin, Yujia, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.
- Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897. <https://doi.org/10.1007/s11431-020-1647-3>
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Radhakrishnan, Ansh, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.
- Raffel, Colin, 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(1):5485–5551.
- Ram, Ori, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*. https://doi.org/10.1162/tac1_a_00605
- Ravi, Selvan Sunitha, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*.
- Rawte, Vipula, Prachi Priya, S. M. Tonmoy, S. M. Zaman, Amit Sheth, and Amitava Das. 2023. Exploring the relationship between LLM hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *arXiv preprint arXiv:2309.11064*.
- Rebuffel, Clément, Marco Roberti, Laure Soulier, Geoffrey Scouteeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*. pages 1–37. <https://doi.org/10.1007/s10618-021-00801-4>
- Ren, Ruiyang, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Wang Haifeng. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.
- Roberts, Adam, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. <https://doi.org/10.18653/v1/2020.emnlp-main.437>
- Robertson, Stephen, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389. <https://doi.org/10.1561/15000000019>
- Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel

- Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Schulman, John. 2023. Reinforcement learning from human feedback: Progress and challenges.
- Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Shi, Chufan, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629. <https://doi.org/10.18653/v1/2024.emnlp-main.489>
- Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 31210–31227.
- Shi, Freda, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. 2022. Natural language to code translation with execution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3546. <https://doi.org/10.18653/v1/2022.emnlp-main.231>
- Shi, Weijia, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023b. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*. <https://doi.org/10.18653/v1/2024.naacl-short.69>
- Shi, Weijia, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023c. REPLUG: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*. <https://doi.org/10.18653/v1/2024.naacl-long.463>
- Si, Chenglei, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 to be reliable. *arXiv preprint arXiv:2210.09150*.
- Sinitsin, Anton, Vsevolod Plokhhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345*.
- Su, Yixuan, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.
- Sun, Kai, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023a. Head-to-tail: How knowledgeable are large language models (LLM)? A.K.A. will LLMs replace knowledge graphs? *arXiv preprint arXiv:2308.10168*. <https://doi.org/10.18653/v1/2024.naacl-long.18>
- Sun, Tianxiang, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855.
- Sun, Tianxiang, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023b. MOSS: Training conversational language models from synthetic data. *Machine Intelligence Research*, 21:888–905. <https://doi.org/10.1007/s11633-024-1502-8>
- Tamkin, Alex, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022. Task ambiguity in humans and language models.
- Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca
- Torabi, Faraz, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957. <https://doi.org/10.24963/ijcai.2018/687>
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al.

- 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Umapathi, Logesh Kumar, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.
- Varshney, Neeraj, 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 preprint arXiv:2307.03987*.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wan, Fanqi, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge verification to nip hallucination in the bud. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2633. <https://doi.org/10.18653/v1/2024.emnlp-main.152>
- Wang, Chaojun and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. *arXiv preprint arXiv:2005.03642*. <https://doi.org/10.18653/v1/2020.acl-main.326>
- Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.
- Wang, Hongmin. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322. <https://doi.org/10.18653/v1/W19-8639>
- Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.
- Wang, Yizhong, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? Exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.
- Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508. <https://doi.org/10.18653/v1/2023.acl-long.754>
- Wang, Zhenhailong, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023d. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*. <https://doi.org/10.18653/v1/2024.naacl-long.15>
- Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*.
- Wei, Jason, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Wei, Jerry, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.
- Wei, Jason, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Wei, Jerry, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Long-form factuality in large language models. *CoRR*, abs/2403.18802. <https://doi.org/10.48550/ARXIV.2403.18802>
- Wu, Jian, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linqun Liu, et al.

- 2023a. On decoder-only architecture for speech-to-text and large language model integration. *arXiv preprint arXiv:2307.03917*. <https://doi.org/10.1109/ASRU57964.2023.10389705>
- Wu, Miao, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023b. Do PLMs know and understand ontological knowledge? *arXiv preprint arXiv:2309.05936*. <https://doi.org/10.18653/v1/2023.acl-long.173>
- Wu, Yuhao, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024. LongGenBench: Benchmarking long-form generation in long context LLMs. *arXiv preprint arXiv:2409.02076*.
- Xiao, Yijun 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*, pages 2734–2744. <https://doi.org/10.18653/v1/2021.eacl-main.236>
- Xie, Jian, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.
- Xiong, Miao, 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 preprint arXiv:2306.13063*.
- Xu, Canwen, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*. <https://doi.org/10.18653/v1/2023.emnlp-main.385>
- Xu, Weijia, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023b. Understanding and detecting hallucinations in neural machine translation via model introspection. *CoRR*, abs/2301.07779. <https://doi.org/10.48550/ARXIV.2301.07779>
- Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In the *Eleventh International Conference on Learning Representations*.
- Ye, Qinghao, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Yin, Zhangyue, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*. <https://doi.org/10.18653/v1/2023.findings-acl.551>
- Yu, Jifan, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023a. KoLA: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.
- Yu, Wenhao, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.
- Yue, Xiang, Boshi Wang, Kai Zhang, Zirui Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*. <https://doi.org/10.18653/v1/2023.findings-emnlp.307>
- Zarrieß, Sina, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. *Information*, 12(9):355. <https://doi.org/10.3390/info12090355>
- Zeng, Aohan, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130B: An open bilingual pre-trained model. In the *Eleventh International Conference on Learning Representations*.
- Zeng, Qingcheng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Do reasoning models show better verbalized calibration? <https://doi.org/10.48550/arXiv.2504.06564>
- Zha, Yuheng, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348. <https://doi.org/10.18653/v1/2023.acl-long.634>
- Zhang, Junhao, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye, and Yaowei Zheng. 2025. Lost-in-the-middle in long-text generation: Synthetic dataset,

- evaluation framework, and mitigation. *arXiv preprint arXiv:2503.06868*.
- Zhang, Lunjun, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. <https://doi.org/10.48550/arXiv.2408.15240>
- Zhang, Michael J. Q. and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543. <https://doi.org/10.18653/v1/2025.findings-naacl.306>
- Zhang, Muru, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.
- Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Zhang, Shaolei, Tian Yu, and Yang Feng. 2024. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949. <https://doi.org/10.18653/v1/2024.acl-long.483>
- Zhang, Shuo, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023c. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.
- Zhang, Xuchao, Menglin Xia, Camille Couturier, Guoqing Zheng, Saravan Rajmohan, and Victor Ruhle. 2023d. Hybrid retrieval-augmented generation for real-time composition assistance. *arXiv preprint arXiv:2308.04215*.
- Zhang, Yue, Leyang Cui, Wei Bi, and Shuming Shi. 2023e. Alleviating hallucinations of large language models through induced hallucinations.
- Zhao, Ruochen, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*. <https://doi.org/10.18653/v1/2023.acl-long.320>
- Zhao, Theodore, Mu Wei, J. Samuel Preston, and Hoifung Poon. 2023b. Automatic calibration and error correction for large language models via Pareto optimal self-supervision. *arXiv preprint arXiv:2306.16564*.
- Zhao, Wayne Xin, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*.
- Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023c. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Zhao, Yiran, Jinghan Zhang, I. Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. FELM: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.
- Zhao, Zheng, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *CoRR*, abs/2009.13312. <https://doi.org/10.18653/v1/2020.findings-emnlp.203>
- Zheng, Ce, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*. <https://doi.org/10.18653/v1/2023.emnlp-main.296>
- Zheng, Rui, Shihan Dou, Songyang Gao, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Limao Xiong, Lu Chen, et al. 2023b. Secrets of RLHF in large language models part 1: PPO. *arXiv preprint arXiv:2307.04964*.
- Zheng, Shen, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does ChatGPT fall short in providing truthful answers. *arXiv preprint arXiv:2304.10513*.
- Zhong, Ming, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921. <https://doi.org/10.18653/v1/2021.naacl-main.472>
- Zhong, Zexuan, Zhengxuan Wu, Christopher D. Manning, Christopher

- Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*. <https://doi.org/10.18653/v1/2023.emnlp-main.971>
- Zhou, Chunting, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. LIMA: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.
- Zhou, Chunting, Graham Neubig, Jiatao Gu, Mona T. Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 1393–1404. <https://doi.org/10.18653/v1/2021.findings-acl.120>
- Zhou, Wenxuan, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*. <https://doi.org/10.18653/v1/2023.findings-emnlp.968>
- Zhu, Chenguang, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733. <https://doi.org/10.18653/v1/2021.naacl-main.58>
- Zhu, Kaijie, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*. <https://doi.org/10.1145/3689217.3690621>
- Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, et al. 2023a. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.
- Zou, Andy, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.