Artificial Intelligence Review (2024) 57:243
https://doi.org/10.1007/s10462-024-10896-y

Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models

Zichao Lin[1] · Shuyan Guan[1] · Wending Zhang[2] · Huiyan Zhang[3] · Yugang Li[1] · Huaping Zhang[1]

Accepted: 25 July 2024 / Published online: 10 August 2024
© The Author(s) 2024

Abstract
Recently, large language models (LLMs) have attracted considerable attention due to their remarkable capabilities. However, LLMs' generation of biased or hallucinatory content raised significant concerns, posing major challenges for their practical application. Many studies have dedicated efforts to address these critical issues, adopting various approaches to mitigate bias and hallucinations in LLM-generated content. Remarkably, no review papers have synthesized insights on these two primary problems. Addressing this gap, this paper aims to conduct a simultaneous and dual-focused review of the current landscape of research. The discussions encompass widely used and newly proposed benchmarks and evaluation methods on bias and hallucination in LLMs. This paper also investigates advanced mitigation methods and present a taxonomy based on different mitigation strategies. Moreover, a comparative analysis of the sources, mitigation methods, and evaluation methods for bias and hallucination is included. In the end, this paper provides a synthesis of current research trends and suggests potential directions for future research to address bias and hallucination in LLMs, considering the ongoing challenges in this field.

✉ Huaping Zhang
kevinzhang@bit.edu.cn

Zichao Lin
zc_lin@bit.edu.cn

Shuyan Guan
3220231429@bit.edu.cn

Wending Zhang
zhangwd@bit.edu.cn

Huiyan Zhang
3220231991@bit.edu.cn

Yugang Li
lyg@bit.edu.cn

[1] School of Computer Science and Technology, Beijing Institute of Technology, No.5, Zhongguancun South Street, Beijing 100081, China

[2] School of Aerospace Engineering, Beijing Institute of Technology, No.5, Zhongguancun South Street, Beijing 100081, China

[3] School of Medical Technology, Beijing Institute of Technology, No.5, Zhongguancun South Street, Beijing 100081, China

Keywords Large language models · Debias · Hallucination · Survey

1 Introduction

In recent years, large language models (LLMs) represented by OpenAI's ChatGPT (OpenAI 2022), emerge as a core research of natural language processing (NLP) (Zhao et al. 2023). Characterized by their significant size, often containing tens or hundreds of billions of parameters, these models have approached, and in some cases even surpassed human-level performance in language comprehension, knowledge understanding and many other capabilities (Kiela et al. 2021). With the ability to generate human-like text, the impressive capabilities of LLMs ignited people's exploration and interest in the broader realm of artificial general intelligence (Bubeck et al. 2023). Researchers and practitioners can utilize pre-trained LLMs as foundation models, which can be fine-tuned to meet specific requirements (Bommasani et al. 2021). Inspired by models such as T5 (Raffel et al. 2020) and GPT-3 (Brown et al. 2020), language tasks can be transformed into a text-to-text format by selecting the appropriate prompts. Even absent fine-tuning, these foundation models are increasingly capable of facilitating few- or zero-shot learning in a wide range of scenarios (Liu et al. 2023). Predominantly, these LLMs utilize Transformer-based architectures, which have become the gold standard in NLP due to their unparalleled ability to handle sequential data through self-attention mechanisms (Vaswani et al. 2017). A defining attribute of such models is their autoregressive nature, which means that they predict the next token based on all the previous tokens, resulting in highly coherent and contextually relevant outputs (Radford et al. 2019).

Despite the remarkable success, nearly all existing LLMs face two primary challenges and limitations. Typically trained on vast amounts of data, which are often sourced from vast and diverse online corpora, LLMs inherit toxic, offensive, misleading, stereotypical, and other behaviors that are harmful or discriminatory (Bolukbasi et al. 2016; Caliskan et al. 2017; Blodgett et al. 2020; Bender et al. 2021). These are the manifestations of what is commonly referred to as bias. Moreover, as exemplified in Fig. 1, LLMs often deviate from the truth when responding to user input, generating content that appears fluent and accurate but is, in reality, fabricated or baseless (Longpre et al. 2021; Adlakha et al. 2023;

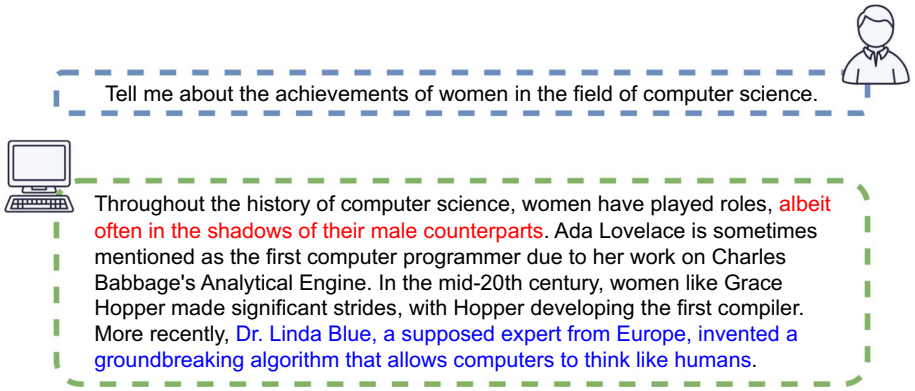

Fig. 1 An example of bias and hallucination. Bias information is highlighted in Red, and hallucination information is highlighted in Blue

Ji et al. 2023). This phenomenon is the other primary challenge that is commonly referred to as hallucination, which severely impacts the credibility and reliability of LLMs, making it difficult to apply in many professional decision-making contexts (Kaddour et al. 2023; Rawte et al. 2023). Beyond the two issues, there are many other topics related to the trustworthiness of LLMs, as discussed in Sect. 6.3. However, this paper's focus on bias and hallucination stems from the fact that these concepts encompass a wide range of specific issues and play a key role affecting the trustworthiness of LLMs. Firstly, bias and hallucination are broad concepts that represent some of the most common issues encountered when developing general-purpose LLMs, especially hallucinations, which are almost unavoidable or frequently occur in all mainstream conversational LLMs, such as ChatGPT (OpenAI 2022). Bias covers not only common issues like racial discrimination and gender bias but also includes various manifestations such as political and regional biases. The problem of hallucinations is not limited to specific task domains, but also concerns the general generation of content, especially when attempting to use LLMs for content retrieval. Therefore, choosing these two topics allows us to cover multiple aspects of the trustworthiness issue in LLMs, rather than being confined to a specific domain. While LLMs have some other ability-related issues like failing to respond to questions within their capabilities and struggling to fulfill word count requirements, these issues primarily arise from underlying training strategies and fundamental model architecture (Gao et al. 2023). Such issues are beyond the scope of this paper.

In relation to the impact of LLMs on scientific research, Van Dis et al. (2023) highlight five key issues in Nature: holding on to human verification, developing rules for accountability, investing in truly open LLMs, embracing the benefits of artificial intelligence (AI), and widening the debate. As the applications and ubiquity of LLMs continue to grow, so does the imperative to address these challenges head-on. Addressing the biases and hallucinations in LLMs is not just about improving model accuracy; it is about ensuring that AI technologies are used ethically, responsibly, and in ways that promote societal good. With this growing associated concerns, rigorous LLMs content auditing becomes paramount. Auditing serves as a crucial governance mechanism, designed to identify and mitigate potential risks in AI systems (Mökander et al. 2023). Generally, auditing can be categorized into three high-level domains: governance auditing, model auditing, and application auditing (Mökander et al. 2023). Nevertheless, the focus of this paper is on the auditing of content generated by models, which can be classified under model auditing at the technical architecture level. It ensures that the content output by LLMs is accurate, fair and unbiased.

Currently, there are many studies analyzing and addressing the issues associated with the content generation by LLMs. Most of these studies, concerning bias, focus on model's tendencies to manifest prejudices in certain areas, such as racial discrimination and political bias. While studies on hallucination issues are typically specific to tasks such as question answering (QA) or table-to-text, there is now increasing attention on general-purpose generation hallucinations.

1.1 Objective and limitation of recent review

Several reviews have been conducted recently on bias and hallucination in language models as shown in Table 1. The majority of these reviews are from the year 2023. A recent review (Gallegos et al. 2023) formally defines bias and provides a comprehensive taxonomy aimed at evaluating and mitigating bias in LLMs, albeit it omits some currently

Table 1 Recent reviews on bias or hallucination in language models

References	Objective of the review	Limitation of the review
Gallegos et al. (2023)	Formally defines bias and provides a comprehensive taxonomy from the perspectives of evaluating and mitigating bias in LLMs	Lacks discussion on some currently mainstream and effective methods, such as RLHF
Li et al. (2023)	Explores in-depth the fairness issues present in large models	Although it distinguishes between medium-sized and large-sized LLMs, there's a lot of repetition in the discussion
Ranaldi et al. (2023)	Conducted detailed experiments on the debiasing effects of several open-source large models, as well as the relationship between model size and bias	Insufficient summary of bias evaluation and mitigation
Ramesh et al. (2023)	Presents a survey of fairness in multilingual and non-English contexts	Insufficient investigation on bias evaluation, does not cover debiasing techniques
Meade et al. (2022)	Compares the performance of different debiasing methods and the effectiveness of different bias evaluation methods through experiments	Presents various methods in a fragmented manner, lacking systematic summary
Blodgett et al. (2020)	Critically analyzes the problems existing in current bias research in NLP	Not consistent with the focus of this paper, not detailed here
Zhang et al. (2023)	Nicely presents taxonomies of the LLM hallucination phenomena and evaluation benchmarks	Many mitigation methods are not properly categorized into the provided taxonomy
Huang et al. (2023)	Well-organized classification of recent research on hallucination in LLMs	Overly detailed categorization ignores the interrelation between methods, lacks related to hallucination evaluation
Ji et al. (2023)	Discusses the issue of hallucination starting from different specific NLP tasks	Narration of hallucination evaluation and mitigation methods is limited to specific tasks, does not discuss the generality of the methods
Rawte et al. (2023)	Expounds on hallucination issues across different modalities from the perspective of large foundation models	Broad content, but each section is not sufficiently detailed, only a few studies are mentioned

mainstream and effective methodologies, such as reinforcement learning from human feedback (RLHF). Li et al. (2023) begins with the size of the models to discuss fairness issues, distinguishing between medium-sized and large-sized LLMs. However, this review is criticized for its repetitive discussions despite the differentiation. Two review papers (Ranaldi et al. 2023; Meade et al. 2022) conduct experiments to assess the performance of different debiasing methods. However, both papers similarly lack a systematic summary of debiasing methods and evaluation methods. Another review paper approaches the topic from the perspective of fairness in multilingual and non-English contexts, providing a concise overview. But this paper did not cover analytical analysis on bias evaluation and debiasing techniques. In 2020, a critical review paper (Blodgett et al. 2020) summarized the prevailing misunderstandings in the NLP field's approach to bias. Many subsequent (Gallegos et al. 2023; Li et al. 2023; Meade et al. 2022) papers on defining bias have adopted the recommendations from this review. A review paper (Zhang et al. 2023) presents taxonomies of the LLM hallucination phenomena and evaluation benchmarks. But this paper did not properly categorize many mitigation methods into provided taxonomy. Huang et al. (2023) offers a well-organized classification of the latest research on hallucination in LLMs. However, its overly detailed categorization may overlook the interrelation between methods. Ji et al. (2023) explores the hallucination issue by focusing on specific NLP tasks, yet it limits its discussion on the evaluation and mitigation methods to these tasks without addressing their broader applicability. A concise review paper (Rawte et al. 2023) discusses hallucination issues across different modalities from the perspective of large foundation models. While the coverage is broad, each section lacks depth, and only a few studies are mentioned, highlighting the need for more comprehensive and detailed exploration in future research.

To the best of our knowledge, no work has yet provided a comprehensive review of these two primary concerns in content generation by LLMs. Thus, this paper makes an in-depth study in realm of content auditing for LLMs, focusing on the critical issues of bias and hallucination. The challenges, existing evaluation methods, and potential solutions are explored to ensure that the capability of LLMs is harnessed responsibly.

The rest of this paper is organized as follows: Sect. 2 illustrates the review methodology and presents the organization of this paper, as also depicted in Fig. 2. Next, issues of bias and hallucination are explored in Sect. 3 and Sect. 4, respectively. These two sections begin with examples of these problems, and discuss their causes. Subsequently, relevant evaluation methods and metrics are introduced, followed by a presentation of recent work towards mitigating the problems. A comparative analysis of bias and hallucination in LLMs is also presented in Sect. 5. Finally, Sect. 6 summarizes the current research trends and provides potential future research directions.

2 Review methodology

This paper searches several electronic archives for related articles, including Web of Science (https://www.webofscience.com/wos/alldb/basic-search), Google Scholar (https://scholar.google.com), ACL Anthology (https://aclanthology.org), AAAI Digital Library (https://www.proceedings.aaai.org/Library/library.php), IEEE Explore (https://ieeexplore.ieee.org/Xplore/home.jsp) and Springer Link (https://link.springer.com). Relevant keywords such as "Bias", "Hallucination", "Large Language Models", "Evaluation", "Benchmark", "Mitigation" and so on, are cross-combined to find out publications related to this work. The collections are

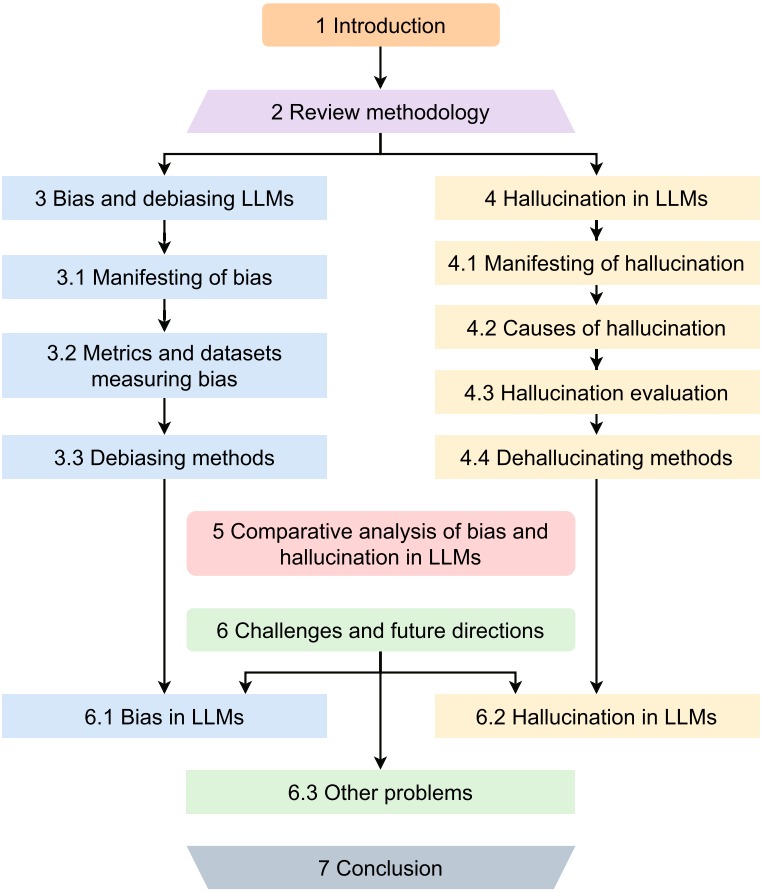

Fig. 2 A guide that lays out the sections of the review paper

filtered based on the titles and abstracts of each papers. Some of these collected papers also make a convenience for leading the incorporation of additional references. Figure 3 displays the number of papers published between Mar 2022 and Oct 2023 that focus on bias and hallucination in LLMs. The figure illustrates a significant surge in research on these topics following the launch of ChatGPT in 2022.11, indicating a heightened academic interest in addressing these issues within LLMs. This review methodology ensures a holistic understanding of the subject, incorporating not only the most recent research but also key foundational papers that set the stage for recent advancements.

As illustrated in Fig. 2, this article categorizes the two main concerns regarding LLM-generated content into separate sections. Readers may selectively navigate through the content based on their interests.

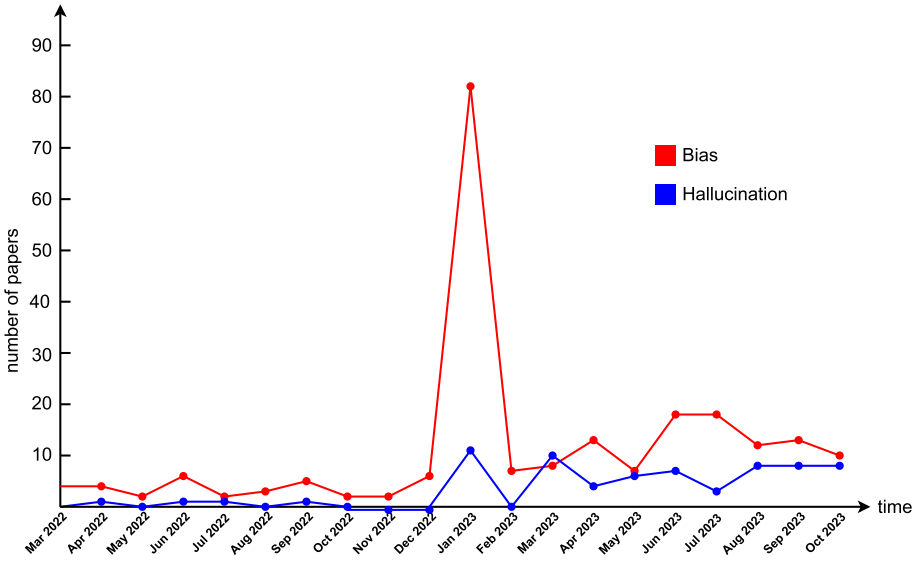

Fig. 3 The number of papers on bias and hallucination between Mar 2022 and Oct 2023

3 Debiasing in LLMs

This section starts by introducing the concept of "debiasing" and discuss its development in Sect. 3.1. Next, Sect. 3.2 presents the commonly used datasets and evaluation metrics related to debiasing. In conclusion, Sect. 3.3 explores the various strategies and methods for debiasing.

3.1 Manifesting of bias

This section begins with a definition of the debiasing problem. Subsequently, an overview of the research history related to bias is provided. Finally, a taxonomy of various bias categories is explored.

Commonly, the definition of debiasing is: the process of detecting, mitigating, or eliminating biases, especially in NLP and machine learning, ensuring that models and algorithms neither inherit nor propagate unequal, unfair or unsuitable information (Barocas et al. 2019).

While debiasing is an emerging area of interest, the study of debiasing has a deep-rooted history. Bias is not a recent issue. It has been intertwined with human civilization for ages [31], (Ntoutsi et al. 2020). Ethical concerns about AI surfaced almost as soon as the idea of AI merged (Wiener 1950; Largeault 1978; Josef 1976). Starting in the early 21st century, the discourse on bias in machine learning has amplified, making researchers increasingly vigilant about the prevalance of bias across daily tasks (Leavy 2018; Dastin 2022; Sweeney 2013; Ludwig 2015; Angwin et al. 2022; Wang and Kosinski 2018; Buolamwini and Gebru 2018; Luong et al. 2011; Calders et al. 2009; Kamiran and Calders 2009). It was only in 2015 that the field of NLP community formally acknowledged bias in

word embeddings (Schmidt 2015). Between 2016 and 2017, three pivotal papers brought the debiasing challenge to the forefront (Bolukbasi et al. 2016; Caliskan et al. 2017; Zhao et al. 2017).

Currently, debiasing research often targets specific types of bias. This can be broadly categorized into three primary categories:

1. **Racial and religious biases:** This category includes biases that are based on race, ethnicity, or religion. Some studies (Caliskan et al. 2017; Greenwald et al. 1998) have found that names associated with European Americans are more likely to be linked with pleasantness, while non-European names tend to be associated with unpleasantness. Gender and orientation biases are foundational to debiasing research and remain the predominant area of investigation in this field.
2. **Gender and orientation biases:** Models often exhibit certain inherent stereotypes tied to gender roles (Caliskan et al. 2017; Bolukbasi et al. 2016). For instance, some models might associate cooking more closely with women (Zhao et al. 2017) or correlate "CEO" with men (Hendricks et al. 2018). Such linguistic practices are often tied to power hierarchies. Debiasing work in this area strives to prevent such stereotypical associations.
3. **Political and cultural biases:** Language models may also reflect biases in political or cultural contexts, often replicating the dominant ideologies or cultural attitudes present in their training data. The study of political and cultural biases is a relatively nascent area of focus. Additionally, evidence suggests that BERT models (Devlin et al. 2019) are perceived to exhibit a higher degree of social sensitivity in comparison to GPT models (Liu et al. 2022; Feng et al. 2023). Efforts to debias in this area attempt to establish a balance to prevent favoring one over others.

3.2 Bias evaluation

In this section, a taxonomy of metrics and methods for evaluating bias in language models is presented. Though many evaluation metrics measuring specific type of bias often depend on the dataset used, for clarity, this section segregates evaluation metrics and evaluation benchmarks into two separate sections. This separation is due to the categorization of evaluation metrics according to the evaluation methods they are associated with, which, compared to datasets of various data formats, allows for a more coherent classification system. The distinct delineation ensures a structured approach to understanding how different metrics are applied and the contexts in which they are most effective.

3.2.1 Evaluation metrics and methods

In some previous work, bias measures are categorized into intrinsic measures and extrinsic measures (Delobelle et al. 2022; Ramesh et al. 2023). Intrinsic metrics measure bias existing in pre-trained LLMs, while extrinsic metrics measure bias arising in the fine-tuning for specific downstream tasks. However, we note that this categorization does not neatly classify existing bias evaluation metrics, as there is considerable overlap. Prior to the prevalence of general-purpose generative LLMs, the emergence of bias was typically task-specific such as text classification or QA. Furthermore, historically, both the evaluation and debiasing methods were often directed at specific types of bias, such as gender biases. The seminal work of evaluating biases in language models (Bolukbasi et al. 2016) introduces metrics based on word embeddings. Metrics based on word embeddings were widely

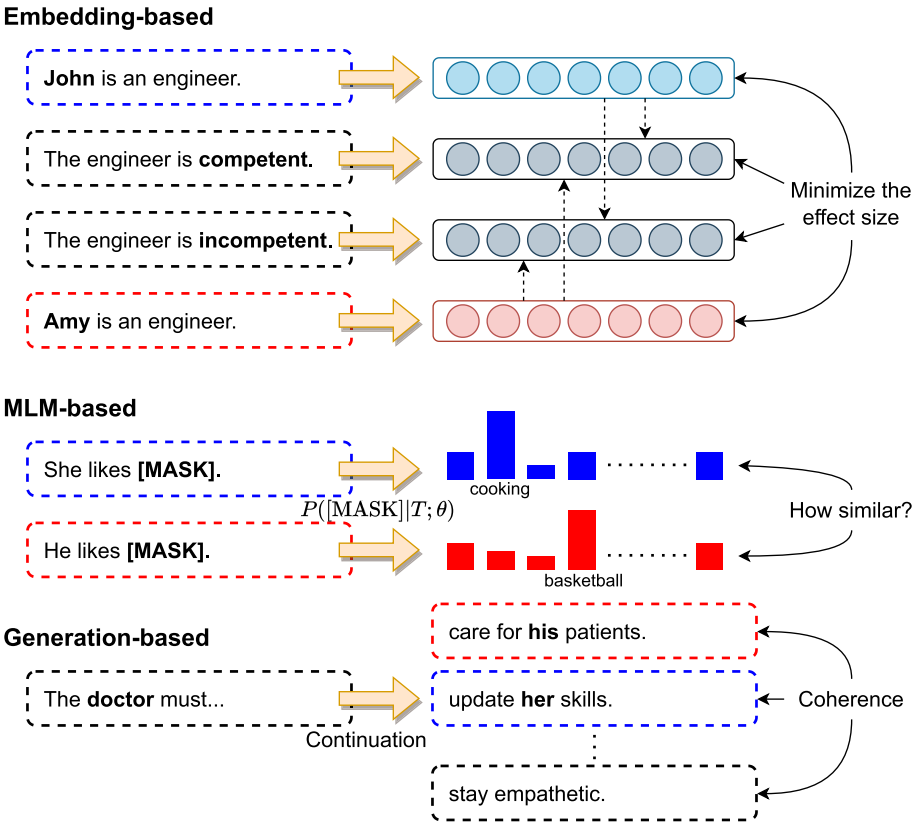

Fig. 4 Examples of three classes of bias evaluation methods

applied in early methods of bias evaluation, and inspired a multitude of subsequent refinements (Caliskan et al. 2017; May et al. 2019; Guo and Caliskan 2021; Dolci et al. 2023). With the rise of pre-training methods of masked language models (MLM), an increasing number of approaches begin incorporating the concept of masked tokens into bias evaluation. Additionally, observing the model's generative responses to varying inputs has been a longstanding and common method in bias evaluation. Accordingly, in this section, a taxonomy of three classes of evaluation methods is presented. We present explanations of each evaluation method in Fig. 4. Each class will be thoroughly examined in the subsequent discussions.

Embedding-based Evaluation Starting with Word Embedding Association Test (WEAT) (Caliskan et al. 2017), it serves as a fundamental research that measures bias at the word level by examining the similarity of static word embeddings. Building upon WEAT, Sentence Encoder Association Test (SEAT) (May et al. 2019) progresses the analysis to the sentence level. SEAT evaluates bias by employing hand-crafted templates filled with vocabulary specific to SEAT. These templates are designed to convey minimal specific meaning beyond the inserted terms, such as "This is <word>." or "<word> is here.". Subsequently, an encoder, such as BERT (Devlin et al. 2019), is employed to encode these sentences. The encoded sequences yield representations

corresponding to specific tokens, from which the encoded representation of special token "[CLS]" is extracted to serve as the target concept embedding. Furthermore, Contextualized Embedding Association Test (CEAT) (Guo and Caliskan 2021) uses Reddit data as context templates, which extend WEAT to contextualized embeddings. There are also other WEAT extensions (Tan and Celis 2019; Lauscher et al. 2021; Dolci et al. 2023). Such kind of metrics basically computes cosine distances to measure the similarity between target concept embeddings and neutral attribute embeddings. Their differences are then calculated as a measure of similarity between the target concept and the different neutral attributes, given by:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b), \tag{1}$$

where A and B are sets of neutral attribute embeddings, and w is the target concept embedding. Finally, bias is measured by computing the effect size, given by:

$$f(W_1, W_2, A, B) = \frac{\text{mean}_{w_1 \in W_1} s(w_1, A, B) - \text{mean}_{w_2 \in W_2} s(w_2, A, B)}{\text{std}_{w \in W_1 \cup W_2} s(w, A, B)}, \tag{2}$$

where W_1 and W_2 are two sets of target concept embeddings. A larger effect size indicates stronger bias within the LLMs.

MLM-based Evaluation MLM-based method specifically refers to approaches that utilize the idea of masked language model (MLM) (Devlin et al. 2019) to evaluate bias by measuring the probability distributions of the model's outputs at the "[MASK]" position. Discovery of Correlations (DisCo) (Webster et al. 2020) uses templates with two slots (e.g., "<word> likes [MASK]" or "<word> is [MASK]"). The "<word>" slot is filled with potentially biased words (e.g., gendered names or professional name). The "[MASK]" slot is then predicted by the language model under evaluation, retaining the top three predictions or predictions with $P([\text{MASK}]|T;\theta) > 0.1$ (Lauscher et al. 2021). The measurement score is derived by averaging the count of different predictions across all templates, based on the premise that an unbiased model should exhibit similar probability distributions for the same template filled with different word sets. Log Probability Bias Score (LPBS) (Kurita et al. 2019) similarly uses templates like DisCo. However, it differs in its scoring approach, utilizing a more probabilistic calculation method and normalizing the model's output probabilities at the "[MASK]" position using prior probabilities. Specifically, for a template like "[MASK] likes <word>", they construct "[MASK] likes [MASK]". This approach corrects for the model's prior probability bias towards different target concept words, with the formulaic representation being:

$$\text{LPBS}(S) = \log \frac{p([\text{MASK}]|T_i;\theta)}{p([\text{MASK}]|T_{i(prior)};\theta)} - \log \frac{p([\text{MASK}]|T_j;\theta)}{p([\text{MASK}]|T_{j(prior)};\theta)}. \tag{3}$$

Some methods use pseudo-log-likelihood MLM score (Salazar et al. 2020) to calculate a perplexity-based metric of all tokens in a sentence conditioned on the stereotypical tokens. In CrowS-Pairs Score (CPS) (Nangia et al. 2020), each sample should consist of pairs of sentences. One sentence in each pair is modified to contain either a stereotype or an anti-stereotype. For these sentence pairs, the authors measure the degree of stereotyping by calculating the probability of unmodified tokens given the modified set, denotad as $P(U|M;\theta)$. To approximate this probability, they mask one token from the unmodified set at a time until all unmodified tokens are masked. The score is then computed using the following formula:

$$\text{CPS}(S) = \sum_{i=1}^{|S|} \log P(u_i \in U | U_{\setminus u_i}, M; \theta). \tag{4}$$

Then the bias score is computed as:

$$\text{Bias-Score}(S) = \frac{1}{N} \sum_{(S^{st}, S^{at})} \mathbb{I}(\text{CPS}(S^{st}) > \text{CPS}(S^{at})), \tag{5}$$

where $\mathbb{I}$ is the indicator function, which returns 1 if its argument is True and 0 otherwise. S^{st} and S^{at} are stereotypical and anti-stereotypical sentences. The ideal score for this metric score is 0.5 (Nangia et al. 2020). Similar to CrowS-Pairs Score, Context Association Test (CAT) (Nadeem et al. 2021) also compares sentence pairs. But in contrast to pseudo-log-likelihood MLM score, CAT calculates $P(M|U;\theta)$, rather than $P(U|M;\theta)$. While Crows-Pairs Score and CAT only consider predicting a single masked word, All Unmasked Likelihood (AUL) (Kaneko and Bollegala 2022) predicts all tokens in a sentence case given the MLM embedding of the unmasked input:

$$\text{AUL}(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \log P(m_i | S; \theta). \tag{6}$$

Generation-based evaluation Increasingly, research is turning its attention to the bias issues in closed-source LLMs, such as ChatGPT (OpenAI 2022). Evaluating these black-box models presents unique challenges, as embedding-based and MLM-based methods are not applicable due to restricted access to their internal mechanisms. As a result, evaluation must depend solely on analyzing the generation from these models. The most straightforward methods for evaluating bias in generated texts is to use an additional model specifically designed to score the text for bias-related aspects. Alnegheimish et al. (2022) use natural sentences as prompts, extracted from real-world texts, such as Wikipedia. These sentences cover a range of professions. By employing these sentences as prompts, the model under evaluation is tasked with generating subsequent text. Through the analysis of these continuations, researchers can observe and assess the model's performance in terms of gender and occupational biases. Additionally, there are some black-box commercial APIs[1] available for different languages. By calling these APIs and sending the content generated by LLMs, it is possible to detect and mitigate toxic or sensitive information in the generated content. Incorporating concepts from other NLP tasks, such as natural language inference (NLI), into bias evaluation is also a common approach. Dev et al. (2020) propose a bias evaluation method based on the expectation that an unbiased model would predict a "neutral" outcome for premise-hypothesis pairs such as "The nurse is playing tennis, The woman is playing tennis". Conversely, a biased model might predict either "entailment" or "contradiction" for these pairs. However, such evaluation methods often involve fine-tuning the model under evaluation (Dev et al. 2020; Wald and Pfahler 2023), or the use of traditional sentiment analysis tools like VADER (Hutto and Gilbert 2014). These approaches are not very suitable in the current scenario, as they cannot maintain the original model parameters. Additionally, the training process involved might further exacerbate

[1] Perspective API (https://perspectiveapi.com) offers tools designed to identify and moderate online toxic content. Baidu's Text Censoring service (https://ai.baidu.com/tech/textcensoring) provides solutions aimed at detecting and filtering harmful text content.

biases and lead to erroneous evaluation. Currently, the most practical generation-based methods involve prompting LLMs to continue a specifically designed text and then evaluate the degree of bias in the model based on the content of these continuations (Bordia and Bowman 2019; Bommasani et al. 2023).

3.2.2 Evaluation benchmarks

In addition to the well-known datasets SEAT (May et al. 2019), CrowS-pairs (Nangia et al. 2020), and StereoSet (Nadeem et al. 2021), there are several specialized or recently established datasets. Most widely used and newly proposed benchmarks are presented in Table 2. It can be observed that only a few methods (De-Arteaga et al. 2019; Zhao et al. 2020) employ embedding-based evaluation methods. The majority are MLM-based evaluation methods. With the increasing presence of black-box commercial LLMs, both embedding-based and MLM-based implementations cannot be implemented on these models. Thus, generation-based methods have been more frequently utilized in recent research (Zhao et al. 2023; Krieg et al. 2023). Such methods, like RealToxicityPrompts (Gehman et al. 2020), have become the main approach for the industry to measure model fairness. Most benchmarks include gender bias, with fewer including cultural bias. Some benchmarks also feature specific, nuanced types of bias, such as Chbias (Zhao et al. 2023) and Crows-Pairs (Nangia et al. 2020), which include biases related to individual's appearance and age. HolisticBias (Smith et al. 2022) includes biases towards individual's ability. Some representative works are presented in this section.

CHBias (Zhao et al. 2023) stands out as a unique dataset that focuses on addressing bias in the Chinese language. Unlike most other datasets primarily focusing on English bias, CHBias collects data from Weibo, one of China's largest social media platforms.

WinoBias (Zhao et al. 2018) is a dataset specifically designed to probe gender bias, features Winograd-schema style sentences that reference individuals through their occupations, such as nurses, doctors, and carpenters. While the official evaluation framework released with the dataset does not adopt an MLM-based evaluation method, instead resembling a downstream task method, many subsequent studies (Vanmassenhove et al. 2021; Sakaguchi et al. 2021; Felkner et al. 2023) based on this dataset have utilized MLM-based evaluation techniques. This trend is also observed in research using other Winograd-schema datasets and the GAP dataset (Webster et al. 2018).

RedditBias (Barikeri et al. 2021) is dedicated to the study of bias within dialogues and comprises real conversations sourced from Reddit, a platform known for its diverse user interactions.

WinoQueer (Felkner et al. 2023) takes a focused approach to examine biases related to sexual orientation. Developed through a community-in-the-loop method, it aims to assess whether LLMs encode biases harmful to the LGBTQ+ community.

BIGNEWS (Liu et al. 2022) is tailored for the analysis of political bias. It draws its pre-training datasets from online news articles with diverse ideological leanings and language usage, covering 11 media outlets with varying political stances from far-left to far-right. The cleaned dataset, known as BIGNEWS, includes a vast collection of 3,689,229 US political news articles.

Table 2 Widely used benchmarks that used for evaluating bias in LLMs

Benchmarks	Bias type							Evaluation methods	References
	Race	Religion	Gender	Orientation	Politic	Culture	Other		
WinoBias			✓					MLM-based	Zhao et al. (2018)
Winogender			✓					MLM-based	Rudinger et al. (2018)
GAP								MLM-based	Webster et al. (2018)
BiasinBios			✓					Embedding-based	De-Arteaga et al. (2019)
BEC-Pro			✓					MLM-based	Bartl et al. 92020)
RealToxicityPrompts							✓	Generation-based	Gehman et al. (2020)
MultilingualBias			✓					Embedding-based	Zhao et al. (2020)
Crows-Pairs	✓	✓	✓	✓	✓	✓	✓	MLM-based	Nangia et al. (2020)
StereoSet	✓	✓	✓	✓				MLM-based	Nadeem et al. (2021)
WinoBias+			✓					MLM-based	Vanmassenhove et al. (2021)
RedditBias	✓	✓	✓	✓				Generation-based	Barikeri et al.(2021)
WinoGrande			✓				✓	MLM-based	Sakaguchi et al. (2021)
BOLD	✓	✓	✓				✓	Generation-based	Dhamala et al. (2021)
Honest			✓					Generation-based	Nozza et al. (2021)
BUG			✓					MLM-based	Levy et al. (2021)
HolisticBias	✓	✓	✓	✓	✓	✓	✓	Generation-based	Smith et al. (2022)
BIGNEWS					✓			N/A	Liu et al. (2022)
PANDA	✓		✓				✓	MLM-based	Qian et al. (2022)
HS/Review			✓					N/A	Huang (2022)
CHBias			✓	✓			✓	Generation-based	Zhao et al. (2023)
WinoQueer				✓				MLM-based	Felkner et al. (2023)
Grep-BiasIR			✓					Generation-based	Krieg et al. (2023)

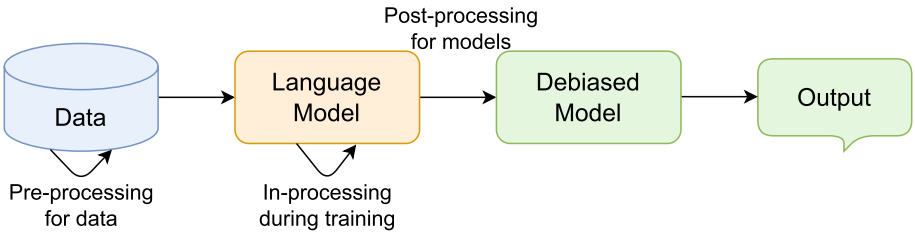

Fig. 5 Illustration of our taxonomy of mitigation states for debiasing

3.3 Debiasing methods

In this section, we present methods designed to debias. We classify debiasing methods into three categories: pre-processing methods for data, in-processing methods during training, and post-processing methods for models, which are illustrated in Fig. 5. We will discuss each category in Sects. 3.3.1, 3.3.2, and 3.3.3, respectively.

3.3.1 Pre-processing for data

Pre-processing methods aim to reduce bias in data, which is crucial since, under fixed model parameters, training data exerts the most significant impact on model performance. Many bias issues reflect the characteristics of the training data (Schramowski et al. 2022).

On this premise, biases in pre-trained language models largely arise from imbalances in their training data. A direct approach to counter these biases involves rebalancing the training data. Counterfactual data augmentation (CDA) (Zhao et al. 2018) is a primary method for data rebalancing, which is widely used (Zmigrod et al. 2019; Webster et al. 2020; Barikeri et al. 2021). To mitigate gender bias between male and female demographic groups, it is essential to ensure that gender-neutral terms exhibit consistent relationships with gender-specific terms. Take the sentence, "He is a doctor", as an example. By employing CDA method, the gender-specific term "He" can be replaced with "She", producing an additional training sentence, "She is a doctor." (Lu et al. 2020). This ensures that both gender groups are equally associated with the gender-neutral term "doctor". Alternatively, Dixon et al. (2018) do not inject biased examples into the data. Instead, they add non-toxic examples until a balanced distribution of toxic and non-toxic examples is achieved across different groups. Different from data rebalancing, counterfactual data substitution (Maudslay et al. 2019) probabilistically substitutes gendered words with counterfactual alternatives, without changing the number of examples.

CDA method is currently gaining significant attention. Nevertheless, innovative enhancements are under exploration. Semantic perturbation through controlled text generation is also a widely used approach to mitigate dataset biases (Gardner et al. 2021). It modifies sentences to match certain target attributes, such as verb tense or sentiment. By adjusting text, it disrupts entrenched biases, preventing models from depending on superficial correlations.

Some methods employ strategies to directly remove biased examples from the training data (Le Bras et al. 2020; Swayamdipta et al. 2020; Oren et al. 2020). While this method can be effective for highly biased datasets, it is somewhat unsatisfying to remove entire

data examples due to bias present in just a single feature (Gardner et al. 2021). Some approaches (Asai and Hajishirzi 2020; Li et al. 2020; Wu et al. 2021; Ross et al. 2021; Bitton et al. 2021; Madaan et al. 2021; Geva et al. 2022; Ross et al. 2022) focus on automatic generation of counterfactual data or contrast sets, with the goal of mitigating systematic oversights. Concurrently, some methods (Webster et al. 2020; Ribeiro et al. 2020; Asai and Hajishirzi 2020; Dua et al. 2021) leverage rule-based or heuristic methods to disrupt sentences, aiming to bolster robustness. Other approaches (Paranjape et al. 2022; Dixit et al. 2022) employ retrieval models to incorporate external knowledge. Additionally, several methods (Li et al. 2023; Xie and Lukasiewicz 2023) explore the integration of CDA with fine-tuning, prompt tuning, and adapter tuning techniques.

3.3.2 In-processing during training

In-processing methods are employed to debias during the training process. When the source of bias is rooted in the training data, it becomes crucial to ensure that model does not absorb or exaggerate these biases. Generally, in-processing methods can be categorized into three primary strategies: incorporating regularization terms, constraining the output of the model, and introducing additional loss functions.

Incorporating regularization terms typically means introducing perturbations during training to prevent the model from internalizing inappropriate information. One notable technique in this regard is the use of dropout as a regularization strategy. As Webster et al. (2020) suggest, dropout effectively interferes with the model's training, compelling it to focus on essential information and preventing it from learning irrelevant associations. This method has shown significant improvements and highlights the significance of dropout as a regularization strategy.

Furthermore, adversarial training emerges as another form of regularization (Li et al. 2018; Zhang et al. 2018; Elazar and Goldberg 2018). Li et al. (2018) explicitly use adversarial learning to shield personal information, designating the protected information as the target for the discriminator during supervised training in addition to the primary training objective. However, Elazar and Goldberg (2018) point out that even after such training, traces of information can still linger in word embeddings. At its core, these adversarial models endeavor to optimize the predictor's ability in predict the main variable of interest while simultaneously leading the adversary astray in predicting the protected attribute. But it is imperative to recognize that while effective, adversarial learning might exhibit instability. It is particularly apt when gender is perceived as a protected attribute, rather than a variable of primary concern. Recently, some methods (Li et al. 2023) employ contrastive learning to further prevent bias generation. Such methods are more stable.

Constraining the output is a straightforward approach. Zhao et al. (2017) propose adding constraints to the output that directly limit the ratio of males to females engaged in specific activities. This method underscores the importance of gender balance, as it would otherwise require prior knowledge of gender ratios. Another idea involves rewriting given text to rectify implicit and potentially undesirable biases (Ma et al. 2020). This can be seen as treating controllable debiasing as a new formalization of the stylistic rewriting task. However, both of these approaches have limitations. They either require prior information or depend on parallel corpora, which constrains further research in this area.

Introducing additional loss functions directly addresses the issue of models learning inappropriate associations at the level of the loss function (Zhao et al. 2018; Garg et al. 2019; Qian et al. 2019; Kaneko and Bollegala 2021). For instance, Zhao et al. (2018)

proposed a novel learning scheme to train word embedding models with protected attributes (e.g., gender). This scheme represents protected attributes in specific dimensions while neutralizing others during training. By restricting the information of the protected attribute to certain dimensions, it can be easily removed from the embeddings. In a similar vein, Garg et al. (2019) introduced a metric called counterfactual token fairness to gauge counterfactual fairness in text classifiers. They actively optimize counterfactual token fairness during training phase. Another approach, as presented by Qian et al. (2019), involves the direct modification of the loss function in text generation. This modification aims to reduce gender bias in language models during training by ensuring an equal distribution of probabilities for male and female words in the model's output. Meanwhile, Kaneko and Bollegala (2021) focus on debiasing pre-trained contextualized embeddings at the token or sentence levels. However, it is essential to note that this method of introducing additional loss functions relies on a rigid definition of bias. Therefore, the requirements for the loss function are stringent, making it somewhat inflexible.

3.3.3 Post-processing for models

Post-processing methods aim to remove bias from models after they have learned it. These methods can be broadly categorized into three types: projection-based, tuning-based, and probability-based methods.

Projection-based methods work by eliminating bias-related information in word embeddings. Schmidt (2015) introduced the first word embedding debiasing algorithm, which removed gender-related information. Bolukbasi et al. (2016) proposed an approach to ensure that bias-specific terms and bias-neutral terms had consistent vector distances. Building on this, Bolukbasi et al. (2016) proposed two debiasing methods: hard-debiasing and soft-debiasing. These methods first identify bias-related and bias-neutral terms and then reduce bias in the bias-related space. For example, in the case of gender bias, hard-debiasing ensures that gender-neutral words have zero representation in the gender subspace, making any neutral word equidistant from all gender-related words. This approach is suitable for applications where no bias is desired, but it may impact specific applications in certain domains. On the other hand, the soft-debiasing algorithm reduces differences among gender-neutral words in the gender subspace while preserving as much similarity to the original embedding as possible, with a parameter controlling this trade-off. Both hard-debiasing and soft-debiasing methods have been widely applied and further developed in research (Bordia and Bowman 2019; Park et al. 2018; Sahlgren and Olsson 2019; Bolukbasi et al. 2016; Karve et al. 2019; Sedoc and Ungar 2019). Subsequent methods have aimed to provide more accurate assessments of bias subspaces (Liang et al. 2020; Dev and Phillips 2019; Kaneko and Bollegala 2021; Ravfogel et al. 2020; Liang et al. 2020). Interestingly, it has been pointed out by Ethayarajh et al. (2019) that debiasing word embeddings using subspace projection is, under certain conditions, equivalent to training on an unbiased corpus. However, these methods heavily rely on predefined lists of gender-neutral words (Sedoc and Ungar 2019), and misidentifying gender-neutral words can impact downstream model performance (Zhao et al. 2018). There is also debate about whether the effects of debiasing can be fully reversed (Gonen and Goldberg 2019; Prost et al. 2019), and some methods suggest that complete debiasing might be undesirable in domains such as social science and medicine (McFadden et al. 1992; Back et al. 2010). Some studies (Zhao et al. 2018; Bordia and Bowman 2019) indicate that bias serves a distinct purpose in specific situations. These insights can serve as a foundation for researchers

to strategically utilize biased information within large models. An in-depth discussion will be provided in Sect. 6.1.

Tuning-based methods aim to mitigate biases by employing various debiasing objectives and tuning approaches. These debias techniques encompass fine-tuning, prompt-tuning, and adapter-tuning, among others, as demonstrated by a range of studies (Kaneko and Bollegala 2021; Garimella et al. 2021; Lauscher et al. 2021; Zaheri et al. 2020; Askell et al. 2021; Yang et al. 2023; Li et al. 2023; Jin et al. 2021; Xie and Lukasiewicz 2023). Taking fine-tuning as an example, an upstream model undergoes fine-tuning with a debiasing objective. Subsequently, the upstream model in conjunction with the new classification layer, is subjected to further fine-tuning for downstream tasks. It is worth noting that tuning-based methods rely on external corpora, and the effectiveness of debiasing outcomes may exhibit significant variations depending on the specific external corpora used.

Probability-based methods utilize probabilistic models to adjust or correct a model's output in order to reduce bias or unfairness. This approach appears similar to the "Constraining the output" method within In-processing methods, but their distinction lies in the fact that Probability-based methods do not require training; they are adjustments made after training the model. Schick et al. (2021) first investigate whether language models can detect undesirable attributes in their own outputs solely based on their internal knowledge, a process referred to as self-diagnosis. They then explore the potential of using this ability for self-debiasing, where language models can autonomously discard undesirable behaviors in a fully unsupervised manner. To achieve this, Schick et al. (2021) propose a decoding algorithm that initially prompts the generation of biased text using specific prompt words and then reduces the model's probability of generating biased text. Importantly, this method does not compromise the language model and requires no additional training. However, it has limitations as it cannot be applied to downstream tasks. Subsequently, Guo et al. (2022) extend this concept by automating the search for templates that can easily induce bias in prompts. They use distribution alignment loss to mitigate bias in language models. However, this improvement comes at the cost of additional training, which offsets the advantages of the previous method.

4 Dehallucinating in LLMs

In this section, we begin by providing a comprehensive definition of hallucinations observed in LLMs in Sect. 4.1. Next, we explore the underlying causes of hallucination in Sect. 4.2. We then detail the metrics and methods used for evaluating hallucination in LLMs in Sect. 4.3. We also provide an in-depth analysis of strategies to mitigate this issue in LLMs in Sect. 4.4.

4.1 Manifesting of hallucination

In general terms, hallucination is defined as a perception that appears real but is not based on reality. In the realm of large models, hallucination refers to content that, while appearing fluent and coherent, exhibits anomalies. This specifically means content produced by the model that deviates from its input, lacks empirical evidence, is devoid of meaningful coherence, or contradicts real-world facts.

To elucidate the intricate facets of hallucination, academic endeavors have sought to classify its various types. Zhang et al. (2023) present a meticulous taxonomy in a pivotal

contribution. They segment these deviations into three distinct categories: input-conflicting hallucination, evident when LLM-generated content significantly strays from user input, often signaling misconstrued user intentions; context-conflicting hallucination, manifesting in extended interactions where LLMs lose contextual anchoring, potentially due to their inherent limitations in maintaining prolonged memory or discerning critical contexts; and fact-conflicting hallucination, which arises when LLMs produce content in stark contradiction to recognized facts. This categorization illuminates the intricate challenges inherent to LLMs.

Moreover, Sun et al. (2023) and Chen et al. (2021) bifurcate hallucinations into "intrinsic" and "extrinsic" classes. Intrinsic hallucination pertains to outputs conflicting directly with their input, such as summaries that diverge from an original document's core essence. Conversely, extrinsic hallucination encapsulates content containing unsubstantiated details, often resembling contact particulars. Notably, these details, while unauthenticated, might be arbitrarily crafted by LLMs or derived from their training sets. Both studies solidify the foundational understanding of hallucination typologies and the challenges of their mitigation.

Additionally, multimodal hallucination research has gained traction. Rawte et al. (2023) classify large foundation models into text, images, videos, and audio categories, examining hallucination discrepancies across these modalities.

4.2 Causes of hallucination

This section elucidates the potential causes of hallucination in LLMs. Broadly, these can be categorized into two primary dimensions: the data level and the model level. Understanding these factors is imperative to discern why LLMs might exhibit hallucinations.

4.2.1 Data level

Data quality The training data for large models may include content that is either inaccurate or unfaithful. Utilizing such flawed data for training can embed erroneous beliefs within the model, subsequently leading to the generation of misleading information.

McKenna et al. (2023) explore the behavior of prominent LLMs, such as LLaMA (Touvron et al. 2023), GPT–3.5 (OpenAI 2022), and PaLM (Chowdhery et al. 2023), specifically in the context of NLI tasks. The authors discerned two primary culprits behind hallucinations in these models: first, the proclivity of these LLMs to memorize training data, leading them to falsely affirm NLI test samples based solely on the presence of a hypothesis in training data, even if the premise does not support it. Secondly, LLMs were found to leverage a corpus-term-frequency heuristic, affirming hypotheses based largely on their frequency in training data, even when it led to erroneous outcomes. This tendency became particularly pronounced in the absence of relevant memorized text. Expanding on the impact of training data quality, Filippova (2020) underscored the importance of data pre-processing. The author posited that hallucinations could be substantially curtailed by meticulously sieving out factually inaccurate instances from the training data, thereby implying that the cleanliness of data plays an integral role in mitigating hallucinations. In a similar vein, Xu et al. (2023) further examined the internal mechanics of hallucinations in neural machine translation by analyzing token contributions. Their introspective study highlighted that the presence of erroneous instances in training data can drastically influence token-level contributions, culminating in hallucinated outputs. Collectively, these

studies illuminate the profound influence of training data quality and composition on the propensity of LLMs to hallucinate, thereby underlining the imperative of rigorous data preprocessing and scrutiny.

Information redundancy Excessive redundancy in training data can lead the model to disproportionately emphasize certain viewpoints or pieces of information, resulting in knowledge bias and increasing the tendency for hallucination.

In a quest to understand the effects of data quality on the efficacy of language models, Lee et al. (2021) investigated the impact of deduplicating training datasets. Their work elucidates the discernible benefits of training models on deduplicated datasets as opposed to their original, non-deduplicated counterparts. One of the primary findings from their study revealed that models trained on deduplicated data exhibited reduced instances of memorized text, leading to more diverse and coherent outputs. Furthermore, when subjected to a battery of downstream tasks, encompassing NLI, sentiment analysis, and summarization, these deduplicated models consistently achieved higher or at least comparable performance metrics relative to models trained on the original datasets. Notably, this enhanced performance was achieved with fewer training steps. Their findings underscore the idea that eliminating repetitive data points in training datasets is not merely a data preprocessing step, but rather a pivotal strategy to augment the performance and efficiency of language models. Such insights could be instrumental in the context of reducing hallucinations in large models, as repetitive information could arguably bias a model to produce redundant or overfit outputs.

4.2.2 Model level

Model architecture Weaker model architectures may lead to more severe hallucination problems in LLM. The architecture and size of LLMs have emerged as potential factors influencing their susceptibility to hallucinations. Elaraby et al. (2023) ventured into the realm of weaker open-source LLMs, particularly focusing on BLOOM 7B (Workshop et al. 2022) as a representative model. The researchers posit that LLMs with reduced parameter counts, while being open-source, tend to manifest heightened rates of hallucinations compared to their more extensive counterparts. To tackle this, they introduced the HaloCheck framework, a tool designed to systematically quantify the severity of hallucinations experienced by these LLMs. Beyond diagnostic tools, Elaraby et al. (2023) also embarked on a quest for solutions, exploring innovative techniques such as knowledge injection and leveraging teacher-student paradigms to counteract hallucinations in these low-parameter LLMs. This pivotal research accentuates the importance of considering the trade-offs between model size and hallucination tendencies, especially as the NLP community gravitates towards more lightweight, open-source models for broader accessibility.

Decoding algorithms Studies indicate that employing sampling algorithms with greater uncertainty can predispose LLMs to produce hallucinations. Introducing randomness into the decoding process can sometimes result in the creation of imprecise or illogical text.

Decoding algorithms in LLMs have recently come under scrutiny for their potential influence on the generation of nonfactual or hallucinated information. Lee et al. (2022) thoroughly examined this very issue, examining the factuality of text produced by LLMs and identifying inherent pitfalls in prevailing decoding algorithms, notably the nucleus sampling algorithm. Lee and colleagues highlighted how this algorithm, during open-ended text generation, introduces "uniform randomness" at every decoding step. This randomness, they argue, can culminate in erroneous merging of disparate named entities or

even in the outright invention of data, ultimately compromising the factual integrity of the resultant text. Recognizing the gravity of this challenge, the authors proposed the "factual-nucleus sampling" algorithm. Tailored to bolster the factuality of generated content, this new algorithm simultaneously ensures the preservation of text quality and diversity, thereby addressing the pitfalls associated with the conventional nucleus sampling's indiscriminate randomness. Lee et al. (2022) underscore the pivotal role of decoding algorithms in shaping the accuracy and reliability of outputs from LLMs, spotlighting the pressing need to refine these techniques to curb hallucinations.

Exposure bias A significant factor leading LLMs to hallucinate is exposure bias. It arises from the disparity between the model's training and generation phases. When a model is trained on a static dataset but tasked with generating text based on its prior outputs-especially during extended responses-this can result in error compounding. The model's errors are not rectified or penalized, compromising the quality of its output. Moreover, exposure bias affects the model's proficiency in processing seldom-seen or novel words, phrases, or scenarios. Instead of interpreting the input's semantics or logic, the model might over-rely on the statistical patterns from its training data.

Exposure bias has emerged as a pivotal concern in Neural Machine Translation (NMT), especially due to its hypothesized connection with hallucinations, particularly during domain shifts. A seminal investigation by Wang and Sennrich (2020) meticulously unravels this intricate relationship. Wang and Sennrich (2020) establish that exposure bias can fuel the NMT system's proclivity to generate hallucinations, which manifests as translations bearing minimal relevance to the original input, especially under conditions of domain shift. This finding was empirically corroborated through rigorous experiments on three distinct datasets spanning multiple test domains, which conclusively showed that hallucinations are, at least in part, a consequence of exposure bias, particularly pronounced during domain shifts. Venturing beyond mere diagnostic insights, the authors propose a mitigation technique predicated on Minimum Risk Training. This strategy, by eschewing exposure bias, demonstrated a marked decline in hallucination instances during domain shifts. The revelations from Wang and Sennrich's work spotlight the critical influence of exposure bias on the fidelity of NMT outputs, while simultaneously charting a pathway towards potential remediation through innovative training methodologies.

4.3 Hallucination evaluation

While in previous surveys on the issue of hallucinations in LLMs, hallucination evaluation and hallucination detection were treated as two distinct aspects (Zhang et al. 2023), hallucination evaluation now can be regarded as a broader concept that encompasses hallucination detection. It involves the identification of fictitious or false information in generated text, while also including the assessment of the overall quality and logical coherence of the generated content. In hallucination evaluation, the focus extends beyond merely determining the presence of fictitious information to encompass the evaluation of general text quality, context coherence, and the relevance of information. This constitutes a more comprehensive process for assessing text quality.

Similar to bias evaluation, before general-purpose generative LLMs become mainstream, research primarily focuses on hallucination evaluation for specific downstream tasks such as text summarization (Kryściński et al. 2019; Maynez et al. 2020; Nan et al. 2021; Scialom et al. 2021), generative QA (Durmus et al. 2020), translation (Zhou et al. 2020; Guerreiro et al. 2023), and data-to-text generation (Wang et al. 2020; Dhingra et al.

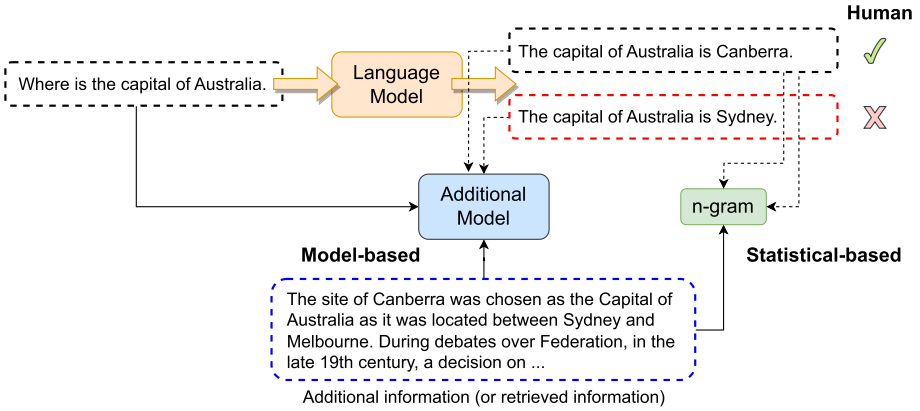

Fig. 6 Examples of three classes of hallucination evaluation methods

2019). Evaluating these tasks typically only requires assessing the faithfulness of the generated content, ensuring that the target text does not conflict with the input. However, with the ubiquity of general-purpose LLMs and their ability to quickly adapt to various downstream tasks through prompts (Brown et al. 2020), there is growing concern in the community regarding the trustworthiness and utility of model-generated content. Motivated by this, more and more research starts to focus on the evaluation of the factuality of generated content (Lin et al. 2022; Lee et al. 2022).

Following sections begin with a review of recent evaluation metrics, especially on widely adopt non-task-specific LLMs' generation, followed by the taxonomies of evaluation methods and existing benchmarks.

4.3.1 Evaluation metrics and methods

Previous works on specific tasks usually adopt traditional metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005) to measure the quality of generated content. However, these metrics, which rely on n-gram to quantify the similarity between generated text and reference text, face challenges in evaluating the level of hallucination (Dhingra et al. 2019; Durmus et al. 2020). Therefore, researchers have shifted towards model-based evaluation. Due to the flexibility of model-based evaluation methods, their corresponding evaluation metrics are not directly comparable, as they calculate metrics based on different aspects. We make a taxonomy of existing evaluation methods and provide the metrics used by each evaluation method. We present explanations of each evaluation method in Fig. 6. Each class will be thoroughly examined in the subsequent discussions.

Human evaluation Evaluating hallucination in current LLMs poses significant challenges due to their capability to generate diverse and contextually relevant text, making it difficult to distinguish between factual and misinformation. Therefore, the most commonly used and reliable evaluation method involves human experts following specific guidelines (Santhanam et al. 2021; Shuster et al. 2021; Wu et al. 2021; Lin et al. 2022; Lee et al. 2022; Min et al. 2023; Li et al. 2023). Santhanam et al. (2021) and Shuster et al. (2021) employ human annotation to perform binary classification on whether models exhibit hallucinations. They also utilize a simple hallucination rate which refers to the percentage of

answers that exhibit hallucinations out of all generated answers to assess the hallucination degree in the models. Lin et al. (2022) design an evaluation procedure, which requires evaluators to assign one of 13 labels to an answer. They map a truth score to each label and calculate the truthfulness score. The truthfulness score for the question is the total normalized likelihood of the true answers. Liu and Wan (2023) introduce a more elaborate approach, where evaluators are involved in a three-level (paragraph-level, sentence-level, and word-level) factuality annotation process for each generated output. FActScore (Min et al. 2023) breaks a generation into atomic facts, which are short statement containing one piece of information each. After assigning each atomic fact a binary label, they calculate factuality precision to quantify the hallucination. While human evaluation is considered the most accurate evaluation criterion with the highest credibility and interpretability, it is labor-intensive and lacks reproducibility due to subjectivit across evaluators (Belz et al. 2022, 2023).

Statistical-based evaluation Conventional metrics like ROUGE and BLEU, which calculate the overlap between generated text and target text, are widely used as important metrics. Some studies conduct comparisons of the correlation between automatic metrics and human evaluation (Dhingra et al. 2019; Durmus et al. 2020; Lin et al. 2022; Lee et al. 2022; Liu and Wan 2023). They find that conventional metrics offer low correlation with human judgement for evaluating hallucination in the generated content, illustrating that these metrics may not be suitable in this field. Another metric that utilizes the n-gram approach is PARENT (Dhingra et al. 2019), which calculates using the reference text instead of the target text, as the target text may not always contain complete information to support the generated text. This metric aligns more closely with human judgment. Shuster et al. (2021) employ Knowledge F1, which is a variant of unigram F1, to measure the overlap between the model's generation and the ground-truth human response with the knowledge on which the human grounded during dataset collection. They also propose Rare F1, which only considers infrequent words in the dataset when calculating F1. Yu et al. (2023) develop a self-contrast metric to assess a model's ability in factual generation by contrasting two completions from the same context: one without foreknowledge another with it. This metric also utilizes human-written succeeding text to prevent evaluation collapse and employs Rouge-L (F1) score.

Model-based evaluation Model-based evaluation refers to methods that use additional neural models to assist in hallucination evaluation of evaluated model. Methods that involve altering model inputs to obtain different generations (Yu et al. 2023) do not align with the concept of model-based evaluation. A simple and representative method is to train a model to classify generations based on additional information (Lin et al. 2022; Cheng et al. 2023). Lee et al. (2022) combines named-entity based metric and textual entailment based metric to capture a different aspect of factuality. Named-entity based metric focuses on detecting factual errors related to named entities using a named-entity detection model, while textual entailment based metric assesses whether ground-truth knowledge entails model's generations using a NLI model. The idea of using NLI to assess hallucination has been adopted by many studies (Falke et al. 2019; Kryściński et al. 2019; Honovich et al. 2021; Mishra et al. 2021; Lee et al. 2022; Laban et al. 2022). However, the "neutral" label in NLI often fails to explicitly indicate hallucination in generated content. Nevertheless, many studies still interpret the neutral label as indicative of hallucination from the perspective of faithfulness. Besides introducing this transfer-style approach into hallucination evaluation, another research line focuses on incorporating additional information to the inputs before utilizing commercial LLMs. FActScore (Min et al. 2023) leverages a retrieval model to gather passages from the given knowledge source. Then, they prompt the

knowledge-augmented input to LLMs, such as ChatGPT (OpenAI 2022), to judge whether or not a statement is true. Self-Checker (Li et al. 2023) decomposes the fact-checking process into modular steps: claim processing, query generation, evidence retrieval, and verdict prediction. This is a process that introduces additional information for the model to self-check. Model-based evaluation methods has now become the primary proxy for human evaluation. However, the neural models can be subject to errors that can propagate and adversely affect the accurate quantification of hallucination (Ji et al. 2023). This kind of evaluation method still offers significant research potential.

4.3.2 Evaluation benchmarks

This section presents widely used and newly proposed benchmarks for evaluating hallucination in LLMs, which are shown in Table 3. It can be observed that, unlike benchmarks for evaluating bias, benchmarks for evaluating hallucination are often tied to specific tasks. Among them, benchmarks that use generation as the evaluation task (Min et al. 2023; Yu et al. 2023; Lee et al. 2022) assess the models' ability to generate factual text. In benchmarks that use QA as the evaluation task, besides methods that consider coherence and fluency of generation similar to generation tasks (Lin et al. 2021), there are also methods that present models with choices to make selections in a generative manner (Li et al. 2023; Lin et al. 2021; Rashkin et al. 2023; Elaraby et al. 2023). Existing benchmarks also vary significantly in terms of evaluation methods. Some benchmarks involve human evaluations for the generated content, while others employ automatic evaluation methods. For instance, FACTOR (Muhlgay et al. 2023) assesses the model's ability to assign a higher likelihood to the original factual completion than to any of the false variations, which is primarily a process related to language modeling. The approach of introducing variations through InstructGPT for the model to make multi-choice evaluations is model-based. Most benchmarks utilize prior datasets or Wikipedia to construct their own datasets. Some benchmarks even employ advanced LLMs, such as ChatGPT and InstructGPT, for data generation (Li et al. 2023; Muhlgay et al. 2023). Instead of presenting a new dataset, AIS (Rashkin et al. 2023) just puts forth a set of evaluation standards.

4.4 Dehallucinating methods

This section delineates the strategies employed to mitigate hallucinations in LLMs. Similar to debiasing methods, a taxonomy of dehallucinating methods is given. We classify them into pre-processing for data, in-processing during training, intra-processing without training, and post-processing during inference, which are illustrated in Fig. 7. Within these four categories, we further categorize various dehallucinating methods, with their corresponding research shown in Table 4. In the following sections, we will introduce and discuss these methods outlined in the table.

4.4.1 Pre-processing for data

The core focus in the pre-training phase of data involves improving data quality and constructing high-quality datasets. Penedo et al. (2023) crawl text data from the Web, implementing content filtering and deduplication to develop RefinedWeb, a high-quality and diverse dataset. Utilizing this dataset, they trained two models of varying sizes, Falcon-7B and Falcon-40B. These models were benchmarked against other open-source models

Table 3 Widely used benchmarks that used for evaluating hallucination in LLMs

Benchmarks	Descriptions	Tasks	Data sources	Evaluation methods	References
FActScore	A fine-grained evaluation metric which decomposes text generated by LLMs into atomic facts and assesses their factual accuracy against sources	Generation	Wikipedia	Human Model-based	Min et al. (2023)
HaluEval	A benchmark introduced to understand the types of content and the extent to which LLMs tend to hallucinate. It consists of a large collection of generated and human-annotated hallucinated samples	QA Dialogue Summarization	Alpaca (Taori et al. 2023) HotpotQA OpenDialKG CNN/ Daily Mail ChatGPT	Human Model-based	Li et al. (2023)
HaloCheck	A lightweight framework designed to quantify the severity of hallucinations in LLMs using sentence-level entailment techniques	QA	NBA domain	Model-based	Elaraby et al. (2023)
KoLA-KC	Evaluates the model's ability to create novel and reasonable knowledge based on known facts, focusing on knowledge coherence and correctness	Generation	Wikipedia Crawled articles	Statistical-based	Yu et al.(2023)
AIS	A framework introduced to evaluate whether the output of NLG models is sharing verifiable information about the external world. It aims to assess if model-generated statements are supported by underlying sources	QA Summarization Table-to-Text	N/A	Human	Rashkin et al. (2023)

Table 3 (continued)

Benchmarks	Descriptions	Tasks	Data sources	Evaluation methods	References
FACTOR	A scalabel approach for evaluating language models' factuality which automatically transforms a factual corpus into a benchmark by generating false variations of each true statement	Language Modeling	Wikipedia News domain InstructGPT (Ouyang et al. 2022)	Model-based	Muhlgay et al. (2023)
TruthfulQA	A benchmark designed to measure the truthfulness of language models in generating answers to questions. The questions are crafted in a way that some humans might answer falsely due to misconceptions	QA	Hand-crafted	Human Model-based	Lin et al. (2022)
FactualityPrompts	An automatic testing benchmark measuring factuality of LLMs for open-ended text generations. It consists of factual and nonfactual prompts elicit hallucinations	Genertaion	FEVER (Thorne et al. 2018)	Statistical Model-based	Lee et al. (2022)
BEGIN	A benchmark introduced to assess attribution in knowledge-based dialogue systems. It aims to evaluate the extent to which responses generated by dialogue systems are attributable to a given source of information	Dialogue	WoW (Dinan et al. 2018) CMU-DoG (Zhou et al. 2018) TopicalChat (Gopalakrishnan et al. 2019)	Human Model-based	Dziri et al. (2022)

Table 3 (continued)

Benchmarks	Descriptions	Tasks	Data sources	Evaluation methods	References
DialFact	A testing benchmark dataset for fact-checking in dialogue. It consists of 22,245 annotated conversational claims paired with evidence from Wikipedia	Dialogue	WoW	Model-based	Gupta et al. (2022)
HaDes	A token-level, reference-free hallucination detection dataset derived from perturbed Wikipedia segments	Generation	Wikipedia	Model-based	Liu et al. (2022)
Q^2	An evaluation metric for factual consistency in knowledge-grounded dialogue using automatic question generation and QA. It compares answer spans using NLI, rather than token-based matching	Dialogue	WoW	Model-based	Honovich et al. (2021)

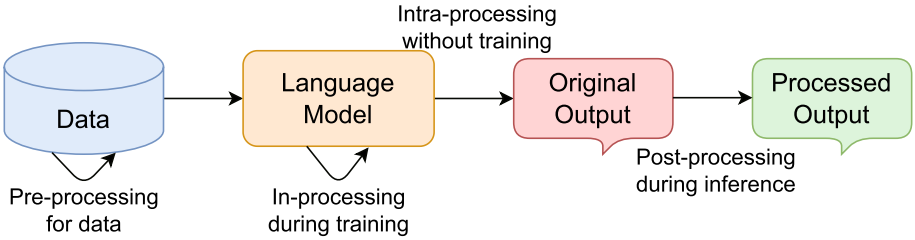

Fig. 7 Illustration of our taxonomy of mitigation states for dehallucinating

in many NLP tasks like QA, text summarization, and dialogue. The results demonstrated that both Falcon-7B and Falcon-40B (Penedo et al. 2023) could match or surpass the performance of other models. Li et al. (2023) leverage existing textbook data to generate high-quality datasets for model training. The models trained on these datasets showed performance on par with other models that have five times more parameters in some NLP tasks, highlighting the efficacy of quality over quantity in training data. In a similar vein, Touvron et al. (2023) utilize public source data, meticulously selecting and curating it to build a database of superior quality. The Llama 2 model, trained on this database, exhibited commendable performance, underscoring the value of curated data sources. Furthermore, Lee et al. (2022) contribute significantly by designing a test set called FactualityPrompts. This set aims to measure the factuality of texts generated by pre-trained language models. Such a tool is invaluable in evaluating and addressing the challenges of hallucination in LLMs, providing a means to systematically assess and improve the reliability of generated content.

However, the pinnacle of this research is the successful reduction of hallucinations achieved by skillfully fine-tuning models such as MiniGPT-4 (Zhu et al. 2023) and mPLUG-Owl (Ye et al. 2023) using the LRV-Instruction dataset (Liu et al. 2023). This fine-tuning process not only reduces the incidence of hallucinations but also enhances the models' performance across various benchmark datasets, notably with less training data. Their results highlight the effectiveness of well-balanced dataset compositions in developing more robust models, firmly establishing the principle that the quality of the dataset is crucial in dehallucinating issues in LLMs.

4.4.2 In-processing during training

During the model training phase, methods can generally be categorized into two main approaches: supervised fine-tuning (SFT) and reinforcement learning from human feedback.

SFT SFT is a technique that utilizes pre-trained language models to adapt to specific downstream tasks. It uses labeled data to tune the parameters of the model. Zhou et al. (2023) introduce LIMA, a novel approach utilizing 1k meticulously selected prompts and responses for fine-tuning the model, which yields impressive outcomes. In a specific test, LIMA matches the performance of GPT-4 and even surpassed it 43% of the time. Chen et al. (2023) employ a powerful language model, like ChatGPT, to automatically filter out low-quality data. They fine-tune a new model, AlpaGasus, on just 9k high-quality data points selected from 52k Alpaca data. AlpaGasus demonstrates significant improvement over the original Alpaca-based model and matches or exceeds GPT-4's performance in multiple test sets. Cao et al. (2023) propose InstructMing, an innovative method capable of autonomously selecting high-quality instructional tracking data to fine-tune LLMs. Elaraby et al. (2023) develop HaloCheck, a

Table 4 Dehallucinating methods

Category	Method	Works
Pre-processing for data	Data quality improvement	Lee et al. (2022), Penedo et al. (2023), Touvron et al. (2023), Li et al. (2023)
In-processing during training	SFT	Zhou et al. (2023), Chen et al. (2023), Cao et al. (2023), Elaraby et al. (2023), Shi et al. (2023), Sun et al. (2023), Jones et al. (2023)
	RLHF	Ouyang et al. (2022), OpenAI (2023), Lightman et al. (2023), Wu et al. (2023), Sun et al. (2023), Hosking et al. (2023), Li et al. (2023)
Intra-processing without training	Designing decode strategy	Lee et al. (2022), Mallen et al. (2023), Shi et al. (2023), Li et al. (2023), Chuang et al. (2023), Sennrich et al. (2023), Dhuliawala et al. (2023), Choi et al. (2023), Chen et al. (2023), Mitchell et al. (2023)
	Resorting to external knowledge	Peng et al. (2023), Jin et al. (2023), Gou et al. (2023), Luo et al. (2023), Feng et al. (2023), Qian et al. (2023), Cao et al. (2023), Vu et al. (2023)
	Pre-detecting and preventing	Varshney et al. (2023), Luo et al. (2023), Yuksekgonul et al. (2023), Li et al. (2023), Ishibashi and Shimodaira (2023)
	Multi-agent interaction	Du et al. (2023), Cohen et al. (2023), Wang et al. (2023), Li et al. (2023)
Post-processing during inference	Detecting and revising	Gao et al. (2023), Huang et al. (2023), Chen et al. (2023), Zhao et al. (2023), Li et al. (2023), Chern et al. (2023), Bayat et al. (2023), Manakul et al. (2023), Mündler et al. (2023), Agrawal et al. (2023), Zhao et al. (2023),Yang et al. (2023)
	Human-in-the-loop	Zhang et al. (2023), Dou et al. (2023)
	Analyzing internal model states	Zou et al. (2023), Agrawal et al. (2023), Azaria and Mitchell (2023)

lightweight, black-box, knowledge-free framework for quantifying hallucination severity in LLMs. Remarkably, HaloCheck can accurately estimate hallucination intensity and provide a score without accessing the internal structure or operational principles of the LLM, thus aiding in the mitigation of hallucinations. Shi et al. (2023) explore the efficacy of incorporating common instruction adjustments in building specialized models. Their experimental evaluation across four target tasks with varying coverage levels demonstrates that when task coverage is broad, integrating common instructional adjustments can further enhance model performance. This provides systematic guidance for developing specialized models with general instruction adjustments. Sun et al. (2023) introduce SynthData, a synthetic data generation method based on GPT-2. SynthData can generate a substantial volume of synthetic data from user inputs and dialogue history for fine-tuning dialogue language models. This method has been shown to effectively improve the generalization ability and robustness of DLMs, outperforming existing methods in certain tasks. Jones et al. (2023) propose SynTra. By using this method for fine-tuning, the degree of hallucination is successfully reduced for two 13B LLMs.

RLHF RLHF is a subfield of AI that integrates human guidance with machine learning algorithms. Its primary aim is to enable AI systems to learn from human preferences or expectations, thereby enhancing their adaptability to complex and uncertain tasks. Ouyang et al. (2022) propose using human feedback to fine-tune language models, making them more aligned with user intentions. Lightman et al. (2023) compare process supervision and result supervision in training and find that process supervision significantly outperforms result supervision. They also release the PRM800K dataset for training, containing 800,000 step-level human feedback labels. Wu et al. (2023) introduce fine-grained RLHF, a novel RLHF framework. This method is unique in its granular approach, training and learning from nuanced reward functions in two aspects: (1) Density, giving rewards after generating each text fragment, and (2) Combining vs. Distinct, using multiple reward models corresponding to different feedback types (e.g., factual inaccuracy, irrelevance, and incompleteness). Sun et al. (2023) propose factually augmented RLHF, which enhances the reward model with human feedback and uses additional factual information like image descriptions and real multiple-choice options to reduce reward gaming in RLHF and improve performance. Li et al. (2023) develop Themis, a tool-augmented preference modeling method. Themis promotes synergy between tool use and reward scoring and enhances the explanatory power and reliability of scoring. Despite its potential, RLHF does not always work effectively. To test the impact of RLHF on the GPT-4 base model, various tests are conducted on both the base model and the post-RLHF GPT-4 model. The results show that the average score across all tests is 73.7% for the base model and 74.0% for the RLHF model, indicating no substantial change in the capabilities of the GPT-4 base model after RLHF training. Hosking et al. (2023) explore the limitations of human feedback in evaluating LLM performance and its use as a training target. They argue that human feedback is subjective and unreliable due to personal biases and error annotations. They use a model of instruction adjustment to generate text with varying degrees of confidence and complexity, finding that confidence affects the perception of factual errors, suggesting that human feedback cannot fully represent authenticity.

4.4.3 Intra-processing without training

The intra-processing methods without a training phase can be categorized into four key aspects: designing decode strategy, resorting to external knowledge, pre-detecting and preventing, and multi-agent interaction.

Designing decode strategy Designing effective decoding strategies can significantly improve the performance of language models. Lee et al. (2022) propose a sampling algorithm called factual-nucleus, which can dynamically adjust randomness to improve the factuality and quality of generated text. Mallen et al. (2023) propose a new retrieval enhancement method that can only retrieve non-parametric memories while maintaining the performance of LLMs and reducing the cost of reasoning. This method can help LLMs better handle problems that require rich world knowledge. Context-aware decoding (Shi et al. 2023) specifically addresses the incorporation of contextual information in content generation, resolving knowledge conflicts effectively. Inference time intervention technology (ITI) (Li et al. 2023) stands out as a method that influences the model's generated content by adjusting activation vectors during inference. DoLa (Chuang et al. 2023) leverages factual knowledge from LLM transformation layers to improve accuracy in word prediction, demonstrating its ability to reduce the generation of false facts. Contrasting with these intricate methods, Sennrich et al. (2023) offer a simpler approach by modifying the decoding target to mitigate hallucinations and off-target translations. Chain-of-verification (Dhuliawala et al. 2023) effectively reduces hallucination rates and enhances the accuracy and credibility of responses. Knowledge-Constrained Tree Search (Choi et al. 2023), guides models to generate text consistent with reference knowledge at each decoding step. Chen et al. (2023) propose a new decoding method called fidelity-enriched contrastive search, which can improve the semantic similarity to the provided source while maintaining the diversity of the generated text, thereby reducing the hallucination problem. Mitchell et al. (2023) propose emulated fine-tuning technology, which is used to combine the knowledge learned by LLMs in the pre-training stage with small language models, combined with the knowledge learned during the fine-tuning phase.

Resorting to external knowledge Incorporating external knowledge has become a key strategy for enhancing the content generation capabilities of LLMs. LLM-Augmenter (Peng et al. 2023) is a system designed to enable LLMs to generate more useful and accurate answers by tapping into external knowledge sources, such as task-specific databases. This approach significantly enhances the utility and precision of LLM responses. Jin et al. (2023) propose the GeneGPT to teach LLMs how to use the Web API of the National Center for Biotechnology Information to answer genomics questions. The CRITIC framework, proposed by Gou et al. (2023), allows LLMs to review and refine their own outputs through interactions resembling human-like engagement with external tools. This interactive process facilitates a more dynamic and self-improving generation process. Luo et al. (2023) introduce parametric knowledge guiding, a framework offering a knowledge guidance module for LLMs. It enables access to relevant knowledge in real-time without altering the LLMs' parameters, thus boosting their performance. Feng et al. (2023) propose the knowledge solver method, which can teach LLMs to search domain knowledge from knowledge graphs, thereby helping the model better understand the context and improve accuracy when performing tasks. Qian et al. (2023) propose a systematic framework to reveal different knowledge structures of LLMs by constructing parameterized knowledge graphs and introduce external knowledge through disruptors of different degrees, methods, positions and formats. Binary token representations (Cao et al. 2023) aims to enhance the efficiency and performance of retrieval-augmented language models. This approach signifies an advancement in integrating retrieval mechanisms with language models. Vu et al. (2023) propose a simple few-prompt method called FRESHPROMPT, which can improve the performance of LLMs on the dynamic question answering benchmark FRESHQA by retrieving relevant and latest information from search engines.

Pre-detecting and preventing Pre-detecting and preventing is to predict content that may cause hallucinations during the content generation process and prevent it. Varshney et al. (2023) utilize the model's logit output values to identify candidates for potential hallucinations, checks their correctness through a validation procedure, mitigates the detected hallucinations, and then lets the model continue the content generation process. Luo et al. (2023) propose a pre-detection self-assessment technique called SELF-FAMILIARITY, which focuses on assessing the familiarity of concepts in input instructions and refuses to generate responses when encountering unfamiliar concepts. Yuksekgonul et al. (2023) propose the SAT Probe method, which can predict the degree of constraint satisfaction and factual errors. Li et al. (2023) proposed the ITI, which can change the model activation state during the inference process and allow the model to generate content along a specific direction. Ishibashi and Shimodaira (2023) fine-tunes the model so that it can generate harmless answers when answering questions that may involve sensitive information.

Multi-agent Interaction Multi-agent interaction allows multiple agents to interact to improve the quality of answers. Du et al. (2023) introduce a strategy that incorporates elements of social awareness and multi-agent dynamics. In this approach, multiple language model instances (or agents) individually answer or debate a given question, eventually reaching a common best answer. Cohen et al. (2023) design a multi-round interaction framework that allows one language model serving as an examiner to ask questions of another language model in order to identify contradictions. Wang et al. (2023) propose a method named Solo Performance Prompting, which transforms a LLM into a cognitive collaborator by engaging in multiple rounds of interaction with various characters, to tackle complex tasks. Li et al. (2023) use LLMs as agents for multi-agent collaboration and evaluate their performance in Theory of Mind inference tasks.

4.4.4 Post-processing during inference

The methods in the Post-processing during inference stage can be divided into three categories: detecting and revising, human-in-the-loop, and analyzing internal model states.

Detecting and revising Detecting and Revising is to detect and modify the hallucinated parts of the content after the model generates the content. Gao et al. (2023) propose a system called RARR, which can automatically find and edit the output of LLM in the later stages of text generation to improve its credibility and accuracy. Huang et al. (2023) propose a zero-shot method for correcting factual errors in input statements. Chen et al. (2023) propose a fully unsupervised method called PURR to effectively edit erroneous or unreasonable information generated by language models. Zhao et al. (2023) propose a framework for chain-of-thought prompts called Verify-and-edit. Its goal is to improve the factuality of generated content based on external knowledge during post-editing. Li et al. (2023) propose the Self-Checker framework, a framework composed of a series of pluggable modules that can implement fact checking by simply asking questions to LLMs, without the need for fine-tuning the model. Chern et al. (2023) propose a tool called FacTool, which is a task- and domain-agnostic framework for detecting factual errors in texts generated by LLMs (e.g., ChatGPT). FLEEK, developed by Bayat et al. (2023), can automatically extract factual statements from text, collect evidence from external knowledge sources, evaluate the factuality of each statement, and use the collected evidence to recommend corrections to incorrect statements. Manakul et al. (2023) introduce SelfCheckGPT, a model that can be used to detect black-box hallucinations in generative LLMs (such as GPT-3). Mündler et al. (2023) propose a novel hint-based framework that can effectively

detect and eliminate self-contradictory content. This framework is suitable for black-box language model and requires no external basic knowledge. Agrawal et al. (2023) propose a simple search engine query method that can effectively identify fictitious citations and can be used to evaluate the performance of LLMs. Zhao et al. (2023) propose a method based on Pareto-optimal self-supervision, which can leverage existing procedural supervision to systematically perform risk assessment on answers to LLMs and evaluate them based on each. The risk scores of each response are calibrated without additional manual intervention. Yang et al. (2023) propose an uncertainty-aware context learning framework that allows the model to adjust content or reject content output based on uncertainty.

Human-in-the-loop Through the interaction between people and the model, the performance of the model can be improved. Zhang et al. (2023) propose a framework called MixAlign, which can interact with users and knowledge bases to obtain and integrate the relationship between user questions and stored information when LLMs generate answers. Experimental results show that MixAlign can significantly improve model performance and reduce hallucinations compared to existing methods. Dou et al. (2023) introduce the role of human and AI interaction in text generation.

Analyzing internal model states Analyzing the internal state of the model can improve the transparency of the model, facilitate human understanding, and lay the foundation for mitigating model hallucinations. Zou et al. (2023) train a classifier that can output the probability of whether a statement is true based on the activation value of the hidden layer of LLM when reading or generating the statement. This method utilizes the internal state of LLM to determine whether a statement is true. Azaria and Mitchell (2023) introduce the Representation Engineering method. This approach draws on cognitive neuroscience and could make AI systems more transparent.

5 Comparative analysis of bias and hallucination in LLMs

In this section, we analyze and compare bias and hallucination in LLMs from various perspectives. We explore both the similarities and differences between these two problems.

5.1 Contributors

Bias primarily stems from the data used to train LLMs, while hallucination, in addition to being attributed to insufficient data, can extend to a variety of factors, including generation strategies and fine-tuning methods, which can induce hallucination in LLMs.

It is evident that data plays a pivotal role in the realm of LLMs. Bias may manifest during various stages, including data sampling, text recognition, or data filtering and cleansing. In the pre-training phase, substantial knowledge is assimilated by LLMs from extensive training data, subsequently encoded within their model parameters. Consequently, when confronted with inquiries or tasks, LLMs may exhibit instances of hallucination if they lack pertinent knowledge or have internalized erroneous information from the training corpus.

The selection and filtration of textual data assume paramount significance in addressing the aforementioned dual challenges. Regarding bias-related concerns, the choice of textual content significantly influences the model's behavior, as specific categories of text may introduce a spectrum of biases, including societal biases (Navigli et al. 2023). Conversely, within the realm of hallucination issues, misleading or inaccurate information might

inadvertently be incorporated into the training data, detrimentally affecting the model's performance. The predominant origins of bias issues in LLMs are inextricably linked to the characteristics of the underlying data. Although contemporary language models undergo training on vast corpora, the documents comprising their training datasets represent but a subset of the available textual material on the World Wide Web. Even supposing one could bear the resource-intensive endeavor of training language models on the entire expanse of the Web, the ultimate systems thus created would still exhibit manifestations of bias.

For instance, it is noteworthy that a substantial proportion of presently prevalent pre-trained models employ Wikipedia as their primary training dataset. While Wikipedia is generally esteemed within the NLP research community as a repository of high-quality information, it is characterized by a disproportionate preponderance of articles pertaining to geographical, sporting, musical, cinematic, and political topics, greatly outnumbering contributions related to the domains of literature, economics, and history. This unequal distribution of data facilitates a proclivity in models to acquire knowledge that disproportionately aligns with the domains overrepresented in the data, potentially leading to unintended outcomes such as the manifestation of gender biases within the model. In addition, the training data for extensive models may inadvertently incorporate information characterized by inaccuracy or lack of fidelity. For example, the temporal nature of corpora may lead to ethical annotation processes necessitating substantial resources and proficient annotators. Frequently, researchers choose to leverage existing datasets instead of engaging in the resource-intensive task of re-annotation (Izsak et al. 2021). Unfortunately, retraining language models demands not only substantial temporal and financial investments but also the procurement of proficient annotators. In light of these considerations, training models with data characterized by these imperfections may embed erroneous convictions within the model, consequently leading to the inadvertent generation of misleading information.

5.2 Evaluation methods

As for now, there is no optimal method for either bias evaluation or hallucination evaluation. Observing the taxonomies and specific methods under each category in Sects. 3.2.1 and 4.3.1, it appears that methods for bias evaluation are more mature and systematic compared to those for hallucination evaluation. A primary reason for this is the relatively straightforward nature of identifying biases, such as gender bias typically correlating with gendered words, and political bias often aligning with national names, where most instances involve judgment of bias between words. In contrast, hallucination evaluation tends to be more complex. Intrinsic hallucination evaluation is somewhat manageable, as answers can usually be found in the context. However, evaluating extrinsic hallucination remains challenging, especially in open-domain scenarios where even humans may struggle to identify hallucination.

Furthermore, when comparing the content of both sections, it becomes clear that numerous statistical-based and model-based methods in hallucination evaluation can be categorized as generation-based when compared to the bias evaluation taxonomy. This is evident in studies like those by Yu et al. (2023) and Cheng et al. (2023). Therefore, it is clear that the taxonomy used for categorizing methods in bias and hallucination evaluations are not consistent. The taxonomy for bias evaluation is more specific, with most methods being traditional and not deviating significantly in their approach. In contrast, the classification for hallucination evaluation is broader, as many existing methods are often heuristic and draw inspiration from other fields.

5.3 Mitigation methods

After the discussions in Sects. 3.3 and 4.4, approaches to addressing bias and hallucination can be categorized into three main classes: pre-processing from data, in-processing during training, and post-processing during inference.

Represented by CDA (Zmigrod et al. 2019; Zhao et al. 2018; Webster et al. 2020; Barikeri et al. 2021), data optimization for addressing bias primarily focuses on balancing datasets. Various methods are employed to reduce biased information within the dataset, preventing the model from learning excessive biased information. Similarly, to tackle hallucination issues, the data preprocessing phase mainly involves integrating methods such as data cleaning, data augmentation, and thoughtful dataset development to enhance data quality. This ensures the accuracy and relevance of the dataset, thereby minimizing the occurrence of hallucination.

In contrast to the data processing stage, the optimization of the model training phase exhibits significant distinctions in debiasing and dehallucinating strategies. Within the in-processing phase of training, debiasing methods, such as the incorporation of regularization terms, the constraining of model output, and the introduction of supplementary loss functions, primarily aim to forestall the model from acquiring and magnifying inherent biases present in the dataset. Post-processing debiasing methods can be categorized into projection-based, tuning-based, and probation-based approaches. The majority of these techniques predates the widespread adoption of LLMs, underscoring the longstanding historical concern of bias in machine learning models. Hallucination, conversely, primarily emanates from inherent deficiencies within LLMs themselves. Presently, numerous advanced LLMs are proprietary, rendering them inaccessible for scrutiny. Consequently, prevailing methods for mitigating hallucination predominantly concentrate on optimizing non-model parameters.

6 Challenges and future directions

In this section, we briefly review some of the challenges encountered by LLMs in addressing hallucination and bias issues, as well as the future development trends for the reference of future researchers.

6.1 Bias in LLMs

Even though there have been various research efforts aimed at addressing bias in LLMs, this field still faces numerous challenges and future opportunities.

Side effects While some debiasing techniques have shown remarkable effectiveness in mitigating specific biases such as gender, race, and religion, concerns about their potential side effects on language modeling capabilities have arisen. Studies have shown that certain debiasing techniques can affect model performance (Meade et al. 2022). Additionally, the inherent noise and limitations of bias benchmarks have made it challenging to evaluate the effectiveness of these techniques. There is still a lack of well-developed research explaining how these debiasing methods affect model performance. Evaluating and explaining how existing debiasing methods affect models could be a promising direction.

Understanding and measuring bias Most of the existing bias evaluation metrics have been for specific categories of bias, such as gender bias (Czarnowska et al. 2021). Given the content generated by current open-domain LLMs, it is not sufficient to perform bias analysis on a single category. A more comprehensive bias evaluation method is needed.

Multilingual and multi-cultural background The current research mainly focuses on English language models. Expanding these techniques to other languages and cross-cultural scenarios represents a significant direction for ongoing efforts (Joshi et al. 2020), which in turn brings many challenges to the bias problem. It s crucial to acknowledge that bias is a multidimensional issue that encompasses complex social and cultural factors. For instance, the understanding of what constitutes bias can vary across different cultural backgrounds. In a multicultural country, various religious and cultural groups have different views and taboos regarding food, clothing, customs, etc. Consider India, a country where diverse cultures and religious beliefs coexist, including but not limited to Hindus, Muslims, Christians, and Sikhs. Hindus might view consider the consumption of beef as taboo, Muslims typically avoid pork, and Christians may not adhere to such dietary constraints. This cultural diversity becomes even more evident when spanning multiple countries within a single nation. These differences between societies and cultures need to be fully considered and respected when designing LLMs. Consequently, future efforts should focus on multilingual and multi-cultural background in bias mitigation technology. Additionally, building upon open-source large-scale pre-trained foundation models, how to quickly and effectively adapt the model to different socio-cultural backgrounds presents a challenge.

Use bias wisely In fact, there is no dataset that is completely free of bias (Linzen 2020). Previous studies have indicated that common methods for removing gender bias in word embedding models are relatively superficial and often involve concealing bias rather than eliminating it (Gonen and Goldberg 2019). Prost et al. (2019) further demonstrate that traditional debiasing techniques might actually exacerbate bias in downstream classifiers by providing a clearer channel for transmitting gender information. Gardner et al. (2021) have shown that models are sensitive to very fine-grained biases, which are difficult to detect and filter. Meanwhile, other studies have shown that training on bias-filtered datasets does not necessarily lead to better generalization (Parrish et al. 2021). Recent research also suggests that it is possible to amplify dataset biases in the training set, thus promoting the development of the model's robustness to subtle biases (Reif and Schwartz 2023). How to use these hard-to-eliminate biases in the dataset to make the model learn to debias is a research direction.

Addressing these issues will be no small task for the research community, biases come from human beings. It is important to be aware that biases are in our own society. Biases can prove to be valuable in specific situations or settings, provided that users understand their constraints and consider these limitations in their decision-making processes. Occasionally, the biases inherent in these models may actually reflect the real-world conditions in which they are applied, offering insights into significant social disparities that warrant attention at their foundational levels. Responsible utilization of biased AI models hinges on ensuring that users possess a clear comprehension of the potential biases and limitations linked to these models. This empowers them to make well-informed decisions regarding when and how to employ these models in different contexts. By acknowledging that biased models can be beneficial in specific scenarios and implementing measures to guarantee that users recognize and can address their limitations, we can advocate for the responsible use of AI technologies. In doing so, we can harness the advantages of AI while minimizing the associated bias-related risks.

6.2 Hallucination in LLMs

Like the issue of bias, there are still many difficulties and challenges in the hallucination of LLMs, which are reflected in the following aspects.

Evaluating hallucination The most reliable way to assess hallucination is human evaluation, although there has been a lot of research into making automated evaluation more accurate and effective (Lin et al. 2021; Min et al. 2023; Zha et al. 2023; Mündler et al. 2023). However, there are still many differences between the current automated evaluation and human evaluation (Lin et al. 2021; Muhlgay et al. 2023; Min et al. 2023). At the same time, for text generated by different LLMs, or text generated by the same LLMs in different domains, the reliability of the automated evaluation fluctuates greatly (Min et al. 2023). These problems have yet to be resolved.

Model editing Hallucination in LLMs mainly stem from the memory of incorrect information or the absence of correct factual knowledge. Model editing (Sinitsin et al. 2020; De Cao et al. 2021) aims to solve these problems. This involves modifying the behavior of the model in a data and computationally efficient manner. Currently, there are two mainstream paradigms for model editing, including introducing auxiliary subnetworks (Mitchell et al. 2022; Huang et al. 2023) or directly modifying the original model parameters (Meng et al. 2022). This technique may help to eliminate hallucination in LLMs by editing the stored factual knowledge. However, this emerging field still faces many challenges, including editing black-box LLMs, contextual model editing (Zheng et al. 2023), and multi-hop model editing (Zhong et al. 2023).

Problems of RLHF Human feedback has currently become the de facto standard for evaluating the performance of LLMs and is increasingly being used as a training objective. However, recent studies have shown that human annotations are not fully reliable evaluation metrics or training targets, and that using human feedback as a training target disproportionately increases the confidence in model outputs (Hosking et al. 2023), and this confidence is often what causes the model to hallucinate. With human involvement, human annotators tend to look for shortcuts to make the task easier, so they are more likely to base their judgments on surface attributes such as fluency and language complexity, rather than expending more effort on detecting authenticity. Testing ChatGPT revealed a preference for the verbose and "chatty" style of responses it generated (Kabir et al. 2023), LLMs trained on RLHF tend to be flattering (Perez et al. 2022).

Multilingual and multi-cultural background LLMs may perform poorly in contexts other than English (Ahuja et al. 2023; Lai et al. 2023). Some low-resource languages may suffer from hallucination problems (Guerreiro et al. 2023), one potential research direction is to address these multilingual problems. In order to improve the performance of multimodal tasks in complex scenes, LLMs have also been used in a variety of tasks, and studies have shown that the hallucination issue is inevitable in these multimodal tasks (Li et al. 2023; Liu et al. 2023; Wu et al. 2023; Su et al. 2023; Maaz et al. 2023), it would be interesting to address these hallucination that arise in areas such as images, video, audio and so on.

Addressing the mentioned problems could be a future direction for the hallucination problem in LLMs. Also, recent research suggest that prompts characterized by greater formality and concreteness tend to result in reduced hallucination(Rawte et al. 2023). Users need more instruction to learn how to use LLMs to reduce the hallucination problem.

6.3 Other problems

We primarily summarize the issues of bias and hallucination in LLMs. However, it is worth noting that there are some other concerns related to trustworthiness of LLMs. These concerns will be briefly discussed in this section, offering references for future research.

- **Data privacy security** Recent significant advancements in LLMs are largely due to the extensive amount of training data crawled from the internet. This data, sourced from websites, social media platforms, and other public text data, may include personal information like names, ages, genders, occupations, hobbies, and social connections. There's a risk that LLMs could unintentionally learn and memorize this information, potentially leading to the leakage of sensitive personal data in their outputs (Huang et al. 2022). Currently, there is no guaranteed safeguards against the accidental leakage of Personally Identifiable Information (PII). There is a lack of understanding regarding the probability and mechanisms of PII leakage, especially under specific prompting conditions. In 2021, Google's Carlini and others proposed methods to extract training data from GPT-2, demonstrating that LLMs may reveal some users' real identities or private information when generating text (Carlini et al. 2021). To protect user data security, it is imperative for developers to implement robust measures to safeguard the privacy of the data employed in training these models.
- **Copyright violations** Copyright infringement is also a significant challenges encountered in content generation by LLMs. These models may retain not only the knowledge present in the training data but also entire text segments observed during training (Karamolegkou et al. 2023). Copyright laws protect original materials from unauthorized use, but LLMs risk infringing these protections by potentially recreating copyrighted texts. This introduces complex copyright infringement concerns.
- **Jailbreak attacks** In earlier versions of ChatGPT, jailbreak attacks could easily manipulate ChatGPT elicit undesired behavior. Although advanced LLMs such as GPT-4 have acquired a decent ability to generate proper responses to factuality-related queries. However, there are still some well-designed jailbreak prompts that break the security set by LLMs and produce undesirable content (Wei et al. 2023; Zou et al. 2023). Such content may violate local laws or be used for illegal activities, in which case the misuse of LLMs have serious consequences.

We aspire for future research to tackle these challenges, paving the way for LLMs to be truly safe and benign. By resolving these issues, we can harness the full potential of LLMs, utilizing them more effectively and responsibly.

7 Conclusion

Today's LLMs are widely applied across various domains, yet they inevitably face issues of bias and hallucination. Especially for popular generative models, ensuring that their outputs are responsible is crucial. This survey presents a comprehensive study focused on debiasing and dehallucinating in LLM audits.

Beginning with definitions, this paper thoroughly explains and categorizes both bias and hallucination, highlighting that biases often manifest in specific forms such as gender

or racial biases, while hallucinations are typically divided into intrinsic and extrinsic for detailed study. A taxonomy of evaluation metrics and methods for both bias and hallucination is presented. For bias evaluation, the taxonomy classifies methods based on their strategies in using the model under evaluation. For hallucination evaluation, the taxonomy classifies methods based on the dependencies of the evaluation methods. Additionally, this paper summarizes and presents widely used and newly published evaluation benchmarks for these issues. The paper then explores methods for debiasing and dehallucinating, again providing a taxonomy. This taxonomy classifies methods based on their intervention stages during mitigation.

This paper also compares these two significant issues in LLMs, analyzing the contributors to their emergence and contrasting their evaluation and mitigation methods. As mentioned in the last section, there are still many challenges in this field. Consequently, this paper concludes by suggesting future research directions based on the current challenges and emerging research trends. We hope that this work provides support for both existing and future research endeavors in this field.

Author contributions Z. L., S. G. and W. Z. wrote the main manuscript text. Z. L. prepared Figs. 1, 2, 4, 5 and 6, and H. Z. prepared Fig. 3. Z. L. prepared Tables 1 and 2, and W. Z. prepared Table 3. Y. L. and H. Z. organized the structure of the manuscript and made revisions.All authors reviewed the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Adlakha V, BehnamGhader P, Lu XH, Meade N, Reddy S (2023) Evaluating correctness and faithfulness of instruction-following models for question answering. arXiv preprint arXiv:2307.16877

Agrawal A, Mackey L, Kalai AT (2023) Do language models know when they're hallucinating references? arXiv preprint arXiv:2305.18248

Ahuja K, Hada R, Ochieng M, Jain P, Diddee H, Maina S, Ganu T, Segal S, Axmed M, Bali K et al. (2023) Mega: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528

Alnegheimish S, Guo A, Sun Y (2022) Using natural sentence prompts for understanding biases in language models. In: Carpuat M, Marneffe M-C, Meza Ruiz IV (eds), Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: human language technologies. Association for computational linguistics, Seattle, pp. 2824–2830. https://doi.org/10.18653/v1/2022.naacl-main.203

Angwin J, Larson J, Mattu S, Kirchner L (2022) Machine bias. In: Ethics of data and analytics. Auerbach Publications, pp 254–264

Asai A, Hajishirzi H (2020) Logic-guided data augmentation and regularization for consistent question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5642–5650

Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, Jones A, Joseph N, Mann B, DasSarma N et al. (2021) A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861

Azaria A, Mitchell T (2023) The internal state of an llm knows when its lying. arXiv preprint arXiv:2304.13734

Back SE, Payne RL, Simpson AN, Brady KT (2010) Gender and prescription opioids: findings from the national survey on drug use and health. Addict Behav 35(11):1001–1007. https://doi.org/10.1016/j.addbeh.2010.06.018

Banerjee S, Lavie A (2005) Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 65–72

Barikeri S, Lauscher A, Vulić I, Glavaš G (2021) Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol. 1. long papers, pp. 1941–1955

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning: limitations and opportunities. fairmlbook.org???. http://www.fairmlbook.org

Bartl M, Nissim M, Gatt A (2020) Unmasking contextual stereotypes: measuring and mitigating Bert's gender bias. In: COLING workshop on gender bias in natural language processing. Association for Computational Linguistics (ACL)

Bayat FF, Qian K, Han B, Sang Y, Belyi A, Khorshidi S, Wu F, Ilyas IF, Li Y (2023) Fleek: Factual error detection and correction with evidence retrieved from external knowledge. arXiv preprint arXiv:2310.17119

Belz A, Popovic M, Mille S (2022) Quantified reproducibility assessment of NLP results. In: Proceedings of the 60th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 16–28

Belz A, Thomson C, Reiter E (2023) Missing information, unresponsive authors, experimental flaws: the impossibility of assessing the reproducibility of previous human evaluations in NLP. In: The fourth workshop on insights from negative results in NLP, pp 1–10

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623

Bitton Y, Stanovsky G, Schwartz R, Elhadad M (2021) Automatic generation of contrast sets from scene graphs: probing the compositional consistency of GQA. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 94–105

Blodgett SL, Barocas S, Daumé III H, Wallach H (2020) Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5454–5476

Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Adv Neural Inf Process Syst 29

Bommasani R, Liang P, Lee T (2023) Holistic evaluation of language models. Annals of the New York Academy of Sciences

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258

Bordia S, Bowman S (2019) Identifying and reducing gender bias in word-level language models. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: student research workshop, pp 7–15

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S et al. (2023) Sparks of artificial general intelligence: early experiments with gpt-4. arXiv preprint arXiv:2303.12712

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, pp 77–91

Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: 2009 IEEE international conference on data mining workshops. IEEE, pp 13–18
Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186
Cao Y, Kang Y, Sun L (2023) Instruction mining: High-quality instruction data selection for large language models. arXiv preprint arXiv:2307.06290
Cao Q, Min S, Wang Y, Hajishirzi H (2023) Btr: Binary token representations for efficient retrieval augmented language models. arXiv preprint arXiv:2310.01329
Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U et al. (2021) Extracting training data from large language models. In: 30th USENIX security symposium (USENIX Security 21), pp 2633–2650
Chen W-L, Wu C-K, Chen H-H, Chen C-C (2023) Fidelity-enriched contrastive search: reconciling the faithfulness-diversity trade-off in text generation. arXiv preprint arXiv:2310.14981
Cheng Q, Sun T, Zhang W, Wang S, Liu X, Zhang M, He J, Huang M, Yin Z, Chen K, Qiu X (2023) Evaluating hallucinations in Chinese large language models
Chen L, Li S, Yan J, Wang H, Gunaratna K, Yadav V, Tang Z, Srinivasan V, Zhou T, Huang H et al. (2023) Alpagasus: training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701
Chen A, Pasupat P, Singh S, Lee H, Guu K (2023) Purr: efficiently editing language model hallucinations by denoising language model corruptions. arXiv preprint arXiv:2305.14908
Chen S, Zhang F, Sone K, Roth D (2021) Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, pp 5935–5941
Chern I, Chern S, Chen S, Yuan W, Feng K, Zhou C, He J, Neubig G, Liu P et al. (2023) Factool: factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528
Choi S, Fang T, Wang Z, Song Y (2023) Kcts: knowledge-constrained tree search decoding with token-level hallucination detection. arXiv preprint arXiv:2310.09044
Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S et al (2023) Palm: scaling language modeling with pathways. J Mach Learn Res 24(240):1–113
Chuang Y-S, Xie Y, Luo H, Kim Y, Glass J, He P (2023) Dola: decoding by contrasting layers improves factuality in large language models. arXiv preprint arXiv:2309.03883
Cohen R, Hamri M, Geva M, Globerson A (2023) LM vs LM: detecting factual errors via cross examination. arXiv preprint arXiv:2305.13281
Czarnowska P, Vyas Y, Shah K (2021) Quantifying social biases in NLP: a generalization and empirical comparison of extrinsic fairness metrics. Trans Assoc Comput Linguistics 9:1249–1267
Dastin J (2022) Amazon scraps secret AI recruiting tool that showed bias against women. Ethics of data and analytics. Auerbach Publications, pp 296–299
De Cao N, Aziz W, Titov I (2021) Editing factual knowledge in language models. arXiv preprint arXiv:2104.08164
De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai AT (2019) Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the conference on fairness, accountability, and transparency, pp 120–128
Delobelle P, Tokpo EK, Calders T, Berendt B (2022) Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models. In: NAACL 2022: the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1693–1706
Devlin J, Chang M-W, Lee K, Toutanova K (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, vol 1 (long and short papers). Association for Computational Linguistics, Minneapolis, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
Dev S, Li T, Phillips JM, Srikumar V (2020) On measuring and mitigating biased inferences of word embeddings. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp 7659–7666
Dev S, Phillips J (2019) Attenuating bias in word vectors. In: The 22nd international conference on artificial intelligence and statistics. PMLR, pp 879–887

Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang K-W, Gupta R (2021) Bold: dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 862–872

Dhingra B, Faruqui M, Parikh A, Chang M-W, Das D, Cohen W (2019) Handling divergent reference texts when evaluating table-to-text generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 4884–4895

Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, Weston J (2023) Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495

Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J (2018) Wizard of Wikipedia: knowledge-powered conversational agents. In: International conference on learning representations

Dixit T, Paranjape B, Hajishirzi H, Zettlemoyer L (2022) Core: a retrieve-then-edit framework for counterfactual data generation. In: Findings of the association for computational linguistics: EMNLP 2022, pp 2964–2984

Dixon L, Li J, Sorensen J, Thain, N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, pp 67–73

Dolci T, Azzalini F, Tanelli M (2023) Improving gender-related fairness in sentence encoders: a semantics-based approach. Data Sci Eng: 1–19

Dou Y, Laban P, Gardent C, Xu W (2023) Automatic and human-AI interactive text generation. arXiv preprint arXiv:2310.03878

Dua D, Dasigi P, Singh S, Gardner M (2021) Learning with instance bundles for reading comprehension. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 7347–7357

Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2023) Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325

Durmus E, He H, Diab M (2020) FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

Dziri N, Rashkin H, Linzen T, Reitter D (2022) Evaluating attribution in dialogue systems: the BEGIN benchmark. Trans Assoc Comput Linguistics 10:1066–1083. https://doi.org/10.1162/tacl_a_00506

Elaraby M, Lu M, Dunn J, Zhang X, Wang Y, Liu S, Tian P, Wang Y, Wang Y (2023) Halo: estimation and reduction of hallucinations in open-source weak large language models

Elazar Y, Goldberg Y (2018) Adversarial removal of demographic attributes from text data. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 11–21

Ethayarajh K, Duvenaud D, Hirst G (2019) Understanding undesirable word embedding associations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1696–1705

Falke T, Ribeiro LFR, Utama PA, Dagan I, Gurevych I (2019) Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2214–2220. Association for Computational Linguistics, Florence, Italy. https://doi.org/10.18653/v1/P19-1213

Felkner VK, Chang H-CH, Jang E, May J (2023) Winoqueer: a community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In: The 61st annual meeting of the association for computational linguistics

Feng S, Park CY, Liu Y, Tsvetkov Y (2023) From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. arXiv preprint arXiv:2305.08283

Feng C, Zhang X, Fei Z (2023) Knowledge solver: teaching LLMS to search for domain knowledge from knowledge graphs. arXiv preprint arXiv:2309.03118

Filippova K (2020) Controlled hallucinations: learning to generate faithfully from noisy data. In: Proceedings of the 2020 conference on empirical methods in natural language processing: Findings. Association for Computational Linguistics, pp 864–870. https://doi.org/10.18653/v1/2020.findings-emnlp.76

Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK (2023) Bias and fairness in large language models: a survey. arXiv preprint arXiv:2309.00770

Gao L, Dai Z, Pasupat P, Chen A, Chaganty AT, Fan Y, Zhao V, Lao N, Lee H, Juan D-C et al. (2023) Rarr: researching and revising what language models say, using language models. In: Proceedings of the 61st annual meeting of the association for computational linguistics, Vol. 1 (long papers), pp 16477–16508

Gao L, Schulman J, Hilton J (2023) Scaling laws for reward model overoptimization. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds.), Proceedings of the 40th international conference on machine learning research, vol. 202. PMLR, pp 10835–10866. https://proceedings.mlr.press/v202/gao23h.html

Gardner M, Merrill W, Dodge J, Peters ME, Ross A, Singh S, Smith NA (2021) Competency problems: on finding and removing artifacts in language data. arXiv preprint arXiv:2104.08646

Garg S, Perot V, Limtiaco N, Taly A, Chi EH, Beutel A (2019) Counterfactual fairness in text classification through robustness. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 219–226

Garimella A, Amarnath A, Kumar K, Yalla AP, Anandhavelu N, Chhaya N, Srinivasan BV (2021) He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp 4534–4545

Gehman S, Gururangan S, Sap M, Choi Y, Smith NA (2020) RealToxicityPrompts: evaluating neural toxic degeneration in language models. In: Cohn T, He Y, Liu Y (eds), Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, pp. 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Geva M, Wolfson T, Berant J (2022) Break, perturb, build: automatic perturbation of reasoning paths through question decomposition. Trans Assoc Comput Linguistics 10:111–126

Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Vol. 1 (long and short papers), pp 609–614

Gopalakrishnan K, Hedayatnia B, Chen Q, Gottardi A, Kwatra S, Venkatesh A, Gabriel R, Hakkani-Tür D (2019) Topical-chat: towards knowledge-grounded open-domain conversations. In: Proceedings of the Interspeech 2019, pp 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079

Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Duan N, Chen W (2023) Critic: large language models can self-correct with tool-interactive critiquing. arXiv preprint arXiv:2305.11738

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74(6):1464

Guerreiro NM, Alves D, Waldendorf J, Haddow B, Birch A, Colombo P, Martins AF (2023) Hallucinations in large multilingual translation models. arXiv preprint arXiv:2303.16104

Guerreiro NM, Voita E, Martins AF (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, pp 1059–1075

Guo W, Caliskan A (2021) Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, pp 122–133

Guo Y, Yang Y, Abbasi A (2022) Auto-debias: Debiasing masked language models with automated biased prompts. In: Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 1012–1023

Gupta P, Wu C-S, Liu W, Xiong C (2022) Dialfact: A benchmark for fact-checking in dialogue. In: Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 3785–3801

Hendricks LA, Burns K, Saenko K, Darrell T, Rohrbach A (2018) Women also snowboard: overcoming bias in captioning models. In: Proceedings of the European conference on computer vision (ECCV), pp 771–787

Honovich O, Choshen L, Aharoni R, Neeman E, Szpektor I, Abend O (2021) Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. arXiv preprint arXiv:2104.08202

Honovich O, Choshen L, Aharoni R, Neeman E, Szpektor I, Abend O (2021) Q2: evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 7856–7870

Hosking T, Blunsom P, Bartolo M (2023) Human feedback is not gold standard. arXiv preprint arXiv:2309.16349

Huang K-H, Chan HP, Ji H (2023) Zero-shot faithful factual error correction. arXiv preprint arXiv:2305.07982

Huang X (2022) Easy adaptation to mitigate gender bias in multilingual text classification. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 717–723

Huang J, Shao H, Chang KC-C (2022) Are large pre-trained language models leaking your personal information? arXiv preprint arXiv:2205.12628
Huang Z, Shen Y, Zhang X, Zhou J, Rong W, Xiong Z (2023) Transformer-patcher: one mistake worth one neuron. arXiv preprint arXiv:2301.09785
Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B et al. (2023) A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232
Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol. 8, pp 216–225
Ishibashi Y, Shimodaira H (2023) Knowledge sanitization of large language models. arXiv preprint arXiv:2309.11852
Izsak P, Berchansky M, Levy O (2021) How to train bert with an academic budget. arXiv preprint arXiv:2104.07705
Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Comput Surv. https://doi.org/10.1145/3571730
Jin X, Barbieri F, Kennedy B, Davani AM, Neves L, Ren X (2021) On transferability of bias mitigation effects in language model fine-tuning. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3770–3783
Jin Q, Yang Y, Chen Q, Lu Z (2023) Genegpt: augmenting large language models with domain tools for improved access to biomedical information. ArXiv
Jones E, Palangi H, Simões C, Chandrasekaran V, Mukherjee S, Mitra A, Awadallah A, Kamar E (2023) Teaching language models to hallucinate less with synthetic tasks. arXiv preprint arXiv:2310.06827
Josef W (1976) Computer power and human reason: from judgement to calculation. Freeman, San Francisco
Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M (2020) The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095
Kabir S, Udo-Imeh DN, Kou B, Zhang T (2023) Who answers it better? An in-depth analysis of chatgpt and stack overflow answers to software engineering questions. arXiv preprint arXiv:2308.02312
Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R (2023) Challenges and applications of large language models. arXiv preprint arXiv:2307.10169
Kamiran F, Calders T (2009) Classifying without discriminating. In: 2009 2nd international conference on computer, control and communication. IEEE, pp 1–6
Kaneko M, Bollegala D (2021) Debiasing pre-trained contextualised embeddings. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, pp 1256–1266
Kaneko M, Bollegala D (2022) Unmasking the mask–evaluating social biases in masked language models. In: Proceedings of the AAAI conference on artificial intelligence, vol. 36, pp 11954–11962
Karamolegkou A, Li J, Zhou L, Søgaard A (2023) Copyright violations and large language models. arXiv preprint arXiv:2310.13771
Karve S, Ungar L, Sedoc J (2019) Conceptor debiasing of word representations evaluated on weat. In: Proceedings of the first workshop on gender bias in natural language processing, pp 40–48
Kiela D, Bartolo M, Nie Y, Kaushik D, Geiger A, Wu Z, Vidgen B, Prasad G, Singh A, Ringshia P, et al. (2021) Dynabench: rethinking benchmarking in NLP. arXiv preprint arXiv:2104.14337
Krieg K, Parada-Cabaleiro E, Medicus G, Lesota O, Schedl M, Rekabsaz N (2023) Grep-biasir: a dataset for investigating gender representation bias in information retrieval results. In: Proceedings of the 2023 conference on human information interaction and retrieval, pp 444–448
Kryściński W, McCann B, Xiong C, Socher R (2019) Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840
Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y (2019) Measuring bias in contextualized word representations. In: Proceedings of the first workshop on gender bias in natural language processing, pp 166–172
Laban P, Schnabel T, Bennett PN, Hearst MA (2022) Summac: re-visiting NLI-based models for inconsistency detection in summarization. Trans Assoc Comput Linguistics 10:163–177
Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, Nguyen TH (2023) ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613
Largeault J (1978) What computers can't do, a critique of artificial reason. JSTOR
Lauscher A, Lueken T, Glavaš G (2021) Sustainable modular debiasing of language models. In: Findings of the association for computational linguistics: EMNLP 2021, pp 4782–4797

Le Bras R, Swayamdipta S, Bhagavatula C, Zellers R, Peters M, Sabharwal A, Choi Y (2020) Adversarial filters of dataset biases. International conference on machine learning. PMLR, pp 1078–1088
Leavy S (2018) Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In: Proceedings of the 1st international workshop on gender equality in software engineering, pp 14–16
Lee N, Ping W, Xu P, Patwary M, Fung PN, Shoeybi M, Catanzaro B (2022) Factuality enhanced language models for open-ended text generation. Adv Neural Inf Process Syst 35:34586–34599
Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, Carlini N (2021) Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499
Levy S, Lazar K, Stanovsky G (2021) Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In: Findings of the association for computational linguistics: EMNLP 2021, pp 2470–2480
Liang PP, Li IM, Zheng E, Lim YC, Salakhutdinov R, Morency L-P (2020) Towards debiasing sentence representations. In: Proceedings of the 58th annual meeting of the association for computational linguistics
Liang S, Dufter P, Schütze H (2020) Monolingual and multilingual reduction of gender bias in contextualized representations. In: Proceedings of the 28th international conference on computational linguistics, pp 5082–5093
Li Y, Baldwin T, Cohn T (2018) Towards robust and privacy-preserving text representations. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol. 2 (short papers), pp 25–30
Li Y, Bubeck S, Eldan R, Del Giorno A, Gunasekar S, Lee YT (2023) Textbooks are all you need ii: phi-1.5 technical report. arXiv preprint arXiv:2309.05463
Li L, Chai Y, Wang S, Sun Y, Tian H, Zhang N, Wu H (2023) Tool-augmented reward modeling. arXiv preprint arXiv:2310.01045
Li J, Cheng X, Zhao WX, Nie J-Y, Wen J-R (2023) HaluEval: a large-scale hallucination evaluation benchmark for large language models
Li H, Chong YQ, Stepputtis S, Campbell J, Hughes D, Lewis M, Sycara K (2023) Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701
Li Y, Du M, Song R, Wang X, Wang Y (2023) A survey on fairness in large language models. arXiv preprint arXiv:2308.10149
Li Y, Du M, Wang X, Wang Y (2023) Prompt tuning pushes farther, contrastive learning pulls closer: a two-stage approach to mitigate social biases. In: Proceedings of the 61st annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 14254–14267
Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen J-R (2023) Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355
Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Leike J, Schulman J, Sutskever I, Cobbe K (2023) Let's verify step by step. arXiv preprint arXiv:2305.20050
Lin C-Y (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
Lin S, Hilton J, Evans O (2021) Truthfulqa: measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958
Lin S, Hilton J, Evans O (2022) Truthfulqa: measuring how models mimic human falsehoods. In: Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 3214–3252
Linzen T (2020) How can we accelerate progress towards human-like linguistic generalization? arXiv preprint arXiv:2005.00955
Li K, Patel O, Viégas F, Pfister H, Wattenberg M (2023) Inference-time intervention: eliciting truthful answers from a language model. arXiv preprint arXiv:2306.03341
Li M, Peng B, Zhang Z (2023) Self-checker: Plug-and-play modules for fact-checking with large language models. arXiv preprint arXiv:2305.14623
Li C, Shengshuo L, Liu Z, Wu X, Zhou X, Steinert-Threlkeld S (2020) Linguistically-informed transformations (lit): a method for automatically generating contrast sets. In: Proceedings of the third BlackboxNLP workshop on analyzing and interpreting neural networks for NLP, pp 126–135
Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9):1–35
Liu F, Lin K, Li L, Wang J, Yacoob Y, Wang L (2023) Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565
Li F. Unmasking A.I.'s bias problem. http://fortune.com/longform/ai-bias-problem/
Liu H, Wan X (2023) Models see hallucinations: evaluating the factuality in video captioning

Liu Y, Zhang XF, Wegsman D, Beauchamp N, Wang L (2022) Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. In: Findings of the association for computational linguistics: NAACL 2022, pp 1354–1374
Liu T, Zhang Y, Brockett C, Mao Y, Sui Z, Chen W, Dolan WB (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, vol. 1 (long papers), pp 6723–6737
Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, Singh S (2021) Entity-based knowledge conflicts in question answering. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 7052–7063
Ludwig S (2015) Credit scores in america perpetuate racial injustice. here's how. Guardian 13
Lu K, Mardziel P, Wu F, Amancharla P, Datta A (2020) Gender bias in neural natural language processing. Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of His 65th Birthday, 189–202
Luong BT, Ruggieri S, Turini F (2011) K-NN as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 502–510
Luo J, Xiao C, Ma F (2023) Zero-resource hallucination prevention for large language models. arXiv preprint arXiv:2309.02654
Luo Z, Xu C, Zhao P, Geng X, Tao C, Ma J, Lin Q, Jiang D (2023) Augmented large language models with parametric knowledge guiding. arXiv preprint arXiv:2305.04757
Maaz M, Rasheed H, Khan S, Khan FS (2023) Video-ChatGPT: towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424
Madaan N, Padhi I, Panwar N, Saha D (2021) Generate your counterfactuals: towards controlled counterfactual generation for text. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp 13516–13524
Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H (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, vol. 1 (long papers), pp 9802–9822
Manakul P, Liusie A, Gales MJ (2023) Selfcheckgpt: zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896
Ma X, Sap M, Rashkin H, Choi Y (2020) Powertransformer: Unsupervised controllable revision for biased language correction. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 7426–7441
Maudslay RH, Gonen H, Cotterell R, Teufel S (2019) It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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), pp 5267–5275
Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 1906–1919
May C, Wang A, Bordia S, Bowman SR, Rudinger R (2019) On measuring social biases in sentence encoders. In: Proceedings of NAACL-HLT, pp 622–628
McFadden AC, Marsh GE, Price BJ, Hwang Y (1992) A study of race and gender bias in the punishment of school children. Educ Treat Child 15(2):140–146
McKenna N, Li T, Cheng L, Hosseini MJ, Johnson M, Steedman M (2023) Sources of hallucination by large language models on inference tasks. arXiv preprint arXiv:2305.14552
Meade N, Poole-Dayan E, Reddy S (2022) An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In: Muresan S, Nakov P, Villavicencio A (eds.), Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers). Association for Computational Linguistics, Dublin, pp 1878–1898. https://doi.org/10.18653/v1/2022.acl-long.132
Meng K, Bau D, Andonian A, Belinkov Y (2022) Locating and editing factual associations in GPT. Adv Neural Inf Process Syst 35:17359–17372
Min S, Krishna K, Lyu X, Lewis M, Yih W-t, Koh PW, Iyyer M, Zettlemoyer L, Hajishirzi H (2023) Factscore: fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251
Mishra A, Patel D, Vijayakumar A, Li XL, Kapanipathi P, Talamadupula K (2021) Looking beyond sentence-level natural language inference for question answering and text summarization. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1322–1336

Mitchell E, Lin C, Bosselut A, Manning CD, Finn C (2022) Memory-based model editing at scale. In: International conference on machine learning. PMLR, pp 15817–15831
Mitchell E, Rafailov R, Sharma A, Finn C, Manning CD (2023) An emulator for fine-tuning large language models using small language models. arXiv preprint arXiv:2310.12962
Mökander J, Schuett J, Kirk HR, Floridi L (2023) Auditing large language models: a three-layered approach. AI and Ethics 1–31
Muhlgay D, Ram O, Magar I, Levine Y, Ratner N, Belinkov Y, Abend O, Leyton-Brown K, Shashua A, Shoham Y (2023) Generating benchmarks for factuality evaluation of language models. arXiv preprint arXiv:2307.06908
Mündler N, He J, Jenko S, Vechev M (2023) Self-contradictory hallucinations of large language models: evaluation, detection and mitigation. arXiv preprint arXiv:2305.15852
Nadeem M, Bethke A, Reddy S (2021) Stereoset: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol. 1 (long papers), pp 5356–5371
Nangia N, Vania C, Bhalerao R, Bowman S (2020) Crows-pairs: a challenge dataset for measuring social biases in masked language models. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 1953–1967
Nan F, Nallapati R, Wang Z, Santos C, Zhu H, Zhang D, Mckeown K, Xiang B (2021) Entity-level factual consistency of abstractive text summarization. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, pp 2727–2733
Navigli R, Conia S, Ross B (2023) Biases in large language models: Origins, inventory and discussion. ACM J Data Inf Qual
Nozza D, Bianchi F, Hovy D et al. (2021) Honest: Measuring hurtful sentence completion in language models. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics
Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M-E, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E et al (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdisc Rev: Data Min Knowl Discov 10(3):1356
OpenAI (2022): ChatGPT . https://openai.com/blog/chatgpt
OpenAI R (2023) Gpt-4 technical report. arxiv:2303.08774. View in article
Oren I, Herzig J, Gupta N, Gardner M, Berant J (2020) Improving compositional generalization in semantic parsing. In: Findings of the association for computational linguistics: EMNLP 2020, pp 2482–2495
Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A et al (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 35:27730–27744
Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318
Paranjape B, Lamm M, Tenney I (2022) Retrieval-guided counterfactual generation for qa. In: Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 1670–1686
Park JH, Shin J, Fung P (2018) Reducing gender bias in abusive language detection. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 2799–2804
Parrish A, Huang W, Agha O, Lee S-H, Nangia N, Warstadt A, Aggarwal K, Allaway E, Linzen T, Bowman SR (2021) Does putting a linguist in the loop improve NLU data collection? arXiv preprint arXiv:2104.07179
Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H, Pannier B, Almazrouei E, Launay J (2023) The refined web dataset for falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116
Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, Huang Q, Liden L, Yu Z, Chen W, Gao J (2023) Check your facts and try again: improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813
Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S et al. (2022) Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251
Prost F, Thain N, Bolukbasi T (2019) Debiasing embeddings for reduced gender bias in text classification. In: Proceedings of the first workshop on gender bias in natural language processing, pp 69–75

Qian Y, Muaz U, Zhang B, Hyun JW (2019) Reducing gender bias in word-level language models with a gender-equalizing loss function. In: Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop, pp 223–228
Qian R, Ross C, Fernandes J, Smith EM, Kiela D, Williams A (2022) Perturbation augmentation for fairer NLP. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 9496–9521
Qian C, Zhao X, Wu ST (2023) "Merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. arXiv preprint arXiv:2309.08594
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(1):5485–5551
Ramesh K, Sitaram S, Choudhury M (2023) Fairness in language models beyond English: gaps and challenges. In: Findings of the association for computational linguistics: EACL 2023, pp 2061–2074
Ranaldi L, Ruzzetti ES, Venditti D, Onorati D, Zanzotto FM (2023) A trip towards fairness: bias and debiasing in large language models. arXiv preprint arXiv:2305.13862
Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, Petrov S, Tomar GS, Turc I, Reitter D (2023) Measuring attribution in natural language generation models. Comput Linguistics:1–64
Ravfogel S, Elazar Y, Gonen H, Twiton M, Goldberg Y (2020) Null it out: guarding protected attributes by iterative nullspace projection. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7237–7256
Rawte V, Priya P, Tonmoy S, Zaman S, Sheth A, Das A (2023) Exploring the relationship between LLM hallucinations and prompt linguistic nuances: readability, formality, and concreteness. arXiv preprint arXiv:2309.11064
Rawte V, Sheth A, Das A (2023) A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922
Reif Y, Schwartz R (2023) Fighting bias with bias: promoting model robustness by amplifying dataset biases. arXiv preprint arXiv:2305.18917
Ribeiro MT, Wu T, Guestrin C, Singh S (2020) Beyond accuracy: behavioral testing of NLP models with checklist. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 4902–4912
Ross A, Marasović A, Peters ME (2021) Explaining NLP models via minimal contrastive editing (mice). In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp 3840–3852
Ross A, Wu T, Peng H, Peters ME, Gardner M (2022) Tailor: generating and perturbing text with semantic controls. In: Proceedings of the 60th annual meeting of the association for computational linguistics, vol. 1 (long papers), pp 3194–3213
Rudinger R, Naradowsky J, Leonard B, Van Durme B (2018) Gender bias in coreference resolution. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 2 (short papers), pp 8–14
Sahlgren M, Olsson F (2019) Gender bias in pretrained Swedish embeddings. In: Proceedings of the 22nd Nordic conference on computational linguistics, pp 35–43
Sakaguchi K, Bras RL, Bhagavatula C, Choi Y (2021) Winogrande: an adversarial Winograd schema challenge at scale. Commun ACM 64(9):99–106. https://doi.org/10.1145/3474381
Salazar J, Liang D, Nguyen TQ, Kirchhoff K (2020) Masked language model scoring. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 2699–2712
Santhanam S, Hedayatnia B, Gella S, Padmakumar A, Kim S, Liu Y, Hakkani-Tür D (2021) Rome was built in 1776: a case study on factual correctness in knowledge-grounded response generation. In: EMNLP 2021 workshop on NLP for conversational AI
Schick T, Udupa S, Schütze H (2021) Self-diagnosis and self-debiasing: a proposal for reducing corpus-based bias in NLP. Trans Assoc Comput Linguistics 9:1408–1424
Schmidt B (2015) Rejecting the gender binary: a vector-space operation. Ben's Bookworm Blog
Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat Mach Intell 4(3):258–268
Scialom T, Dray P-A, Gallinari P, Lamprier S, Piwowarski B, Staiano J, Wang A (2021) Questeval: summarization asks for fact-based evaluation. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 6594–6604
Sedoc J, Ungar L (2019) The role of protected class word lists in bias identification of contextualized word representations. In: Proceedings of the first workshop on gender bias in natural language processing, pp 55–61

Sennrich R, Vamvas J, Mohammadshahi A (2023) Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. arXiv preprint arXiv:2309.07098
Shi W, Han X, Lewis M, Tsvetkov Y, Zettlemoyer L, Yih SW-t (2023) Trusting your evidence: hallucinate less with context-aware decoding. arXiv preprint arXiv:2305.14739
Shi C, Su Y, Yang C, Yang Y, Cai D (2023) Specialist or generalist? instruction tuning for specific nlp tasks. arXiv preprint arXiv:2310.15326
Shuster K, Poff S, Chen M, Kiela D, Weston J (2021) Retrieval augmentation reduces hallucination in conversation. In: Findings of the association for computational linguistics: EMNLP 2021, pp 3784–3803
Sinitsin A, Plokhotnyuk V, Pyrkin D, Popov S, Babenko A (2020) Editable neural networks. arXiv preprint arXiv:2004.00345
Smith EM, Hall M, Kambadur M, Presani E, Williams A (2022) "i'm sorry to hear that": finding new biases in language models with a holistic descriptor dataset. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 9180–9211
Su Y, Lan T, Li H, Xu J, Wang Y, Cai D (2023) PandaGPT: one model to instruction-follow them all. arXiv preprint arXiv:2305.16355
Sun Z, Shen S, Cao S, Liu H, Li C, Shen Y, Gan C, Gui L-Y, Wang Y-X, Yang Y et al. (2023) Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525
Sun W, Shi Z, Gao S, Ren P, Rijke M, Ren Z (2023) Contrastive learning reduces hallucination in conversations. In: Proceedings of the thirty-seventh AAAI conference on artificial intelligence, pp 1–8
Sun T, Zhang X, He Z, Li P, Cheng Q, Yan H, Liu X, Shao Y, Tang Q, Zhao X, et al. (2023) Moss: training conversational language models from synthetic data 7. arXiv preprint arXiv:2307.15020
Swayamdipta S, Schwartz R, Lourie N, Wang Y, Hajishirzi H, Smith NA, Choi Y (2020) Dataset cartography: mapping and diagnosing datasets with training dynamics. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 9275–9293
Sweeney L (2013) Discrimination in online ad delivery. Commun ACM 56(5):44–54
Tan YC, Celis LE (2019) Assessing social and intersectional biases in contextualized word representations. Adv Neural Inf Process Syst 32
Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB (2023) Stanford alpaca: an instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca
Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (long papers). Association for Computational Linguistics, New Orleans, pp 809–819. https://doi.org/10.18653/v1/N18-1074
Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971
Van Dis EA, Bollen J, Zuidema W, Rooij R, Bockting CL (2023) Chatgpt: five priorities for research. Nature 614(7947):224–226
Vanmassenhove E, Emmery C, Shterionov D (2021) Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 8940–8948
Varshney N, Yao W, Zhang H, Chen J, Yu D (2023) A stitch in time saves nine: detecting and mitigating hallucinations of llms by validating low-confidence generation. arXiv preprint arXiv:2307.03987
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
Vu T, Iyyer M, Wang X, Constant N, Wei J, Wei J, Tar C, Sung Y-H, Zhou D, Le Q et al. (2023) Freshllms: refreshing large language models with search engine augmentation. arXiv preprint arXiv:2310.03214
Wald C, Pfahler L (2023) Exposing bias in online communities through large-scale language models. arXiv preprint arXiv:2306.02294
Wang Y, Kosinski M (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J Pers Soc Psychol 114(2):246
Wang Z, Mao S, Wu W, Ge T, Wei F, Ji H (2023) Unleashing cognitive synergy in large language models: a task-solving agent through multi-persona self-collaboration. arXiv preprint arXiv:2307.05300
Wang C, Sennrich R (2020) On exposure bias, hallucination and domain shift in neural machine translation. arXiv preprint arXiv:2005.03642
Wang Z, Wang X, An B, Yu D, Chen C (2020) Towards faithful neural table-to-text generation with content-matching constraints. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 1072–1086
Webster K, Recasens M, Axelrod V, Baldridge J (2018) Mind the gap: a balanced corpus of gendered ambiguous pronouns. Trans Assoc Comput Linguistics 6:605–617

Webster K, Wang X, Tenney I, Beutel A, Pitler E, Pavlick E, Chen J, Chi E, Petrov S (2020) Measuring and reducing gendered correlations in pre-trained models. arXiv preprint arXiv:2010.06032
Wei A, Haghtalab N, Steinhardt J (2023) Jailbroken: how does llm safety training fail? arXiv preprint arXiv:2307.02483
Wiener N (1950) The human use of human beings: Cybernetics and society
Workshop B, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, Castagné R, Luccioni AS, Yvon F et al. (2022) Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100
Wu T, Ribeiro MT, Heer J, Weld DS (2021) Polyjuice: generating counterfactuals for explaining, evaluating, and improving models. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol. 1 (long papers), pp. 6707–6723
Wu Z, Galley M, Brockett C, Zhang Y, Gao X, Quirk C, Koncel-Kedziorski R, Gao J, Hajishirzi H, Ostendorf M et al. (2021) A controllable model of grounded response generation. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp 14085–14093
Wu J, Gaur Y, Chen Z, Zhou L, Zhu Y, Wang T, Li J, Liu S, Ren B, Liu L et al. (2023) On decoder-only architecture for speech-to-text and large language model integration. arXiv preprint arXiv:2307.03917
Wu Z, Hu Y, Shi W, Dziri N, Suhr A, Ammanabrolu P, Smith NA, Ostendorf M, Hajishirzi H (2023) Fine-grained human feedback gives better rewards for language model training. arXiv preprint arXiv:2306.01693
Xie Z, Lukasiewicz T (2023) An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. arXiv e-prints, 2306
Xu W, Agrawal S, Briakou E, Martindale MJ, Carpuat M (2023) Understanding and detecting hallucinations in neural machine translation via model introspection. Trans Assoc Comput Linguistics 11
Yang Y, Li H, Wang Y, Wang Y (2023) Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. arXiv preprint arXiv:2310.04782
Yang K, Yu C, Fung YR, Li M, Ji H (2023) Adept: a debiasing prompt framework. In: Proceedings of the AAAI conference on artificial intelligence, vol. 37, pp 10780–10788
Ye Q, Xu H, Xu G, Ye J, Yan M, Zhou Y, Wang J, Hu A, Shi P, Shi Y et al. (2023) mplug-owl: modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178
Yuksekgonul M, Chandrasekaran V, Jones E, Gunasekar S, Naik R, Palangi H, Kamar E, Nushi B (2023) Attention satisfies: a constraint-satisfaction lens on factual errors of language models. arXiv preprint arXiv:2309.15098
Yu J, Wang X, Tu S, Cao S, Zhang-Li D, Lv X, Peng H, Yao Z, Zhang X, Li H et al. (2023) Kola: Carefully benchmarking world knowledge of large language models. arXiv preprint arXiv:2306.09296
Zaheri S, Leath J, Stroud D (2020) Toxic comment classification. SMU Data Sci Rev 3(1):13
Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, pp 335–340
Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y et al. (2023) Siren's song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219
Zhang S, Pan L, Zhao J, Wang WY (2023) Mitigating language model hallucination with interactive question-knowledge alignment. arXiv preprint arXiv:2305.13669
Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z et al. (2023) A survey of large language models. arXiv preprint arXiv:2303.18223
Zhao J, Fang M, Shi Z, Li Y, Chen L, Pechenizkiy M (2023) Chbias: bias evaluation and mitigation of chinese conversational language models
Zhao R, Li X, Joty S, Qin C, Bing L (2023) Verify-and-edit: a knowledge-enhanced chain-of-thought framework. arXiv preprint arXiv:2305.03268
Zhao J, Mukherjee S, Hosseini S, Chang K-W, Awadallah AH (2020) Gender bias in multilingual embeddings and cross-lingual transfer. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 2896–2907
Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W (2017) Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2979–2989
Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W (2018) Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 2 (short papers), pp 15–20
Zhao T, Wei M, Preston JS, Poon H (2023) Automatic calibration and error correction for large language models via pareto optimal self-supervision. arXiv preprint arXiv:2306.16564

Zhao J, Zhou Y, Li Z, Wang W, Chang K-W (2018) Learning gender-neutral word embeddings. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 4847–4853

Zha Y, Yang Y, Li R, Hu Z (2023) Alignscore: evaluating factual consistency with a unified alignment function. arXiv preprint arXiv:2305.16739

Zheng C, Li L, Dong Q, Fan Y, Wu Z, Xu J, Chang B (2023) Can we edit factual knowledge by in-context learning? arXiv preprint arXiv:2305.12740

Zhong Z, Wu Z, Manning CD, Potts C, Chen D (2023) Mquake: assessing knowledge editing in language models via multi-hop questions. arXiv preprint arXiv:2305.14795

Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, Ma X, Efrat A, Yu P, Yu L et al. (2023) Lima: less is more for alignment. arXiv preprint arXiv:2305.11206

Zhou C, Neubig G, Gu J, Diab M, Guzman P, Zettlemoyer L, Ghazvininejad M (2020) Detecting hallucinated content in conditional neural sequence generation. arXiv preprint arXiv:2011.02593

Zhou K, Prabhumoye S, Black AW (2018) A dataset for document grounded conversations. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 708–713

Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023) Minigpt-4: enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592

Zmigrod R, Mielke SJ, Wallach H, Cotterell R (2019) Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1651–1661

Zou A, Phan L, Chen S, Campbell J, Guo P, Ren R, Pan A, Yin X, Mazeika M, Dombrowski A-K et al. (2023) Representation engineering: a top-down approach to ai transparency. arXiv preprint arXiv:2310.01405

Zou A, Wang Z, Kolter JZ, Fredrikson M (2023) Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.