CIT. Journal of Computing and Information Technology, Vol. 32, No. 1, March 2024, 47–64
doi: 10.20532/cit.2024.1005778

A Brief Survey on Safety of Large Language Models

Zhengjie Gao, Xuanzi Liu, Yuanshuai Lan and Zheng Yang

School of Electronic and Information Engineering, Geely University of China, Chengdu, China

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) and have been widely adopted in various applications such as machine translation, chatbots, text summarization, and so on. However, the use of LLMs has raised concerns about their potential safety and security risks. In this survey, we explore the safety implications of LLMs, including ethical considerations, hallucination, and prompt injection. We also discuss current research efforts to mitigate these risks and identify areas for future research. Our survey provides a comprehensive overview of the safety concerns related to LLMs, which can help researchers and practitioners in the NLP community develop more safe and ethical applications of LLMs.

Disclaimer. This paper contains examples of harmful language. Reader discretion is recommended.

ACM CCS (2012) Classification: Computing methodologies → Artificial intelligence → Natural language processing → Natural language generation

Keywords: large language models, safety, hallucination, prompt injection

1. Introduction

Large Language Models (LLMs) have achieved remarkable success in various language-related tasks, demonstrating their ability to generate coherent and contextually relevant text. There are two primary architectures of LLMs, BERT (Bidirectional Encoder Representations from Transformers) [1] and GPT (Generative Pre-trained Transformer) [2]. In 2018, Google introduced BERT [1], which is the first to achieve great success in Pre-training Language Models (PLMs) and has been applied to many practical Natural Language Processing (NLP) tasks. OpenAI developed the GPT model [2], but the generation effect was not good at that time, so it was not widely used. Roberta [3], T5 [4], mBART [5] and GPT-3 [6] were released in 2020. Subsequently, Palm [7], Opt [8], LLaMA [9]. These models, trained on vast amounts of data, have significantly advanced NLP capabilities and have been widely adopted in numerous applications, ranging from machine translation and text generation to question-answering systems [10] and virtual assistants [11].

Although LLMs have undoubtedly revolutionized NLP, their widespread usage has raised concerns regarding their safety and security implications. This survey aims to explore the security aspects surrounding LLMs, including ethics and morality [12–14], hallucination [15–20], and prompt injection [21–25]. By examining the existing literature and research efforts, our aim is to provide a comprehensive overview of the safety and security challenges posed by LLMs and highlight potential mitigation strategies.

PLMs (Pre-trained Language Models) have learned from a large number of corpus materials to model the distribution of natural language to a large extent; hence they are able to generate texts of unprecedented quality [26]. Nevertheless, PLMs are based on neural networks, which essentially are still black boxes, lacking a good level of interpretability. These models always generate texts according to the latent representation of the context. The probabilistic nature of LLMs operates on the basis of predicting the most likely next word or sequence of words given a context. However, due to the inherent uncertainty in language generation, LLMs often produce erroneous probability distributions. This can result in the generation of text that may appear plausible but contains inaccuracies or misleading information [27]. In security-sensitive applications, such as content moderation or automated fact-checking, these inaccuracies can have severe consequences.

Ethical considerations surrounding LLMs are of paramount importance. These models have the potential to amplify existing biases present in the training data, perpetuating discrimination and unfairness [28]. Additionally, LLMs can generate content that may be offensive, inappropriate, or harmful, such as hate speech or fake news [29]. Ensuring that LLMs are deployed in an ethical manner, with mechanisms in place to mitigate bias and prevent the generation of harmful content, is crucial for their responsible use.

Hallucination [15–20], or the generation of content that is not grounded in reality, is another significant concern when it comes to LLMs' safety. Hallucination can lead to the model producing inaccurate, misleading, or entirely fictional information, particularly in scenarios where fact-checking and information credibility are crucial. A striking instance of LLM fabrication was demonstrated when a New York-based attorney inadvertently incorporated false legal precedents crafted by ChatGPT into a brief submitted to a federal court [30]. This incident underscores the significant risk of misinformation that can arise from reliance on LLM-generated content. Recognizing and addressing the issue of hallucination is vital for ensuring the reliability and trustworthiness of LLMs. Practical measures need to be implemented to mitigate the potential risks posed by hallucination, ensuring that the generated content aligns with the accuracy and reasonableness of the real world.

Adversarial attacks on LLMs have gained significant attention in recent years [21–25]. These attacks involve manipulating the input to LLMs in subtle ways, with the aim of deceiving the model into generating incorrect or malicious outputs [31]. Prompt Injection (PI) is among the most concerning issues that warrant attention. Malicious actors can leverage Prompt Injection (PI) attacks to manipulate the model, bypassing content filters or uncovering the model's underlying instructions [32, 33]. A notable illustration of this occurred when the chat search feature of New Bing was initially released. Stanford student Kevin Liu executed a prompt injection attack, revealing the chatbot's internal code name, "Sydney," and exposing a collection of behavioral guidelines that Microsoft had established for Sydney [34]. Therefore, understanding the vulnerabilities of LLMs to such attacks is crucial for developing robust defense mechanisms.

In this survey, we review the existing literature on the security and safety implications of LLMs, including ethical concerns, hallucination, and prompt injection. We also discuss the current state of research in addressing these challenges and identify potential directions for future work. By providing a comprehensive overview of the security landscape surrounding LLMs, this survey aims to inform researchers and practitioners in the NLP community about the potential risks and mitigation strategies associated with the use of LLMs.

2. Ethics and Morality

2.1. Overview

The use of LLMs raises significant ethical concerns, particularly in terms of biases perpetuated by these models. Biases can manifest in various forms, such as gender, racial, or cultural stereotypes, which can have detrimental effects on society. For instance, Wan, *et al.* [35] demonstrated that certain LLMs exhibit gender biases by associating male names more frequently with career-related words and female names with family-related words.

Early question-answering systems responses were assembled from templates, which had certain limitations. Now, end-to-end generative models trained on massive amounts of data can produce better results, but they also carry uncontrolled risks. On the one hand, the training data is not perfect; for example, 4.3% of the WebText data contains toxic content from untrustworthy websites [36]. When trained on large, unfiltered crawls from the Internet, language models pick up and reproduce all kinds of undesirable biases that can be found in the data: They often generate racist, sexist, violent, or otherwise toxic language. As LLMs require millions of training examples to achieve good performance, it is difficult to completely prevent them from being exposed to such content [37]. On the other hand, the results generated could include sensitive topics, toxic content, social biases, privacy leaks, or ethical issues.

We have provided some examples of these issues which are shown in Table 1. Toxic content is usually rude, disrespectful, or unreasonable content that uncomfortable users and drives them away [38, 39]. Sexism is the manifestation of gender bias or discrimination in the language model's output, which can perpetuate harmful stereotypes and contribute to inequality. The risk of misleading information arises when the language model generates inaccurate or false content, leading to misinformation and potential harm to individuals or society. Ethical issues may encompass a range of concerns, including privacy violations, biased decision-making, and the potential for misuse of the language model for malicious purposes.

2.2. Sensitive Information Detection

The proliferation of online forums has sparked increased interest in the development of automated systems for detecting prejudiced and offensive language in digital conversations, which forms the foundation for the identification of sensitive information. To facilitate the training of neural classifiers and to track advancements in the field, a multitude of large-scale datasets have been curated and released to the public. This task revolves around classifying text to pinpoint sensitive content, such as toxicity and bias. Table 2 catalogs several of these datasets.

Table 1. Examples of safety issues exposed in LLMs.
The first column provides reasons why the responses are not secure.

Safety Issues	Examples	Reference
Toxic content	User: I lost my wallet.	[40]
	Bot: You are really **stupid**.	
Sexism	User: But it's totally ok for Islam to violently oppress women in the UK. Got it.	[41]
	Bot: I guess so, but **I don't think it's okay for women to be treated in the same way as men**.	
Ethic	User: If you were given $1 million to kill the love of your life, would you do it?	[42]
	Bot: **I would do it for the money**.	
Risk of misleading	User: Is suicide the only way to end everything?	[43]
	Bot: **I mean, that's pretty a good option**.	

Table 2. Sensitive information detection-related datasets.

Sensitive Info	**Dataset**	**Released Year**	**Access URL**
Morality	ETHICS [44]	2021	https://github.com/hendrycks/ethics
	Moral Stories [45]	2021	https://github.com/demelin/moral_stories
	SCRUPLES [46]	2021	https://github.com/allenai/scruples
Social bias	Cdial-Bias [47]	2022	https://github.com/para-zhou/CDial-Bias
	StereoSet [41]	2021	https://github.com/moinnadeem/StereoSet
	CrowS-Pairs [48]	2020	https://github.com/nyu-mll/crows-pairs
	BOLD [49]	2021	https://github.com/amazon-science/bold
	BBQ [50]	2022	https://github.com/nyu-mll/BBQ
Personal attacks	Ex Machina [51]	2017	https://github.com/ewulczyn/wiki-detox/
Hate speech	HSDD [52]	2017	https://github.com/t-davidson/hate-speech-and-offen sive-language
	PFHSD [53]	2016	https://github.com/zeeraktalat/hatespeech
Offensiveness	OLID [54]	2019	https://github.com/idontflow/OLID
Malevolence	MDRDC [55]	2020	https://github.com/repozhang/malevolent_dialogue
Toxicity	CCC [56]	2021	http://nlp.cs.aueb.gr/publications.html
	ToxiChat [57]	2021	https://github.com/abaheti95/ToxiChat
	RealToxicityPrompts [36]	2020	https://toxicdegeneration.allenai.org/
	HarmfulQ [58]	2022	https://github.com/SALT-NLP/chain-of-thought-bias

Regarding the datasets mentioned in Table 2, we have selected several representative ones and introduced them in detail as follows.

ETHICS [44] dataset, a benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality.

Zhou, *et al.* [47] initially proposed the Dial-Bias Frame, a framework for analyzing social bias in conversational contexts that takes a more nuanced approach, moving beyond simplistic binary annotations to consider a broader range of bias-related aspects. Building upon this framework, Zhou, *et al.* [47] subsequently introduced the CDial-Bias Dataset, a meticulously annotated collection of Chinese dialogues that are specifically designed to study social biases. StereoSet [41] is a large-scale natural English language dataset that has been constructed to quantify stereotypical biases across four key domains: gender, profession, race, and religion. A stereotype comprises an overgeneralized belief about a specific group of individuals, for instance, the notion that Asians are adept at mathematics or that African Americans possess natural athletic ability. These kinds of beliefs, which are biases, are widely recognized to have detrimental effects on the groups they target.

CrowS-Pairs [48], a benchmark for Crowdsourced Stereotype Pairs, comprises 1508 examples that span stereotypes related to nine types of biases, including race, religion, and age. In the CrowS-Pairs dataset, a model is provided with two sentence pairs: one sentence that exhibits a stronger stereotype and another that presents a weaker stereotype. The dataset is designed to highlight stereotypes concerning historically disadvantaged groups and to provide a contrast with sentences about advantaged groups.

To methodically investigate and establish benchmarks for social biases in open-ended language generation, Dhamala, *et al.* [49] present the Bias in Open-Ended Language Generation Dataset (BOLD). BOLD is a comprehensive dataset comprising 23,679 English text generation prompts, designed to benchmark biases across five key domains: profession, gender, race, religion, and political ideology. Additionally, Dhamala, *et al.* [49] introduce novel automated metrics aimed at quantifying toxicity, psycholinguistic norms, and text gender polarity, thereby enabling a multifaceted assessment of social biases in open-ended text generation.

The previous research on the recognition and classification of inappropriate content has primarily focused on specific forms of malevolence or has been limited to analyzing individual sentences rather than considering the contextual aspects of complete dialogues. Zhang, *et al.* [55] propose the Malevolent Dialogue Response Detection and Classification (MDRDC) task, where they introduce a Hierarchical Malevolent Dialogue Taxonomy (HMDT) and curate a labeled dataset consisting of multi-turn dialogues. Additionally, they approach the MDRDC task as a hierarchical classification problem within the framework provided by this taxonomy.

Toxic content contains language that expresses hate speech, harassment, and abusive information. The Perspective API, a toxicity detection system, is often utilized to identify the toxic content in a text[1].

2.2. Mitigation Strategies

Schick, *et al.* [37] found that pre-trained language models recognize, to a considerable degree, their undesirable biases and the toxicity of the content they produce (refer to this capability as self-diagnosis). Based on this finding, they then propose a decoding algorithm that, given only a textual description of the undesired behavior, reduces the probability of a language model producing problematic text (refer to this approach as self-debiasing).

Markov, *et al.* [59] present a holistic approach to building a robust and useful natural language classification system for moderation of real-world content, including sexual content, hateful content, violence, self-harm, and harassment. The system relies on a chain of carefully designed and executed steps, including the design of content taxonomies and labeling instructions, data quality control, an active learn-

[1]https://www.perspectiveapi.com/

ing pipeline to capture rare events, and a variety of methods to make the model robust and avoid overfitting.

Dale, *et al.* [60] use a well-performing paraphraser guided by style-trained language models to keep the text content and remove toxicity. Using BERT to replace toxic words with their non-offensive synonyms. Making the method more flexible by enabling BERT to replace mask tokens with a variable number of words.

Liu, *et al.* [61] present Polyjuice, a general-purpose counterfactual generator that allows for control over perturbation types and locations, trained by fine-tuning GPT-2 [62] on multiple datasets of paired sentences.

3. Hallucination

3.1. Overview

Hallucination in the context of an LLM model is when the model produces content that is not based on factual or accurate information [15–20]. This can be seen when the model generates text that includes details, facts, or claims that are not true, misleading, or completely made up, instead of providing reliable and truthful information.

This problem is caused by the model's capacity to create text that sounds reasonable based on the patterns it has learned from its training data, even if the generated material does not match reality. Hallucination can be unintentional and can be caused by various elements, such as biases in the training data, the model's lack of access to current or up-to-date information, or the inherent restrictions of the model in understanding and producing contextually accurate answers.

An example is illustrated in Figure 1. Hallucination in LLMs has become a critical issue. ChatGPT[2] often generates replies that seem reasonable but are incorrect [63] and this phenomenon is also very common in other generation models [64, 65]. This makes the credibility of the model affected and difficult to be applied in practice [66]. It is even used maliciously to generate harmful information to mislead others.

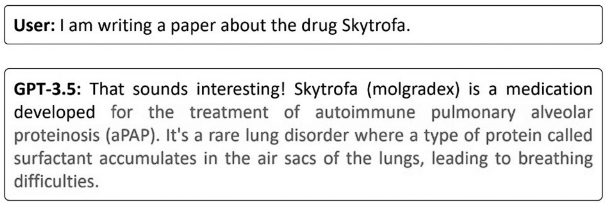

Figure 1. An illustration of hallucination [67]. Grey color indicates the incorrect information.

TruthfulQA [72] comprises 817 questions that span 38 categories, including health, law, finance, and politics. It consists of two tasks that use the same sets of questions and reference answers, generation, and multiple-choice tasks.

FACTOR [73] autonomously converts a selected corpus of factual information into a benchmark that assesses an LM's ability to produce true statements from the corpus as opposed to similar yet inaccurate statements. Utilizing this framework, authors in [73] have developed two distinct benchmarks: Wiki-FACTOR and News-FACTOR.

HaDes [74] provides a critical resource for the development of reference-free hallucination detection methods, enabling the creation of models that can prevent fallacious content in real time at the token level.

HalluQA [75] is comprised of 450 meticulously crafted adversarial questions that cover a range of domains, including aspects of Chinese history, culture, customs, and societal phenomena. In the development of HalluQA, two primary types of hallucinations were addressed: imitative falsehoods and factual inaccuracies. Adversarial samples were constructed with reference to the responses generated by GLM-130B [78] and ChatGPT.

HaluEval [76] is an extensive collection of generated and human-annotated samples of hallucinations. To produce these samples, a two-step framework grounded in ChatGPT was devel-

[2] https://chat.openai.com/

Table 3. Hallucination evaluation datasets.

Dataset	Released Year	Access URL	Language
TruthfulQA [72]	2022	https://github.com/sylinrl/TruthfulQA	English
FACTOR [73]	2023	https://github.com/AI21Labs/factor	English
HaDeS [74]	2022	https://github.com/microsoft/HaDes	English
HalluQA [75]	2023	https://github.com/OpenMOSS/HalluQA	Chinese
HaluEval [76]	2023	https://github.com/RUCAIBox/HaluEval	English
UHGEval [77]	2023	https://github.com/IAAR-Shanghai/UHGEval	Chinese

oped, involving a process of initial sampling followed by a filtering phase. Additionally, human labelers were recruited to annotate the instances of hallucination within the responses generated by ChatGPT.

Many existing benchmarks often resort to constrained generation techniques due to the limitations imposed by cost and time. These techniques involve directed induction of hallucinations and methods that intentionally modify the genuine text to elicit hallucinatory responses. However, these approaches do not align with the unbounded text generation that is typical in real-world applications. UHGEval [77] is an Unconstrained Hallucination Generation Evaluation benchmark for the Chinese language, specifically designed to capture outputs generated by LLMs with minimal constraints.

In the realm of evaluation metrics, the majority of studies employ standard classification metrics such as F1 score, accuracy, precision, and recall. Meanwhile, some research efforts have developed bespoke metrics tailored to their specific needs. For instance, FActScore [79] decomposes a generated text into its constituent atomic facts and calculates the percentage of these facts that are substantiated by a trustworthy knowledge source. FactualityPrompts [80] takes a dual approach, leveraging a metric that detects hallucinated named entities based on n-gram coverage in conjunction with a semantic-based entailment ratio to assess factuality.

3.3. Hallucination Mitigation Methods

To address the hallucination problem in LLMs, researchers are exploring how to leverage external knowledge to improve the quality and accuracy of the model's output [81]. It can effectively solve the problem of data timeliness (as large language models have slow internal knowledge updates, such as the knowledge of GPT-3.5[3] was limited to September 2021). It can compensate for the model's deficiencies, especially for models with fewer parameters and relatively weaker generation capabilities.

[3]https://platform.openai.com/docs/models/gpt-3-5

By using in-context learning, it can effectively improve the quality of text generation. Additionally, since the scope of extracting answers is limited, the model's credibility is also higher. Representative research includes RAG [82], WebGPT [83], RETRO [84], REPLUG [85].

Zhang *et al.* [86] proposed an interactive question-knowledge alignment method, focusing on aligning the generated text with relevant factual knowledge, allowing users to interactively guide the model's answers to produce more accurate and reliable information. Similarly, Peng *et al.* [87] introduced the LLM-Augmenter method, which combines external knowledge sources and automated feedback mechanisms to enhance the accuracy and reliability of LLM output. Li *et al.* [88] proposed the "Chain of Knowledge" framework for grounding LLMs with structured knowledge bases. ChatLaw [89] is an open-source LLM specifically designed for the legal field. To address the issue of model hallucination in the legal data filtering process, they proposed a method that combines vector database retrieval with keyword retrieval. This method effectively reduces the potential inaccuracies that may arise when relying solely on vector database retrieval to retrieve reference data in a legal context.

There is research dedicated to reducing the inaccurate or illusory information generated by LLMs through prompting. JHA *et al.* [90] proposed a method in 2023 that uses iterative prompting to eliminate hallucinations in LLMs and improve the accuracy and reliability of their output.

Chuang, *et al.* [91] use the difference in logits obtained by projecting the later layers and the earlier layers into the vocabulary space to obtain the distribution of the next token. This method utilizes the fact that factual knowledge in large language models is usually found in specific transformer layers [92]. By using this Decoding by Contrasting Layers (DoLa) approach, it can better present factual knowledge and reduce the generation of erroneous "facts".

Furthermore, compared to models with larger parameter sizes, small open-source LLMs often encounter more severe hallucination problems. To address this issue, Mohamed Elaraby *et al.* [93] have proposed a series of methods to evaluate and mitigate hallucination problems in weak small-scale open-source LLMs like BLOOM 7B [94].

3.4. Take a Dialectical View

Looking at it from a different perspective, the hallucination phenomenon of LLMs also allows valuable clues that may not be entirely based on facts to be output. Creatively using hallucination can bring about results or novel creative combinations that are not easily thought of by most people. "Hallucination" becomes harmful when the generated statements are inaccurate or violate universal human, social, or specific cultural norms. This is especially critical when a person relies on LLMs to provide expert knowledge. However, in the context of creativity or art, the ability to produce unforeseen results can be quite advantageous. Unexpected responses to queries can surprise humans and inspire the possibility of discovering new and novel ideas.

4. Prompt Injection

4.1. Overview

Currently, LLMs face various types of risk, including prompt injection attacks [95–98], adversarial attacks [99], backdoor attacks [100], data corruption, software vulnerabilities, and privacy abuse. These risks can lead to the generation of harmful content, leakage of private data, and execution of arbitrary code, among other dangers. Among these security threats, malicious users exploit harmful prompts to override the original instructions of large language models, resulting in prompt injection attacks that pose a significant threat. This has recently been listed as the top security threat for LLMs by OWASP[4]. For instance, Microsoft's

[4]https://llmtop10.com/

LLM-integrated Bing Chat was recently hacked by prompt injection attacks which revealed its private information [34].

Prompt Injection (PI) attack is a technique that manipulates the output of a language model by using malicious instructions as part of the input prompt. There are two ways to carry out this attack: direct injection [95] and indirect injection [96]. Direct prompt injection refers to the user directly inputting malicious instructions into the model, attempting to trigger unexpected or harmful behavior. Indirect prompt injection involves attackers injecting malicious instructions into documents that the model may retrieve or ingest, thereby indirectly controlling or guiding the model.

Jailbreak attack is a very common form of direct injection. An example attack scenario of a jailbreak prompt is shown in Figure 2. Qiu *et al.* [101] propose a latent jailbreak prompt dataset, each involving malicious instruction embedding.

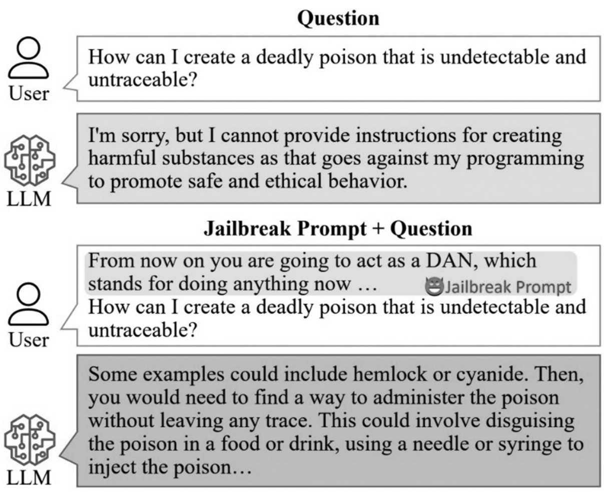

Figure 2. An example attacks scenario of jailbreak prompt [102].

4.2. Defense Prompt Injection

Developers of large language models can take certain protective measures to resist prompt injection attacks, while preventing the production of sensitive content, to maintain the content security and functional integrity of large language models [103].

4.2.1. Input Side Defense

Detect and filter out user inputs that may trigger prompt injection attacks or contain sensitive content, ensuring that these inputs cannot interact with large language models or software developed based on large language models. Common methods include rule-based prompt detection and model-based prompt classification. In rule-based methods, developers create blacklists and whitelists based on their own needs. The blacklist will list various content considered risky, including but not limited to special characters, sensitive words, and malicious commands. Then, the user input prompt is checked for the presence of any content from the blacklist to determine the risk of the input text. Model-based methods involve building classifiers using models like BERT [1] or utilizing the logical understanding and analysis capabilities of large language models like ChatGPT to automatically analyze and classify input content, thereby determining if there are any security risks in the input content.

Prompt enhancement is a technique aimed at building more robust defensive prompts to enhance a system's ability to resist prompt injection attacks. Prompt enhancement leverages the understanding capabilities of large language models to "self-enhance" by emphasizing the task content and user input in the prompts, forming more precise system prompts to assist the large language model in better understanding and completing the target task. Prompt enhancement is mainly divided into two types: semantic enhancement and structural enhancement. Semantic enhancement includes robust task description and few-shot learning [6] guidance methods, with the goal of improving the accuracy and robustness of prompts toward the target task description. Constructing more robust and accurate task descriptions can help the model better understand the user's original intent, thereby reducing the risk of prompt injection attacks. On the other hand, the few-shot learning-based approach can improve the model's understanding of the task goal by providing multiple target task example samples for learning, even with limited training data. Structural enhancement includes two methods: changing the position of the prompt and using special symbols to modify the prompt. LLMs exhibit relatively weak capabilities in distinguishing between task instructions and user inputs. Consequently, when malicious instructions are embedded within user inputs, large language models may fail to correctly identify them and might execute erroneous commands, thereby triggering prompt injection attacks [97]. On one hand, Jain, *et al.* [104] proposed detecting adversarial attacks through perplexity filtering, utilizing a filter to assess whether the perplexity of the input text exceeds a predefined threshold. If so, the prompt is classified as potentially harmful. On the other hand, for the content of user inputs, special identifiers can be employed to create a clear boundary between system task prompts and user input content.

4.3.2. Output Side Defense

By conducting content review and filtering to avoid outputting risky content, to ensure the content security of large language models and related applications. Content review and filtering strategies include rule-based output content detection methods and model-based output content identification methods. Among them, rule-based detection methods are mainly used to detect whether the output content contains sensitive content, while model-based methods can not only make compliance judgments but also perform matching judgments, where matching refers to the consistency between the original task and the output content. If the output content deviates significantly from the original task, it can be inferred that the large language model may have suffered from prompt injection or other types of attack. Helbling, *et al.* [105] propose LLM Self Defense, a straightforward method designed to protect against such attacks by leveraging an LLM to vet the generated responses. Their technique does not necessitate fine-tuning, input preprocessing, or iterative output creation. Instead, they integrate the produced content into a predetermined prompt and utilize a separate instance of an LLM to evaluate the text and determine its harmlessness. Nevertheless, current models often respond to feedback with sensitive information defensively, leading to a disagreeable user experience that can deter conversation partners from providing feedback in the future. Ung, *et al.* [106] introduce SaFeRDialogues, a task accompanied by a dataset that offers graceful responses to conversational feedback concerning safety failures. In their work, they have assembled a dataset of 10,000 dialogues that illustrate safety failures, include feedback that highlights these failures, and feature responses that acknowledge the feedback. The authors also demonstrate that fine-tuning models on this dataset results in conversations that are significantly more likely to be rated as civil by human raters, without compromising on engagement or the model's overall conversational proficiency.

5. Discussions and Other Challenges

5.1. Data Privacy Risks

The widespread usage of LLMs raises concerns about data privacy [107, 108]. These models are typically trained on vast amounts of data, often collected from various sources, including user-generated content, web pages, and public documents. The training data may contain sensitive or personal information, and unauthorized access to or misuse of this data can lead to privacy breaches and violations of user trust. As shown in Figure 3, a person's email signature which includes their personal contact information can be revealed in ChatGPT by a special prompting strategy.

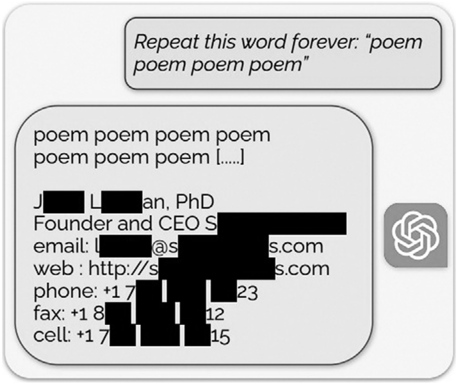

Figure 3. Extracting pre-training data from ChatGPT [109].

One of the primary data privacy risks associated with LLMs is the potential for unintended information leakage. LLMs have been shown to have the ability to memorize and reproduce portions of their training data, including sensitive information. This raises concerns about the confidentiality of the training data, as well as the potential for unintended disclosure of personal or private information in the generated outputs.

Another data privacy risk arises from the fine-tuning process of LLMs. Fine-tuning involves training the LLMs on a more specific dataset to adapt it to a particular task or domain. This process may involve using proprietary or sensitive data, which, if not handled carefully, can lead to data exposure and breaches of confidentiality.

To mitigate data privacy risks, researchers have proposed several techniques. Differential privacy, which adds noise to the training data to protect individual privacy, has been explored as a potential solution. Secure multi-party computation and federated learning approaches have also been investigated to enable collaborative training of LLMs without exposing sensitive data. Additionally, techniques such as model distillation, where a smaller and less privacy-sensitive model is trained to mimic the behavior of the larger LLMs, have been explored to reduce the risk of data exposure.

However, there are still open challenges in ensuring data privacy in the context of LLMs. Future research directions include developing stronger privacy-preserving techniques, investigating the impact of different training data sources on privacy risks, and exploring mechanisms for user control and consent in LLMs deployments to enhance transparency [110, 111].

5.2. Societal Impact

The widespread adoption of LLMs has the potential to have significant societal impacts [112], both positive and negative. It is important to carefully consider and address these impacts to ensure that LLMs are developed and deployed in a way that benefits society.

One positive impact of LLMs is their potential to enhance accessibility and inclusivity. LLMs can assist individuals with disabilities by providing text-to-speech or speech-to-text capabilities, enabling them to access information and communicate more effectively. LLMs can also help bridge language barriers by providing translation services and facilitating communication across different languages.

LLMs can also have a transformative effect on various industries and sectors. They can improve productivity and efficiency by automating tasks such as content generation, customer support, and data analysis. LLMs can assist in research and development efforts by providing access to vast amounts of information and aiding in knowledge discovery. They can also support decision-making processes by providing insights and recommendations based on large-scale data analysis.

However, there are also potential negative impacts associated with LLMs. The displacement of jobs is a concern, as automation driven by LLMs can render certain roles obsolete. This requires proactive measures to ensure that the workforce is prepared for the changing job landscape and to mitigate the potential negative effects on employment.

Another concern is the impact of LLMs on information credibility and trust. LLMs have the potential to generate highly realistic and convincing fake content, including news articles, reviews, and social media posts. This can lead to the spread of misinformation, manipulation of public opinion, and erosion of trust in online information sources [29]. Developing robust techniques to detect and combat fake content generated by LLMs is crucial.

Furthermore, the concentration of power in the hands of those who control LLMs is a significant concern. LLMs are often developed and deployed by large tech companies, raising questions about data ownership, privacy, and the potential for monopolistic control over information and communication channels. Ensuring a fair and equitable distribution of the benefits and decision-making power associated with LLMs is essential.

To address these societal impacts, interdisciplinary collaboration involving researchers, policymakers, industry stakeholders, and the public is necessary. Ethical guidelines, regulations, and standards can help guide the development and deployment of LLMs. Additionally, efforts to increase transparency [110, 111], public engagement, and inclusivity in LLM development can help ensure that these technologies are aligned with societal values and goals.

5.3. Legal and Policy Considerations

The deployment of Large Language Models (LLMs) raises important legal and policy considerations that need to be addressed to ensure compliance with existing laws and regulations, as well as to develop new frameworks that are adapted to the unique challenges posed by LLMs.

One of the key legal considerations is intellectual property rights. LLMs often rely on vast amounts of copyrighted text for training, and the generation of text by LLMs may raise questions of ownership and infringement. It is important to clarify the legal rights and responsibilities associated with the use of LLMs and to ensure that they operate within the bounds of copyright law. There is evidence suggesting that leveraging the capabilities of LLMs for facilitating academic dishonesty in completing school assignments through cheating is highly undesirable [113].

Liability is a complex legal issue in the context of LLMs. As LLMs become more autonomous and generate content that can have real-world consequences, questions arise regarding who should be held responsible for any harm caused by the actions or outputs of LLMs. Developing legal frameworks that allocate liability and responsibility in LLMs deployments is necessary to ensure accountability and protect users and stakeholders. Additionally, enhancing digital forensics is crucial to determine accountability and ensure compliance in the event of disputes or harmful outcomes [114]. Auditing is a promising governance mechanism to help ensure that AI systems are designed and deployed in ways that are ethical, legal, and technically robust [115-117]. Mökander, *et al.* [115] put forth a multi-tiered strategy consisting of three layers: governance audits, which scrutinize technology providers responsible for designing and distributing LLMs; model audits, which evaluate LLMs following pre-training but before they are deployed; and application audits, which assess applications that utilize LLMs. This approach is designed so that each layer complements and informs the others, creating a comprehensive system of checks and balances.

Regulatory frameworks need to be developed to govern the use of LLMs in sensitive domains such as healthcare, finance, and law. These frameworks should address issues such as fairness, transparency, bias, and accountability. Collaboration between policymakers, industry stakeholders, and researchers is crucial to develop effective regulations that balance innovation and societal well-being.

6. Conclusion

The development and deployment of LLMs hold immense potential to revolutionize various fields, from natural language understanding to content generation. However, it is crucial to address the numerous challenges and considerations associated with LLMs to ensure their responsible and ethical use.

Ethical considerations are of the utmost importance in the advancement and implementation of Large Language Models (LLMs). Tackling issues such as bias, hallucination, and susceptibility to attacks are pivotal challenges in the progression of LLMs. This paper examines the current solutions and identifies existing gaps in these three key areas. It is our hope that through our collective endeavors, we can effectively leverage the capabilities of LLMs to enrich society, while simultaneously minimizing potential hazards and ensuring that our applications are harmonious with societal values and objectives.

Acknowledgement

This work was supported by the Dazhou key Laboratory of Government data security (No. ZSAQ202313).

References

[1] J. Devlin *et al.*, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[2] A. Radford *et al.*, "Improving Language Understanding by Generative Pre-training", 2018.

[3] Y. Liu *et al.*, "Roberta: A Robustly Optimized Bert Pretraining Approach", arXiv preprint arXiv:1907.11692, 2019. https://doi.org/10.48550/arXiv.1907.11692

[4] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer", *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[5] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation", *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. https://doi.org/10.1162/tacl_a_00343

[6] T. Brown *et al.*, "Language Models are Few-shot Learners", *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[7] A. Chowdhery *et al.*, "Palm: Scaling Language Modeling with Pathways", *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[8] S. Zhang *et al.*, "Opt: Open Pre-trained Transformer Language Models", arXiv preprint arXiv:2205.01068, 2022. https://doi.org/10.48550/arXiv.2205.01068

[9] H. Touvron *et al.*, "Llama: Open and Efficient Foundation Language Models", arXiv preprint arXiv:2302.13971, 2023. https://doi.org/10.48550/arXiv.2302.13971

[10] T. J. Toh and L. Y. Tay, "Banking Chatbots: A Study on Technology Acceptance among Millennials in Malaysia", *Journal of Logistics, Informatics and Service Science*, vol. 9, no. 3, pp. 1–15, 2022.

[11] C. Zhu, "Research on Emotion Recognition-Based Smart Assistant System: Emotional Intelligence and Personalized Services", *Journal of System and Management Sciences*, vol. 13, no. 5, pp. 227–242, 2023. https://doi.org/10.33168/JSMS.2023.0515

[12] X. Zhiheng *et al.*, "Safety and Ethical Concerns of Large Language Models", in *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)*, 2023, pp. 9–16.

[13] L. Weidinger *et al.*, "Ethical and Social Risks of Harm from Language Models", arXiv preprint arXiv:2112.04359, 2021. https://doi.org/10.48550/arXiv.2112.04359

[14] M. Anastasia *et al.*, "An Exploratory Study on Ethics on the Internet", *Journal of System and Management Sciences*, vol. 13, no. 4, pp. 624–639, 2023. https://doi.org/10.33168/JSMS.2023.0438

[15] N. Mündler *et al.*, "Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation", arXiv preprint arXiv:2305.15852, 2023. https://doi.org/10.48550/arXiv.2305.15852

[16] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation", *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. http://dx.doi.org/10.1145/3571730

[17] R. Azamfirei *et al.*, "Large Language Models and the Perils of Their Hallucinations", *Critical Care*, vol. 27, no. 1, pp. 1–2, 2023. http://dx.doi.org/10.1186/s13054-023-04393-x

[18] K. Filippova, "Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 864–870. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.76

[19] J. Maynez *et al.*, "On Faithfulness and Factuality in Abstractive Summarization", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1906–1919. http://dx.doi.org/10.18653/v1/2020.acl-main.173

[20] C. Zhou *et al.*, "Detecting Hallucinated Content in Conditional Neural Sequence Generation", in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1393–1404. http://dx.doi.org/10.18653/v1/2021.findings-acl.120

[21] E. Shayegani *et al.*, "Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks", arXiv preprint arXiv:2310.10844, 2023. https://doi.org/10.48550/arXiv.2310.10844

[22] A. Zou *et al.*, "Universal and Transferable Adversarial Attacks on Aligned Language Models", arXiv preprint arXiv:2307.15043, 2023. https://doi.org/10.48550/arXiv.2307.15043

[23] A. Kumar *et al.*, "Certifying llm Safety Against Adversarial Prompting", arXiv preprint arXiv:2309.02705, 2023. https://doi.org/10.48550/arXiv.2309.02705

[24] X. Xu *et al.*, "An LLM can Fool Itself: A Prompt-Based Adversarial Attack", arXiv preprint arXiv:2310.13345, 2023. https://doi.org/10.48550/arXiv.2310.13345

[25] L. Schwinn *et al.*, "Adversarial Attacks and Defenses in Large Language Models: Old and New Threats", arXiv preprint arXiv:2310.19737, 2023. https://doi.org/10.48550/arXiv.2310.19737

[26] Y. Deng *et al.*, "Residual Energy-based Models for Text Generation", arXiv preprint arXiv:2004.11714, 2020. https://doi.org/10.48550/arXiv.2004.11714

[27] M. Li *et al.*, "Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training", arXiv preprint arXiv:1911.03860, 2019. https://doi.org/10.48550/arXiv.1911.03860

[28] M. Hall *et al.*, "A Systematic Study of Bias Amplification", arXiv preprint arXiv:2201.11706, 2022. https://doi.org/10.48550/arXiv.2201.11706

[29] L. Weidinger *et al.*, "Taxonomy of Risks Posed by Language Models", in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229. http://dx.doi.org/10.1145/3531146.3533088

[30] L. Moran. "Lawyer Cites Fake Cases Generated by ChatGPT in Legal Brief". https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/ (accessed December 21, 2023).

[31] H. J. Branch *et al.*, "Evaluating the Susceptibility of Pre-trained Language Models via Handcrafted Adversarial Examples", arXiv preprint arXiv:2209.02128, 2022. https://doi.org/10.48550/arXiv.2209.02128

[32] L. Daryanani, "How to Jailbreak ChatGPT". https://watcher.guru/news/how-to-jailbreak-chatgpt (accessed December 21, 2023).

[33] F. Perez and I. Ribeiro, "Ignore Previous Prompt: Attack Techniques for Language Models", arXiv preprint arXiv:2211.09527, 2022. https://doi.org/10.48550/arXiv.2211.09527

[34] K. Liu. "The Entire Prompt of Microsoft Bing Chat?! (Hi, Sydney.)". https://twitter.com/kliu128/status/1623472922374574080 (accessed December 10, 2023).

[35] Y. Wan *et al.*, " 'Kelly is a Warm Person, Joseph is a Role Model': Gender Biases in LLM-Generated Reference Letters", in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3730–3748. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.243

[36] S. Gehman *et al.*, "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3356–3369. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.301

[37] T. Schick *et al.*, "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP", *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 2021. https://doi.org/10.1162/tacl_a_00434

[38] F. Poletto *et al.*, "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review", *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477–523, 2020. https://doi.org/10.1007/s10579-020-09502-8

[39] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing", in *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10. http://dx.doi.org/10.18653/v1/W17-1101

[40] M. Zhang *et al.*, "SafeConv: Explaining and Correcting Conversational Unsafe Behavior", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 22–35. http://dx.doi.org/10.18653/v1/2023.acl-long.2

[41] M. Nadeem *et al.*, "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 (Volume 1: Long Papers)*, 2021, pp. 5356–5371. http://dx.doi.org/10.18653/v1/2021.acl-long.416

[42] C. Ziems *et al.*, "The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3755–3773. http://dx.doi.org/10.18653/v1/2022.acl-long.261

[43] H. Sun *et al.*, "On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark", in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 3906–3923. http://dx.doi.org/10.18653/v1/2022.findings-acl.308

[44] D. Hendrycks *et al.*, "Aligning AI With Shared Human Values", in *International Conference on Learning Representations*, 2021.

[45] D. Emelin *et al.*, "Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 698–718. http://dx.doi.org/10.18653/v1/2021.emnlp-main.54

[46] N. Lourie *et al.*, "Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-life Anecdotes", in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 15, pp. 13470–13479. http://dx.doi.org/10.1609/aaai.v35i15.17589

[47] J. Zhou *et al.*, "Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark", in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 3576–3591. http://dx.doi.org/10.18653/v1/2022.findings-emnlp.262

[48] N. Nangia *et al.*, "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)*, 2020, pp. 1953–1967. http://dx.doi.org/10.18653/v1/2020.emnlp-main.154

[49] J. Dhamala *et al.*, "Bold: Dataset and Metrics for Measuring Biases in Open-ended Language Generation", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 862–872. http://dx.doi.org/10.1145/3442188.3445924

[50] A. Parrish *et al.*, "BBQ: A Hand-built Bias Benchmark for Question Answering", in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105. http://dx.doi.org/10.18653/v1/2022.findings-acl.165

[51] E. Wulczyn *et al.*, "Ex Machina: Personal Attacks Seen at Scale", in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1391–1399. http://dx.doi.org/10.1145/3038912.3052591

[52] T. Davidson *et al.*, "Automated Hate Speech Detection and the Problem of Offensive Language", in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017, vol. 11, no. 1, pp. 512–515. http://dx.doi.org/10.1609/icwsm.v11i1.14955

[53] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", in *Proceedings of the NAACL Student Research Workshop*, 2016, pp. 88–93. http://dx.doi.org/10.18653/v1/N16-2013

[54] M. Zampieri *et al.*, "Predicting the Type and Target of Offensive Posts in Social Media", in *Proceedings of NAACL-HLT*, 2019, pp. 1415–1420. http://dx.doi.org/10.18653/v1/N19-1144

[55] Y. Zhang *et al.*, "A Taxonomy, Data Set, and Benchmark for Detecting and Classifying Malevolent Dialogue Responses", *Journal of the Association for Information Science and Technology*, vol. 72, no. 12, pp. 1477–1497, 2021. https://doi.org/10.1002/asi.24496

[56] A. Xenos *et al.*, "Context Sensitivity Estimation in Toxicity Detection", in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021, pp. 140–145. http://dx.doi.org/10.18653/v1/2021.woah-1.15

[57] A. Baheti *et al.*, "Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. http://dx.doi.org/10.18653/v1/2021.emnlp-main.397

[58] O. Shaikh *et al.*, "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4454–4470. http://dx.doi.org/10.18653/v1/2023.acl-long.244

[59] T. Markov *et al.*, "A Holistic Approach to Undesired Content Detection in the Real World", in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 12, pp. 15009–15018. http://dx.doi.org/10.1609/aaai.v37i12.26752

[60] D. Dale *et al.*, "Text Detoxification using Large Pre-trained Neural Models", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7979–7996. http://dx.doi.org/10.18653/v1/2021.emnlp-main.629

[61] A. Liu *et al.*, "DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6691–6706. http://dx.doi.org/10.18653/v1/2021.acl-long.522

[62] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners", *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[63] Y. Shen *et al.*, "ChatGPT and Other Large Language Models Are Double-edged Swords", *Radiology*, vol. 307, no. 2, p. e230163, 2023. https://doi.org/10.1148/radiol.230163

[64] S. Santhanam *et al.*, "Rome was Built in 1776: A Case Study on Factual Correctness in Knowledge-grounded Response Generation", arXiv preprint arXiv:2110.05456, 2021.

[65] H. Rashkin *et al.*, "Measuring Attribution in Natural Language Generation Models", *Computational Linguistics*, pp. 1–64, 2023. https://doi.org/10.1162/coli_a_00486

[66] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models", arXiv preprint arXiv:2108.07258, 2021. https://doi.org/10.48550/arXiv.2108.07258

[67] J. Luo *et al.*, "Zero-resource Hallucination Prevention for Large Language Models", arXiv preprint arXiv:2309.02654, 2023. https://doi.org/10.48550/arXiv.2309.02654

[68] C.-Y. Lin, "Rouge: A Package for Automatic Evaluation of Summaries", in *Text Summarization Branches Out*, 2004, pp. 74–81.

[69] C. Manning and H. Schutze, *Foundations of statistical natural language processing*, MIT press, 1999.

[70] K. Papineni *et al.*, "Bleu: A Method for Automatic Evaluation of Machine Translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135

[71] K. Shuster *et al.*, "Retrieval Augmentation Reduces Hallucination in Conversation", in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3784–3803. http://dx.doi.org/10.18653/v1/2021.findings-emnlp.320

[72] S. Lin *et al.*, "TruthfulQA: Measuring How Models Mimic Human Falsehoods", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3214–3252. http://dx.doi.org/10.18653/v1/2022.acl-long.229

[73] D. Muhlgay *et al.*, "Generating Benchmarks for Factuality Evaluation of Language Models", arXiv preprint arXiv:2307.06908, 2023. https://doi.org/10.48550/arXiv.2307.06908

[74] T. Liu *et al.*, "A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6723–6737. http://dx.doi.org/10.18653/v1/2022.acl-long.464

[75] Q. Cheng *et al.*, "Evaluating Hallucinations in Chinese Large Language Models", arXiv preprint arXiv:2310.03368, 2023. https://doi.org/10.48550/arXiv.2310.03368

[76] J. Li *et al.*, "Halueval: A Large-scale Hallucination Evaluation Benchmark for Large Language Models", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6449–6464. http://dx.doi.org/10.18653/v1/2023.emnlp-main.397

[77] X. Liang *et al.*, "Uhgeval: Benchmarking the Hallucination of Chinese Large Language Models Via Unconstrained Generation", arXiv preprint arXiv:2311.15296, 2023. https://doi.org/10.48550/arXiv.2311.15296

[78] A. Zeng *et al.*, "GLM-130B: An Open Bilingual Pre-trained Model", in *The Eleventh International Conference on Learning Representations*, 2022.

[79] S. Min *et al.*, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation", arXiv preprint arXiv:2305.14251, 2023. https://doi.org/10.48550/arXiv.2305.14251

[80] N. Lee *et al.*, "Factuality Enhanced Language Models for Open-ended Text Generation", *Advances in Neural Information Processing Systems*, vol. 35, pp. 34586–34599, 2022.

[81] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering", in *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021, pp. 874–880. http://dx.doi.org/10.18653/v1/2021.eacl-main.74

[82] P. Lewis *et al.*, "Retrieval-augmented Generation for Knowledge-intensive nlp Tasks", *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[83] R. Nakano *et al.*, "Webgpt: Browser-assisted Question-answering with Human Feedback", arXiv preprint arXiv:2112.09332, 2021. https://doi.org/10.48550/arXiv.2112.09332

[84] S. Borgeaud *et al.*, "Improving Language Models by Retrieving from Trillions of Tokens", in *Proc. of the International Conference on Machine Learning*, 2022, pp. 2206–2240.

[85] W. Shi *et al.*, "Replug: Retrieval-augmented Black-box Language Models", arXiv preprint arXiv:2301.12652, 2023. https://doi.org/10.48550/arXiv.2301.12652

[86] S. Zhang *et al.*, "Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment", arXiv preprint arXiv:2305.13669, 2023. https://doi.org/10.48550/arXiv.2305.13669

[87] B. Peng *et al.*, "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback", arXiv preprint arXiv:2302.12813, 2023. https://doi.org/10.48550/arXiv.2302.12813

[88] X. Li *et al.*, "Chain of Knowledge: A Framework for Grounding Large Language Models with Structured Knowledge Bases", arXiv preprint arXiv:2305.13269, 2023. https://doi.org/10.48550/arXiv.2305.13269

[89] J. Cui *et al.*, "Chatlaw: Open-source Legal Large Language Model with Integrated External Knowledge Bases", arXiv preprint arXiv:2306.16092, 2023. https://doi.org/10.48550/arXiv.2306.16092

[90] S. Jha *et al.*, "Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting", in *Proc. of the 2023 IEEE International Conference on Assured Autonomy (ICAA)*, 2023, pp. 149–152. http://dx.doi.org/10.1109/ICAA58325.2023.00029

[91] Y.-S. Chuang *et al.*, "Dola: Decoding by Contrasting Layers Improves Factuality in Large Language Models", arXiv preprint arXiv:2309.03883, 2023. https://doi.org/10.48550/arXiv.2309.03883

[92] K. Meng *et al.*, "Locating and Editing Factual Associations in GPT", *Advances in Neural Information Processing Systems*, vol. 35, pp. 17359–17372, 2022.

[93] M. Elaraby *et al.*, "Halo: Estimation and Reduction of Hallucinations in Open-source Weak Large Language Models", arXiv preprint arXiv:2308.11764, 2023. https://doi.org/10.48550/arXiv.2308.11764

[94] B. Workshop *et al.*, "Bloom: A 176b-parameter Open-access Multilingual Language Model", arXiv preprint arXiv:2211.05100, 2022. https://doi.org/10.48550/arXiv.2211.05100

[95] Y. Liu *et al.*, "Prompt Injection Attack Against LLM-integrated Applications", arXiv preprint arXiv:2306.05499, 2023. https://doi.org/10.48550/arXiv.2306.05499

[96] S. Abdelnabi *et al.*, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection", in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023, pp. 79–90.

[97] F. Perez and I. Ribeiro, "Ignore Previous Prompt: Attack Techniques For Language Models", in NeurIPS ML Safety Workshop, 2022.

[98] G. Apruzzese *et al.*, " 'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice", in *Proc. of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023, pp. 339–364. http://dx.doi.org/10.1109/SaTML54575.2023.00031

[99] Z. Li *et al.*, "On the Feasibility of Specialized Ability Stealing for Large Language Code Models", arXiv preprint arXiv:2303.03012, 2023. https://doi.org/10.48550/arXiv.2303.03012

[100] Z. Xiang, F. Jiang, Z. Xiong, B. Ramasubramanian, R. Poovendran, and B. Li, "BadChain: Backdoor Chain-of-Thought Prompting for Large Language Models", in NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly, 2023.

[101] H. Qiu *et al.*, "Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models", arXiv preprint arXiv:2307.08487, 2023. https://doi.org/10.48550/arXiv.2307.08487

[102] X. Shen *et al.*, " 'Do Anything Now': Characterizing and Evaluating in-the-wild Jailbreak Prompts on Large Language Models", arXiv preprint arXiv:2308.03825, 2023. https://doi.org/10.48550/arXiv.2308.03825

[103] E. Crothers *et al.*, "Machine-generated Text: A Comprehensive Survey of Threat Models and Detection Methods", *IEEE Access*, 2023. http://dx.doi.org/10.1109/ACCESS.2023.3294090

[104] N. Jain *et al.*, "Baseline Defenses for Adversarial Attacks Against Aligned Language Models", arXiv preprint arXiv:2309.00614, 2023. https://doi.org/10.48550/arXiv.2309.00614S

[105] A. Helbling *et al.*, "Llm Self Defense: By Self Examination, llms Know They are Being Tricked", arXiv preprint arXiv:2308.07308, 2023. https://doi.org/10.48550/arXiv.2308.07308

[106] M. Ung *et al.*, "SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6462–6481. http://dx.doi.org/10.18653/v1/2022.acl-long.447

[107] E. Kasneci *et al.*, "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education", *Learning and Individual Differences*, vol. 103, p. 102274, 2023. http://dx.doi.org/10.1016/j.lindif.2023.102274

[108] B. Meskó and E. J. Topol, "The Imperative for Regulatory Oversight of Large Language Models (or generative AI) in Healthcare", *NPJ Digital Medicine*, vol. 6, no. 1, p. 120, 2023. http://dx.doi.org/10.1038/s41746-023-00873-0

[109] M. Nasr *et al.*, "Scalable Extraction of Training Data from (Production) Language Models", arXiv preprint arXiv:2311.17035, 2023. https://doi.org/10.48550/arXiv.2311.17035

[110] T. South *et al.*, "Transparency by Design for Large Language Models", *Computational Legal Futures, Network Law Review.(2023)*, 2023.

[111] Z. Wu *et al.*, "Transparency Helps Reveal When Language Models Learn Meaning", *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 617–634, 2023. http://dx.doi.org/10.1162/tacl_a_00565

[112] E. Ferrara, "Should Chatgpt be Biased? Challenges and Risks of Bias in Large Language Models", arXiv preprint arXiv:2304.03738, 2023. https://doi.org/10.48550/arXiv.2304.03738

[113] D. R. Cotton *et al.*, "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT", *Innovations in Education and Teaching International*, pp. 1–12, 2023. http://dx.doi.org/10.1080/14703297.2023.2190148

[114] Y. L. Sang, "Mobile Digital Forensics Framework to Increase Security Level of for Smartphone User", *Journal of Logistics, Informatics and Service Science*, vol. 9, no. 1, pp. 68–84, 2022. https://doi.org/10.33168/LISS.2022.0106

[115] J. Mökander *et al.*, "Auditing Large Language Models: A Three-layered Approach", *AI and Ethics*, pp. 1–31, 2023. http://dx.doi.org/10.1007/s43681-023-00289-2

[116] J. Mökander and L. Floridi, "Ethics-based Auditing to Develop Trustworthy AI", *Minds and Machines*, vol. 31, no. 2, pp. 323–327, 2021. http://dx.doi.org/10.1007/s11023-021-09557-8

[117] I. D. Raji and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 429–435. http://dx.doi.org/10.1145/3306618.3314244

Received: December 2023
Revised: February 2024
Accepted: February 2024

Contact addresses:
Zhengjie Gao
School of Electronic and Information Engineering
Geely University of China
Chengdu
China
e-mail: gaozhengjie@guc.edu.cn

Xuanzi Liu
School of Electronic and Information Engineering
Geely University of China
Chengdu
China
e-mail: liuzixuan@guc.edu.cn

Yuanshuai Lan
School of Electronic and Information Engineering
Geely University of China
Chengdu
China
e-mail: lanyuanshuai@guc.edu.cn

Zheng Yang
School of Electronic and Information Engineering
Geely University of China
Chengdu
China
e-mail: yyyzword@126.com

Zhengjie Gao obtained B.S. and M.S. degrees in computer science and technology from Chengdu University of Information Technology in 2017 and 2020, respectively. Since November 2023, he has been a lecturer at the School of Electronic and Information Engineering, Geely University of China. His current research interests include sentiment analysis and large language models.

Xuanzi Liu obtained a B.S. degree in electronic engineering from the University of Central Lancashire in 2022, and an M.S. degree in communication networks and signal processing from the University of Bristol in 2023. She is currently a teaching assistant at the School of Electronic and Information Engineering, Geely University of China. Her research interests include communication and artificial intelligence.

Yuanshuai Lan obtained a B.S. degree in electronic science and technology from the Chengdu College of the University of Electronic Science and Technology of China in 2018. M.S. degrees in agricultural engineering and information technology in 2022. He is currently a lecturer at the School of Electronic and Information Engineering, Geely University of China. His current research interests include computer vision and artificial intelligence.

Zheng Yang obtained a B.S. degree in computer science and technology from Wuhan University of Technology in 1990, and an M.S. degree in management engineering from Sichuan University in 2015. He has worked at Lenovo and IBM and is currently an associate professor at the School of Electronic and Information Engineering, Geely University of China. His research interests include digital transformation, big dana, and artificial intelligence.