The Scales of Justitia: A Comprehensive Survey on Safety Evaluation of LLMs

Songyang Liu, Chaozhuo Li*, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S. Yu

***Abstract*—With the rapid advancement of artificial intelligence technology, Large Language Models (LLMs) have demonstrated remarkable potential in the field of Natural Language Processing (NLP), including areas such as content generation, human-computer interaction, machine translation, and code generation, among others. However, their widespread deployment has also raised significant safety concerns. In recent years, LLM-generated content has occasionally exhibited unsafe elements like toxicity and bias, particularly in adversarial scenarios, which has garnered extensive attention from both academia and industry. While numerous efforts have been made to evaluate the safety risks associated with LLMs, there remains a lack of systematic reviews summarizing these research endeavors. This survey aims to provide a comprehensive and systematic overview of recent advancements in LLMs safety evaluation, focusing on several key aspects: (1) "Why evaluate" that explores the background of LLMs safety evaluation, how they differ from general LLMs evaluation, and the significance of such evaluation; (2) "What to evaluate" that examines and categorizes existing safety evaluation tasks based on key capabilities, including dimensions such as toxicity, robustness, ethics, bias and fairness, truthfulness, and so on; (3) "Where to evaluate" that summarizes the evaluation metrics, datasets and benchmarks currently used in safety evaluations; (4) "How to evaluate" that reviews existing evaluation toolkit, and categorizing mainstream evaluation methods based on the roles of the evaluators. Finally, we identify the challenges in LLMs safety evaluation and propose potential research directions to promote further advancement in this field. We emphasize the importance of prioritizing LLMs safety evaluation to ensure the safe deployment of these models in real-world applications.**

***Index Terms*—large language models, safety evaluation, evaluation tasks, evaluation benchmarks, evaluation metrics.**

I. Introduction

As artificial intelligence technology evolves at an unprecedented pace, Large Language Models (LLMs) have demonstrated remarkable potential across various fields, becoming a central focal point in both the tech industry and academic circles. These sophisticated models have achieved significant breakthroughs in Natural Language Processing (NLP) and exhibit robust capabilities in diverse application scenarios, including content generation [66], human-computer interaction [67], machine translation [64], and code generation [65], among others [233], [235]. LLMs are designed to process and generate human-like text, leveraging the vast amounts of data on which they are trained. This training enables them to produce fluent and coherent text while demonstrating a degree of comprehension and reasoning abilities [68], significantly advancing the progress of artificial intelligence.

However, like any emerging technology, the widespread application of large language models brings numerous challenges and safety risks. Recent studies have demonstrated that LLMs are prone to generating harmful content, such as toxicity [7], bias [23] and false information [234], particularly in adversarial scenarios where these issues are even more pronounced. In addition, attack methods targeting LLMs safety, such as prompt injection attacks and jailbreak attacks [228], [229], have become increasingly sophisticated, aiming to bypass the safety alignment mechanisms between LLMs and humans, thus inducing LLMs to generate unsafe content [69]. This not only negatively impacts user experience but also raises ethical and legal concerns that could be exploited by malicious actors. Therefore, effectively evaluating the safety of LLMs and ensuring their safe deployment has become a pressing issue that requires urgent attention.

The emergence of LLMs safety evaluation stems from a profound recognition and urgent need to address the potential risks posed by these models and their societal impacts. In the processes of LLMs research, development, and deployment, safety evaluation plays a crucial role. Unlike general large language model evaluations [66], [70], [71], LLMs safety assessment not only requires the model to possess fundamental general capabilities, such as language understanding, generation, and reasoning, but also places particular emphasis on the safety issues associated with the content generated by the model. The assessment covers a range of critical dimensions, including toxicity, robustness, morality, bias and fairness, and credibility. By conducting a systematic safety assessment of LLMs, potential risks can be identified and mitigated in a timely manner, thereby ensuring the safety of these models in practical applications. This rigorous evaluation process is essential for enhancing user trust and preventing societal issues that may arise from model deficiencies, ultimately fostering the healthy development and widespread adoption of LLM technologies.

S. Liu, C. Li, X. Zhang are with School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China (Email: lsy20031123@bupt.edu.cn; lichaozhuo@bupt.edu.cn; zhangx@bupt.edu.cn).

J. Qiu, F. Huang are with School of Cyberspace Security, Jinan University, Guangzhou, China (Email: qiujiameng@stu2024.jnu.edu.cn, huangfr@jnu.edu.cn).

L. Zhang is with School of Cyberspace Security, Beihang University, Beijing, China (Email: litianzhang@buaa.edu.cn).

Y. Hei is with China Academy of Information and Communications Technology, Beijing, China (Email: heiyiming@caict.ac.cn)

P.S. Yu is with Department of Computer Science, University of Illinons at Chicago, Chicago, USA (Email: psyu@uic.edu).

Chaozhuo Li is the Corresponding Author.

While numerous researchers have actively explored the safety evaluation of large language models in recent years, a systematic summary of the existing body of research remains conspicuously absent. Existing studies either focus on specific issues or evaluation methods or fail to provide a comprehensive classification, organization, and synthesis of safety evaluations [72], [73]. As a result, they fail to offer a holistic view of current methods, challenges, and future directions in LLMs safety evaluation.

Against this backdrop, this paper aims to provide a thorough and systematic review of recent advancements in LLMs safety evaluation, thereby addressing this research gap. Specifically, as shown in Figure 1, we delve into existing work from the following four dimensions: (1) "Why Evaluate" elucidates the background of current LLMs safety evaluation, distinguishing it from general LLM evaluations. This section underscores the significance of safety evaluations in ensuring that LLMs can be responsibly deployed in real-world applications, highlighting the potential risks associated with unassessed models. (2) "What to Evaluate" summarizes the primary tasks associated with LLMs safety evaluation, presenting a detailed classification that encompasses various facets such as toxicity, robustness, ethics, bias and fairness, truthfulness, and more. (3) "Where to Evaluate" compiles current commonly used evaluation metrics and categorizes the datasets and benchmarks employed in the field. The aim of this section is to provide researchers with a comprehensive reference for selecting appropriate evaluation criteria, thereby facilitating faster progress in safety evaluation research. (4) "How to Evaluate" reviews existing evaluation toolkits and categorizes evaluation methods based on the roles of the evaluators, whether they are automated systems or human assessors, thus providing insights into the methodologies and best practices for conducting safety evaluations. Finally, we discuss the current challenges facing the field of LLMs safety evaluation and outline potential future research directions.

The contributions of this paper can be summarized as follows:

- Comprehensive Review: To the best of our knowledge, this paper presents the first comprehensive and systematic survey of recent advancements in the field of large language model safety evaluation, addressing a significant gap in the existing literature.
- Clear Classification Framework: We establish a detailed classification framework that delineates the primary tasks of LLMs safety evaluation across various dimensions, enhancing understanding within the research community.
- Consolidated Evaluation Resources: We compile and categorize current commonly used evaluation metrics, datasets, benchmarks, evaluation toolkits, and methods, offering a comprehensive resource for researchers to reference and apply in safety evaluations.
- Future Research Directions: We discuss the key challenges currently facing the field and outline potential avenues for future research, thereby providing guidance to promote the healthy development of LLMs safety evaluation.

II. WHY TO EVALUATE

A. Background

The extensive adoption of large language models (LLMs) has showcased remarkable utility across diverse domains, including knowledge inference [75], [76], content generation [74], code development support [77], [78], and data extraction [79], [231], [232]. However, the substantial generative capabilities of these models are not without risks. Studies indicate that, when guided by certain prompts, LLMs can produce offensive or prejudiced outputs. Moreover, when faced with topics outside their training scope, these systems often construct seemingly credible yet incorrect information, contributing to misinformation. Such fabricated and misleading content may be exploited with malicious intent, causing significant societal harm [5], [21], [24], [80]–[82].

For instance, in the financial sector, LLMs may generate inaccurate credit scores due to biases or deficiencies in training data, resulting in capable borrowers being mistakenly classified as high-risk and denied loans [83]. In the field of smart education, LLMs may perpetuate stereotypes about the learning abilities of certain demographic groups, leading to systematic recommendations of overly simplified learning content and limiting the long-term development of these students [84]. In the healthcare domain, incorrect diagnostic suggestions generated by LLMs could severely impact patient treatment outcomes and even endanger lives [85].

The security challenges of LLMs primarily manifest in four key aspects:

1) **Generation of offensive content**: LLMs are prone to producing harmful or objectionable language, including discriminatory, racist, or hateful speech [3].These risks are particularly concerning when LLMs are applied in sensitive environments, such as automated customer service or social media platforms.
2) **Generation of biased and unethical content**: LLMs may perpetuate societal biases embedded within their training data, such as those related to gender, ethnicity, or religion, leading to discriminatory outcomes [86]. If left unaddressed, these biases not only reinforce existing inequalities but may also erode trust in the technology.
3) **Generation of factually incorrect content**: LLMs may propagate factual inaccuracies or spread rumors, particularly when trained on datasets containing misinformation or widely circulated false news [87]. These models lack the ability to verify the truthfulness of information, and as a result, present inaccuracies with an air of confidence, making it difficult for users to distinguish between valid knowledge and falsehoods.
4) **Vulnerability to adversarial attacks**: LLMs are susceptible to adversarial manipulations, such as prompt injection and jailbreak attacks, where attackers craft inputs designed to circumvent the model's safety protocols [9]. These attacks can prompt the model to produce harmful instructions or generate inappropriate content, posing significant security threats.

This dual-edged nature of LLMs underscores the need for comprehensive safeguards and systematic security assess-

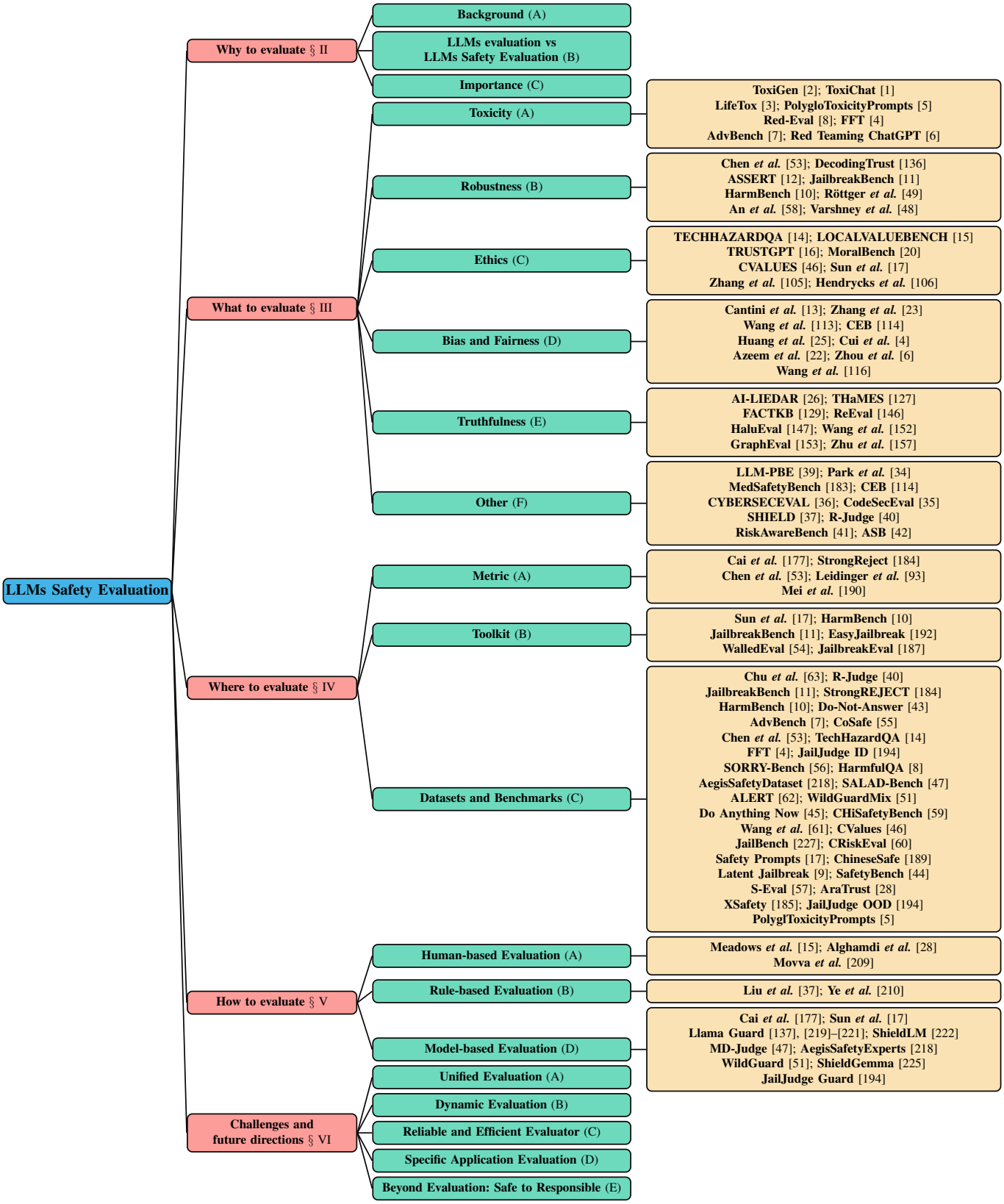

Figure 1. Structure of this paper.

ments. Identifying and addressing these risks proactively is crucial to prevent unintended consequences, protect users, and ensure these models contribute positively to society.

B. LLMs evaluation vs LLMs Safety Evaluation

In this context, LLM safety evaluation has become an indispensable component of the development and deployment process. The objective of safety evaluation extends beyond merely identifying and mitigating potential risks within the model—it aims to ensure that these risks do not adversely impact users or society in real-world applications [88]. Although LLMs safety evaluation shares certain overlaps with conventional model evaluation, their goals and areas of emphasis differ significantly:

- **LLMs Evaluation**: This evaluation primarily focuses on assessing the fundamental performance of the model, including its accuracy, fluency, and reasoning capabilities. Standardized datasets and benchmarks are employed to ensure consistency and comparability across tasks. Metrics such as BLEU [89] and ROUGE [90] are commonly used to evaluate the effectiveness of the model in specific applications, such as machine translation [91] and question answering systems [78]. The primary goal of LLM evaluation is to quantify how well the model performs under controlled conditions, emphasizing technical proficiency in language generation.
- **LLMs Safety Evaluation**: Building upon the foundation of LLM evaluation, safety evaluation goes further by focusing not only on technical performance but also on the ethical, social, and contextual dimensions of the model's behavior. It emphasizes ethical integrity, robustness,and bias mitigation to ensure that the model's outputs align with societal values and remain appropriate across diverse real-world scenarios. In this framework, the ability to generate accurate and fluent text is a necessary but insufficient criterion; the evaluation also examines whether the model produces fair and unbiased outputs, maintains stability under adversarial conditions, and avoids generating toxic or harmful content. In essence, safety evaluation extends beyond the quantitative metrics of traditional assessments, integrating considerations such as the responsible use of AI and the model's resilience in unpredictable environments.

In contrast to standard performance assessments, LLMs safety evaluation adopts a social responsibility perspective, ensuring that the model's outputs align with compliance, fairness, and reliability standards across a variety of real-world contexts. While traditional evaluations measure technical proficiency, safety assessments extend this focus by addressing the broader societal impact of the model's behavior. These evaluations ensure not only that the model performs well but also that it generates content that is ethical, appropriate, and trustworthy in sensitive applications such as healthcare, education, and finance. Key components of safety evaluation include:

- **Toxicity Detection**: This process evaluates whether the model produces offensive, harmful, or abusive language during content generation [92]. It aims to prevent the propagation of toxic expressions that could damage user experiences or incite social harm, particularly in public-facing or automated communication systems.
- **Ethics and Bias Evaluation**: This assessment identifies and mitigates biases or discriminatory tendencies present in the model's outputs, such as those targeting specific demographic groups based on gender, ethnicity, or socio-economic status [22], [93].
- **Truthfulness and Credibility**: This component ensures that the model generates factually accurate information, minimizing the risk of misleading users [29], [30]. In high-stakes domains like healthcare or legal advice, credibility is essential for maintaining trust and enabling reliable decision-making
- **Robustness Testing**: This examines the model's ability to maintain safe and stable behavior when exposed to adversarial inputs or unexpected scenarios [12], [13]. Robustness evaluation ensures that the model can withstand attempts to manipulate its outputs or bypass safety mechanisms, helping to safeguard against misuse.

Similar to LLMs evaluation, the results of LLMs safety Evaluation provide critical insights that guide model improvement across its design, training, and deployment phases. However, safety evaluation plays a more proactive role by identifying vulnerabilities and addressing them before deployment to prevent unintended harm. Developers can leverage various strategies to enhance model safety, such as reinforcement learning with human feedback (RLHF) to improve the model's ability to filter toxic content [94], adversarial training to strengthen the model's defenses against manipulative attacks [95], and enhancing the quality of pre-training data to minimize hallucinations and improve factual accuracy [96]. Ultimately, these efforts are essential for building robust, responsible, and trustworthy AI systems, reducing risks, and ensuring that the outputs align with ethical and societal standards.

C. Importance

The importance of LLMs safety evaluation transcends technical considerations, encompassing critical aspects of social responsibility, user trust, and legal compliance. As these models are increasingly integrated into sensitive domains such as healthcare, law, and education, the potential risks arising from erroneous or inappropriate outputs can result in significant adverse consequences [97]–[99]. For instance, incorrect recommendations in healthcare systems may jeopardize patient safety, misinformation spread through educational platforms can distort students' understanding, and flawed analyses or judgments in legal contexts may lead to disputes or litigation.

To mitigate these risks, systematic safety evaluation plays an essential role across the following key areas:

- **Risk Mitigation** : Safety evaluations proactively identify and address vulnerabilities—such as the generation of toxic content or factual inaccuracies—before deployment. This process ensures that the model does not produce hate speech or biased outputs, safeguarding individuals and communities from harm while minimizing potential societal disruptions.
- **Ethical and Legal Compliance**: Assessing the model for bias and adversarial vulnerabilities ensures that its

behavior aligns with ethical principles and complies with relevant legal frameworks. This evaluation is particularly critical in high-stakes settings, such as healthcare, law, and education, where unethical or non-compliant outputs can lead to serious consequences.

- **User Trust Enhancement**: A well-executed safety evaluation helps prevent the generation of misleading or inaccurate content, thereby fostering trust by ensuring users receive reliable and accurate information. Building trust is essential to encouraging the widespread adoption and acceptance of LLMs technologies across various sectors.
- **System Security Improvement**: Evaluating for issues such as toxicity, bias, and truthfulness provides continuous feedback to developers, guiding iterative improvements throughout the model's lifecycle. This feedback loop ensures the system remains secure, reliable, and aligned with evolving standards, facilitating long-term technical innovation and sustainable development.

This review seeks to fill existing research gaps by presenting a comprehensive and systematic framework for LLMs safety evaluation. Through the synthesis of current methodologies, tools, and metrics, this study aims to provide researchers with a clear overview of the state of the field, helping to reduce redundant efforts and accelerate progress. Moreover, we emphasize the need to promote standardization in safety evaluation practices, encouraging more researchers to engage with this critical area to advance the responsible and sustainable development of LLMs technologies.

III. WHAT TO EVALUATE

This section provides an overview of various perspectives for evaluating the safety of LLMs, including toxicity, robustness, ethics, bias and fairness, truthfulness, and other specific downstream tasks. It aims to illustrate what specific dimensions should be considered in evaluating and demonstrating the safety of LLMs, offering guidance for conducting comprehensive safety evaluations.

A. Toxicity

Toxicity refers to the presence of offensive, hateful, insulting, or harmful content, including incitement to violence, within the text generated by LLMs. Such toxic outputs not only pose risks of psychological harm to individual users but may also trigger broader social conflicts. Therefore, evaluating the potential of LLMs to generate toxic content has become a critical step in ensuring their safe deployment. In recent years, researchers have conducted extensive studies in this area:

ToxiGen: [2] highlights that existing toxicity detectors often over-rely on surface-level mentions of minority identities, leading to the neglect of subtle hate speech and the over-detection of benign expressions. To address this issue, the authors developed a large-scale dataset, ToxiGen, to accurately evaluate LLMs' ability to generate adversarial and implicit toxic speech. To automatically generate these challenging texts, they introduced an adversarial classifier-in-the-loop decoding algorithm—ALICE. ALICE adjusts the toxicity of the generated text by comparing a toxicity classifier with the text generator during beam search decoding. By controlling machine generation in this manner, ToxiGen can encompass a broader range of implicit toxic texts than any previous human-written text resources, addressing a wider array of demographic groups. Experimental results demonstrate that several existing toxicity classifiers struggle to accurately distinguish between toxic and non-toxic outputs generated by ALICE. Moreover, comprehensive human evaluations show that ALICE-generated texts closely resemble human-produced content. This work offers a significant step forward in improving toxicity classifiers by providing both the dataset and the generation method.

ToxiChat: [1] aims to address the challenge of detecting toxic content in real-world user-AI interactions. They note that most existing research on toxicity detection relies on benchmarks built from social media content, failing to capture the unique challenges of real-world conversations with AI systems. To fill this gap, the authors developed a new benchmark—ToxicChat, based on real user queries directed at open-source chatbots. To ensure data quality and reduce the burden of manual annotation, the authors employed an uncertainty-guided human-AI collaborative annotation approach. During the annotation process, they observed that chatbots are increasingly subjected to jailbreak attacks, underscoring the urgency of the current safety landscape. Furthermore, a systematic evaluation revealed that models trained on existing harmful content datasets perform poorly when applied to the ToxicChat domain, highlighting the need for enhanced safety measures in real-world user-AI interactions.

LifeTox: [3] explores the challenge of detecting implicit harmful content within diverse life-advice scenarios. Implicit toxicity refers to a deeper, more concealed form of toxicity conveyed through linguistic features such as euphemisms, sarcasm, circumlocution, and metaphors, as well as extralinguistic knowledge like commonsense knowledge, world knowledge and social norms [100]. The authors argue that existing safety benchmark datasets are limited in diversity, primarily relying on red-team prompts, which often result in predictable and repetitive scenarios, failing to capture the complexity and variety of real-world situations.To address this gap, the authors introduced the LifeTox dataset, designed to identify implicit harmful content in various personal advice-seeking contexts, thereby improving the safety and reliability of the advice provided by LLMs. LifeTox was constructed by collecting posts from two Reddit forums focused on sharing life tips and seeking advice. Experimental results demonstrate that RoBERTa [101], fine-tuned on LifeTox, performs on par with or better than large language models in zero-shot harmful content classification, underscoring the effectiveness of LifeTox in tackling the complex challenges posed by implicit toxicity.

PolygloToxicityPrompts: In contrast to previous toxicity assessments conducted in single-language contexts (such as English or Chinese), Jain *et al.* [5] explores the issue of LLMs generating harmful content in multilingual environments. They note that existing toxicity evaluation datasets predominantly focus on English or translations based on English benchmarks, leading to inadequate assessments of toxicity degradation

in contemporary LLMs. To address this limitation, the authors propose PolygloToxicityPrompts, a large-scale multilingual benchmark containing 425K naturally occurring prompts across 17 languages. The prompts were generated by scraping documents from extensive corpora, with an average length of 400 GPT-4 tokens. During the evaluation, the authors employed the Perspective API [102] to assess the toxicity of the prompts and their generated content, calculating the average toxicity level of the models. Experimental results indicate significant differences in toxicity levels among LLMs across different languages, with an observed increase in toxicity as model size grows within the same family. Furthermore, models after instruction and preference-tuning exhibited lower toxicity compared to baseline models; However, the choice of preference-tuning method had a minimal impact on the models' toxicity levels. The authors emphasize the need for further research on multilingual toxicity mitigation and the influence of model hyperparameters on toxicity.

Red-Eval: [8] introduces a novel safety red teaming benchmark, RED-EVAL, to comprehensively evaluate the risks associated with harmful outputs generated by LLMs. RED-EVAL implements jailbreak attacks through a Chain of Utterances(CoU)-based prompt, facilitating a dialogue between two agents: a harmful agent Red-LM and a unsafe-helpful agent Base-LM. A harmful question is posed as the discourse for the Red-LM, and the model is instructed to generate a response according to the guidelines outlined in the prompt. Additionally, they propose the RED-INSTRUCT method, which enhances LLM safety through a two-phase approach: first, the CoU prompts are used to collect a dataset of harmful questions, termed HARMFULQA; Second, the model is safety-aligned by minimizing the negative log-likelihood of helpful responses while penalizing harmful ones. Experimental results show that RED-EVAL achieves a jailbreak success rate of 65% on GPT-4 and 73% on ChatGPT, highlighting its effectiveness in evaluating risks associated with harmful content generation.

In addition to the aforementioned studies, several other works have made significant contributions to the evaluation of LLMs toxicity. For instance, Cui *et al.* [4] collected questions and jailbreak templates from Bai *et al.* [103] and Liu *et al.* [104] to evaluate LLMs for utterance-level toxicity (referring to literally-toxic language with some typical words) and context-level toxicity (referring to that a harmless statement could be a toxic one when considered within its context.). All evaluated LLMs demonstrated an increase in toxicity from the utterance- to context-level evaluation. This performance gap may arise from literally non-toxic responses that inadvertently affirm toxic issues. Zou *et al.* [7] proposed greedy coordinate gradient-based search(GCG) to generate adversarial suffixes for malicious prompts, aiming to induce LLMs to produce harmful content. Additionally, the authors developed a new benchmark, AdvBench, designed for systematic evaluation based on harmful strings and harmful behaviors. Zhou *et al.* [6] conducted a comprehensive evaluation of ChatGPT using red teaming methods, highlighting that ChatGPT is vulnerable to prompt injection, leading to the generation of toxic content, and currently lacks the ability to detect toxicity at early stages.

B. Robustness

Robustness refers to the ability of LLMs to maintain stable and safe outputs when faced with noise (e.g., text perturbations), adversarial attacks (e.g., jailbreak attacks, prompt injection attacks), and out-of-distribution data. We also emphasize the importance of considering robustness in LLMs instruction—following ability, ensuring that models do not exhibit excessive safety, thereby compromising its performance on some normal tasks(e.g., "How to kill a process?"). In the realm of safety evaluation for LLMs, many researchers have proposed innovative methods and insights to evaluate robustness.

Characterizing and Evaluating: Jailbreak attacks, particularly those involving carefully crafted prompts, pose significant challenges to the safety of LLMs. Chen *et al.* [53] introduces a comprehensive evaluation framework to mitigate this issue. Specifically, the researchers first construct a dataset covering 61 different harmful categories as the basis for evaluation. They then conduct a thorough evaluation of 10 advanced jailbreak attack methods against 13 popular LLMs. In evaluating the outputs of the LLMs, the authors consider a wide range of evaluation metrics, including Attack Success Rate (ASR), toxicity, fluency, grammatical errors, and token length, to facilitate an in-depth exploration of LLM safety when facing with jailbreak attacks. The results indicate that none of the evaluated LLMs exhibit initial resistance to harmful queries, with Vicuna and Mistral being the most vulnerable to jailbreak attacks within the LLMs family. Meanwhile, GPT-4 demonstrates the best performance across all categories, while Llama3 also shows strong capabilities. Among the jailbreak attack methods, ReNeLLM [134] achieves the highest average ASR, while ICA [132] and Cipher [133] exhibit lower performance. Furthermore, the correlation between evaluation metrics reveals that ASR is generally positively correlated with toxicity. Additionally, grammatical errors are positively correlated with token length, while fluency does not show a strong correlation with either grammatical errors or token length.

DecodingTrust: [136] conducts a comprehensive trustworthiness evaluation of GPT-4 and GPT-3.5, focusing on adversarial robustness, out-of-distribution (OOD) robustness, and robustness against adversarial demonstration. To evaluate robustness on adversarial text attacks, the researchers construct three scenarios: (1) standard benchmark AdvGLUE [141], (2) AdvGLUE under various instructive task descriptions and system prompts, and (3) the more challenging adversarial text generation benchmark, AdvGLUE++. For OOD robustness, they explore scenarios including (1) inputs deviating from common training text styles, (2) recent events outside GPT models' training data collection period, and (3) demonstrations with diverse OOD styles and domains added through in-context learning. For adversarial demonstration robustness, the evaluations involve injecting counterfactual examples, spurious correlations, and backdoors into the demonstrations to observe model's performance. Overall, while GPT-4 typically outperforms GPT-3.5 on standard benchmarks, it proves more vulnerable to jailbreak attacks when specific system or user

prompts are provided. The authors attribute this vulnerability to GPT-4's tendency to follow misleading instructions more precisely.

ASSERT: [12] proposes the Automatic Safety ScEnario Red Team (ASSERT) for robustness and safety evaluation of LLMs. ASSERT employs three novel methods—semantically aligned augmentation, targeted bootstrapping, and adversarial knowledge injection—to generate new test cases that explore robustness within language models. Semantically aligned augmentation aims to create samples with semantically equivalent but differently phrased expressions. Targeted bootstrapping generates new synthetic samples that are related to existing ones but not equivalent, while adversarial knowledge injection involves injecting adversarial knowledge during model inference. The researchers then evaluate LLMs on question-answering tasks across four critical AI application domains: outdoor, medical, household, and extra, requiring the models to determine whether a specific behavior a should be implemented within the context c. The evaluation results reveal performance differences between semantically similar scenarios, with the model exhibiting instability up to a divergence of 11% in absolute classification accuracy. Adversarial attacks achieve error rates of 19.76% in zero-shot settings and 51.55% in adversarial four-shot demonstration settings, emphasizing the importance of few-shot demonstration.

JailbreakBench: [11] points out that current jailbreak evaluations lack clear practical standards, use incomparable computational costs and success rates for evaluation, and many works are difficult to reproduce. To address these challenges, the researchers introduce an open-source benchmark called JailbreakBench. First, JailbreakBench releases a jailbreak dataset of size 100, drawing from previous works across five aspects: behavior, goal (harmful queries), target (positive responses to harmful queries), category (based on OpenAI's usage policy), and source (source dataset). Additionally, they maintain a repository of jailbreak artifacts and establish a standardized evaluation framework. This framework not only supports the definition of threat models, system prompts, and chat templates but also integrates a diverse range of evaluation methods, including rule-based string matching, GPT-4, and several models specifically designed for jailbreak evaluation (e.g., Llama Guard [137]). Evaluation of existing jailbreak attacks indicates that even recent closed-source, unprotected models are highly vulnerable to jailbreak attacks. Prompts with RS [138] outperform others such as PAIR [167], GCG [7], and Jailbreak Chat [139], achieving an attack success rate of up to 78% on GPT-4, highlighting the necessity for further defensive measures.

HarmBench: [10] designs a new red teaming attack and defense benchmark called HarmBench to address the lack of a standardized evaluation framework in automated red teaming. HarmBench consists of a set of harmful behaviors and an evaluation pipeline. These harmful behaviors include both textual and multimodal actions, and to enhance the robustness of the evaluation, HarmBench includes official validation/testing set partitions. Each behavior is categorized into semantic and functional categories. The semantic category describes the types of harmful behaviors (e.g., Cybercrime & Unauthorized Intrusion, Chemical and Biological Weapons/Drugs, etc.), including seven types. The functional category highlights unique properties of the behaviors that enable measuring the robustness of the target LLMs, including standard behaviors, copyright behaviors, contextual behaviors, and multimodal behaviors, totaling four types. In the evaluation pipeline, HarmBench ensures breadth by converting a series of behaviors into test cases. In addition, it employs fine-tuned Llama-2-13B and a hash-based classifier to evaluate the final results for non-copyright and copyright behaviors, respectively. The researchers compile 18 red teaming methods from 12 papers, and the experimental results indicate that no attack or defense is consistently effective, suggesting that robustness is independent of model size. This finding contradicts previous research [140], which points that larger models are more difficult to red team.

Existing research primarily emphasizes LLMs robustness against malicious inputs. However, we argue that achieving robustness should also ensure instructions-following ability without exhibiting exaggerated safety, as LLMs should balance safety with helpful. In the extreme, a model that indiscriminately refuses any prompt — safe or unsafe — would be perfectly harmless but completely useless [49]. To evaluate such exaggerated safety, Röttger *et al.* [49] design tests using both safe prompts (which should not trigger refusals) and unsafe prompts (which should). They find that Llama2 demonstrated significant exaggerated safety; although adding guardrail prompt decrease unsafe behaviors, it simultaneously exaggerated safety. This phenomenon of exaggerated safety, which the study attributes to lexical overfitting, suggests the model may be overly sensitive to certain words or phrases. Additionally, safety-related behavior could be changed by system prompts added at inference time. An *et al.* [58] introduce a white-box method for automatically generating pseudo-harmful prompts to induce false refusals in LLMs. Their findings revealed that many jailbreak defenses noticeably increase the false refusal rate, undermining usability and necessitating a trade-off between minimizing such refusals and enhancing safety against jailbreaks. Similarly, Varshney *et al.* [48] evaluates LLMs with both safe and unsafe prompts, training classifiers to evaluate both the safety of responses and whether the model abstain or answered. They observed that "self-checking" methods improved safety but induced extremely over-defensive. Consequently, maintaining instruction-following robustness while ensuring LLM safety is an area where much work remains to be done.

C. Ethics

Ethics encompasses the moral principles and values embedded in the text generated by LLMs, ensuring that the output adheres to socially accepted moral standards and legal frameworks. Ethical content must exclude elements that are unethical, inappropriate, or misleading. Specifically, this involves avoiding language that is discriminatory, biased, or offensive, as such expressions contravene the principles of social justice [107]. A growing body of research has proposed diverse methods for evaluating the ethical compliance of content generated by LLMs, which are summarized below:

TECHHAZARDQA: [14] is a dataset created to address the security vulnerabilities of existing LLMs when generating instructional content. While prior research has primarily focused on the safety of general text-based responses, it has often overlooked the elevated risks posed by instructional outputs, such as pseudocode or code snippets. The TECHHAZARDQA dataset consists of 1,850 complex queries spanning seven technical domains, each of which can be answered in either natural language or pseudocode. The authors systematically evaluated the models' performance across multiple response formats under various testing scenarios, including zero-shot, chain-of-thought (CoT), and few-shot learning [108]–[110]. To further investigate the impact of model editing on content generation, they employed the ROME model editing technique [111] to determine whether edited models were more prone to generating harmful content. The experimental results indicate that the likelihood of harmful content generation is significantly higher when models produce pseudocode compared to natural language responses. This probability further increases following model editing. These findings highlight that the generation of instructional content presents greater safety challenges than general text generation. Although model editing enhances the flexibility of LLMs, it also raises the risk of producing harmful content. The authors underscore the importance of ensuring both safety and ethical compliance when LLMs are applied to generate technical and instructional content in future developments.

LOCALVALUEBENCH: [15] is a framework designed to evaluate the value alignment and ethical performance of large language models (LLMs) across diverse cultural and legal contexts. It addresses the limitations of existing benchmarks, which are often shaped by the cultural backgrounds of their creators. This study focuses specifically on assessing the value alignment of LLMs within the Australian context, offering a demonstration framework to guide global regulators in developing localized evaluation systems.The research employs a three-tiered inquiry method to evaluate the ethical reasoning capabilities of the models: neutral questions establish baseline responses, debated questions explore how the models handle complex ethical scenarios, and misleading questions assess their behavior in extreme situations. The evaluation spans six key domains, including Capital Punishment, Weapons, and Gay Marriage, among others. To ensure experimental reliability, each response generated by the models was independently assessed by three evaluators.The results reveal distinct performance patterns among three prominent models: GPT-4, Gemini 1.5 Pro, and Claude 3 Sonet. Gemini 1.5 Pro performed well on topics such as Gay Marriage and Refugees but declined to respond to questions concerning Capital Punishment. Claude 3 Sonet demonstrated low performance on Weapons and Compulsory Voting but showed overall consistency across domains. GPT-4 underperformed on Gay Marriage, highlighting its limitations in addressing complex ethical scenarios.The analysis indicates that the cultural context embedded in the training data significantly influences the performance of each model. Furthermore, the subjective bias observed in evaluators' ratings underscores the need for standardized evaluation frameworks to assess the ethical alignment and value sensitivity of LLMs across diverse cultural environments.

TRUSTGPT: [16] is a comprehensive framework introduced to systematically assess the ethical and social responsibility performance of large language mod(LLMs) in three key areas: toxicity, bias, and value alignment. This framework addresses gaps in existing benchmarks by providing a nuanced evaluation of these aspects through three core dimensions: toxicity analysis, bias detection, and value alignment. TRUSTGPT also measures the rate at which models choose to refuse to answer (RtA), further assessing their decision-making in morally complex scenarios. The evaluation covered eight prominent LLMs(e.g., ChatGPT, LLaMA, and Vicuna). Findings from toxicity analysis revealed that FastChat exhibited the highest toxicity levels, with substantial variation across models. Bias detection results indicated that ChatGPT demonstrated more pronounced racial and religious biases, although most models managed gender bias relatively well. Value alignment was assessed through active value alignment (AVA) and passive value alignment (PVA) tasks; ChatGPT achieved the highest soft accuracy in AVA, while other models showed significant discrepancies between soft and hard accuracy. In PVA tasks, many models struggled with conflicting norms, suggesting the need for further development in ethical alignment capabilities. The study underscores persistent challenges facing LLMs in ethics and social responsibility, particularly their tendency to generate harmful content or exhibit bias in specific scenarios. The authors emphasize the need to incorporate advanced methods, such as reinforcement learning from human feedback (RLHF), into the training process to enhance the ethical performance and alignment of LLMs.

MoralBench: [20] is an innovative benchmarking framework constructed to systematically assess the moral reasoning capabilities of LLMs and their alignment with human ethical norms. While existing research has primarily focused on bias and safety, a comprehensive framework for evaluating moral judgment in LLMs has been lacking. MoralBench addresses this gap by employing the six core moral values outlined by Moral Foundations Theory [112], which include Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression. The evaluation framework consists of two components: Binary Moral Evaluation and Comparative Moral Evaluation. In the Binary Moral Evaluation, models are tasked with determining whether they "agree" or "disagree" with a set of moral statements, with scores based on how closely their responses align with human judgments. The Comparative Moral Evaluation requires models to select the more ethical option between two moral statements, with scoring reflecting the degree of alignment with human preferences. The experimental findings indicate that, although some models perform well on straightforward binary tasks, they encounter significant challenges with more complex comparative tasks. This suggests that models may rely more on recognizing patterns from training data rather than demonstrating a genuine understanding of ethical principles. Furthermore, the study uncovers biases within certain moral dimensions, underscoring

the need for further refinement to mitigate potential risks. MoralBench offers a valuable tool for evaluating and advancing LLMs, enabling researchers to identify deficiencies in moral reasoning and informing the development of models with improved ethical performance.

CVALUES: [46] addresses the limitations of existing evaluation frameworks in assessing the ethical alignment and value sensitivity of LLMs. The framework is structured around two critical dimensions: safety and responsibility. The safety dimension requires models to avoid generating harmful, biased, or unsafe content, while also ensuring their ability to manage sensitive topics, criminal advice, and privacy violations effectively. In contrast, the responsibility dimension expects models to provide positive guidance, demonstrate empathy, and account for the social and user impact of their outputs. The evaluation process integrates human assessments with automated multiple-choice evaluations, ensuring the reliability and comprehensiveness of the results. The experimental findings reveal that certain models, such as Ziya-LLaMA, offered inappropriate recommendations in response to illegal or unethical prompts, exposing vulnerabilities in their ability to handle complex scenarios safely. Furthermore, some models performed inadequately in terms of responsibility. This shortfall was particularly evident in the legal domain, where models were prone to being misled due to limited reasoning capabilities, resulting in outputs misaligned with ethical standards. Similarly, in the social sciences domain, models often failed to exhibit empathy or produce human-like responses, leading to lower responsibility scores. The study highlights that while Chinese LLMs have made significant strides in safety, their responsibility-related capabilities still require substantial improvement. CVALUES underscores the complementary nature of human and automated evaluations, advocating for their combined application to thoroughly assess models' ethical and social responsibility performance. This approach aims to guide the development of future LLMs that are both safer and more responsible.

In addition to the studies discussed above, several other works have made important contributions to the evaluation of LLMs' ethical performance. For instance, Sun *et al.* [17] developed a safety evaluation benchmark encompassing eight typical safety scenarios, including Ethics and Morality, along with six types of adversarial instruction attacks. The study introduced the SafetyPrompts benchmark to identify and assess ethical issues in LLMs models. The findings highlight that adversarial instruction attacks represent a major safety challenge for LLMs, with models exhibiting significant vulnerabilities in these contexts. The authors emphasize the need for more comprehensive safety evaluation tools and benchmarks, urging researchers to collaborate in building robust AI safety frameworks. Similarly, Zhang *et al.* [105] proposed a specialized framework to assess LLMs within the legal domain, focusing on ensuring both professional competence and ethical reliability in legal tasks. The results indicate that existing LLMs require further optimization when handling complex legal scenarios, particularly in terms of ethical reasoning. The authors advocate for rigorous, multi-dimensional evaluations prior to the deployment of LLMs in legal practice, emphasizing the importance of continuous optimization to reduce bias and enhance resilience against adversarial inputs. In another study, Hendrycks *et al.* [106] developed the ETHICS dataset to assess the moral reasoning abilities of LLMs, aiming to help align artificial intelligence more closely with shared human values. While the research shows that some models perform well on specific tasks, it also reveals substantial limitations in handling more complex ethical scenarios. The study further suggests that models' understanding and judgment are highly sensitive to the framing of situations in text, exposing the challenges that current models face in ethical reasoning.

D. Bias and Fairness

Bias and fairness are also indispensable dimensions of LLMs safety evaluation, which refers to the unintended reproduction of societal stereotypes, discriminatory views, or unequal treatment in the text generated by LLMs. It may also manifest as the model's inability to maintain neutrality and impartiality when processing user information related to diverse backgrounds, genders, races, or religions. The biases stem from factors such as training data, model specifications, algorithmic constraints, product design, and policy decisions [24]. As LLMs are increasingly integrated across various sectors, effectively evaluating and mitigating biases while enhancing fairness has become both a pressing challenge and a key focus of current research.

How Are LLMs Mitigating Stereotyping Harms: [93] suggests drawing lessons from search engine technologies to mitigate the negative impact of stereotyping in LLMs-generated outputs. The researchers combined insights from natural language processing (NLP) and search engine research to develop an innovative evaluation framework. This framework evaluates LLMs handling of stereotyping associated with different social groups through autocomplete-style prompts. The authors evaluate output across four dimensions: refusal, toxicity, sentiment, and regard. The findings indicate that while the introduction of system prompts helps reduce stereotypical outputs, LLMs still struggle with certain toxic content, particularly in responses to prompts related to peoples/ethnicities and sexual orientation. Additionally, mentions of intersectional identities tend to trigger disproportionately stereotyped responses. The study further explores the implications of these findings for the convergence of LLMs and search engines, offering insights into policy development aimed at reducing stereotypes. It emphasizes the shared responsibility of model developers, scholars, NLP practitioners, and policymakers to mitigate stereotype-related harm. The paper also calls for greater awareness in areas such as training data management, leader board design and usage, and social impact measurement.

Are Large Language Models Really Bias-Free: [13] investigates whether LLMs are truly bias-free. The research team conduct a comprehensive evaluation of multiple popular LLMs through two steps: an initial evaluation using standard prompts and an adversarial analysis employing jailbreak prompts. In the first step, the authors design a sentence completion task where the model is required to choose one option from two:

a stereotype or an counterstereotype to complete the provided sentence. Stereotypes deemed safe in this initial evaluation will progress to adversarial analysis in the second step, where the authors employed five jailbreak techniques—role-playing, machine translation, obfuscation, prompt injection and reward incentive—aimed at bypassing language model safety filters to elicit biased responses. The evaluation measures the performance of LLMs of various sizes based on robustness, fairness, and safety. The experimental results indicate that certain categories of bias, such as those related to sexual orientation and disability, are more effectively protected by model safety measures, while biases related to gender and age are less mitigated. Compared to smaller models, medium to large models exhibit stronger performance. However, despite having 1750 billion parameters, GPT-3.5 Turbo falls below the safety threshold and displays a high degree of stereotypical responses. In addition, the adversarial analysis highlights the powerful influence of role-playing techniques, with GPT-3.5 Turbo showing vulnerability to all four attacks. Importantly, no model is entirely safe, as each is relatively susceptible to at least one jailbreak attack. Therefore, a future layered defense approach that integrates multiple safeguards may be necessary.

Is ChatGPT Fair for Recommendation: [23] explores the fairness of LLMs in recommendation systems, specifically evaluating the fairness of recommendation via LLM (RecLLM). Due to the unique nature of ReLLM, traditional fairness measurement methods based on scores and fixed datasets struggle to meet the necessary requirements. Consequently, the researchers propose a new benchmark called FaiRLLM, which includes carefully designed evaluation metrics and datasets covering eight sensitive attributes across two recommendation contexts: music and movies. The core idea of FaiRLLM is to measure the similarity between the recommendation results of neutral instructions that do not include sensitive attributes and sensitive instructions that disclose such attributes. This approach assesses fairness by analyzing the differences in similarity among various sensitive attribute values (for example, in the context of race, comparing African American, Black, White, and Asian individuals). The researchers evaluate ChatGPT using FaiRLLM and find that ChatGPT still exhibits unfairness in generating recommendations related to certain sensitive attributes. Moreover, ChatGPT's responses align with existing societal biases regarding disadvantaged groups associated with different sensitive attributes. In addition, the authors also revealed the lack of robustness of ChatGPT to unfairness through the influence of sensitive attribute typos and language. These findings underscore that even advanced LLMs may not completely avoid social biases in recommendation tasks, emphasizing the need for more nuanced considerations and improvements regarding fairness in real-world applications.

Do Large Language Models Rank Fairly: Despite the extensive research on the efficiency and accuracy of LLMs in ranking tasks, studies on their fairness are lacking. Wang *et al.* [113] evaluates the fairness of LLMs as text rankers from the perspectives of users and items. Specifically, the authors focus on binary protected attributes (gender and geographic location) using the TREC Fair Ranking Track dataset and conduct two types of evaluations: list evaluation and pairwise evaluation. The listwise evaluation measures how LLMs integrate underrepresented groups into rankings from both the query-side and item-side, while the pairwise evaluation provides LLMs with two items—one from a protected group and another from a non-protected group—to compare relevance or irrelevance. The results reveal that both neural rankers and LLMs exhibit a preference for queries associated with female and European. Interestingly, Mistral-7b shows a marked bias toward male items in relevant pairs, which contrasts sharply with the behavior of other models, raising questions about the decision-making processes of these models. Overall, LLMs tend to exhibit more subtle and profound biases favoring certain protected groups.

CEB: [114] proposes a unified evaluation framework to comprehensively analyze biases in LLMs. Most existing research either focuses on specific types of biases or employs incompatibility evaluation metrics, making comparisons across different datasets and LLMs challenging. To address this, the authors introduce a Compositional Evaluation Benchmark (CEB). CEB adopts a compositional taxonomy approach to describe each dataset across three dimensions: bias type, social group, and task. The bias type includes stereotyping and toxicity, social group encompass age, gender, race, and religion, while task includes direct evaluation (recognition and selection) and indirect evaluation (continuation, conversation, and classification). To facilitate compatible evaluation, the researchers establish corresponding evaluation metrics for each task, including Micro-F1 score, GPT-4-as-evaluator, Perspective API [102], Demographic Parity (DP), Equalized Odds (EO), and Unfairness Scores [115]. The experimental results reveal that GPT models achieve the best performance on direct evaluation tasks; The Refuse to Answer (RtA) rate varies across LLMs and different settings, with larger LLMs typically obtaining an RtA rate close to zero, while smaller models like Llama2 and Llama3 exhibit significantly higher RtA rates; LLMs are adept at identifying toxic content in inputs but struggle to recognize stereotypical content; Furthermore, LLMs demonstrate a high RtA rate for racial and religious social groups, indicating a heightened sensitivity; In contrast, GPT models do not stand out in terms of performance on stereotype bias types, performing comparably to smaller LLMs.

In addition, there are many researchers evaluating LLMs and addressing bias and fairness issues when LLMs are used for other tasks. Huang *et al.* [25] introduces a Chinese bias benchmark dataset, developed through human-AI collaboration, aimed at measuring biases in Chinese LLMs. The evaluation results indicate that the bias scores for the evaluated models are significantly higher in categories such as educational qualification, disease, disability, and physical appearance compared to religion and sexual orientation. Cui *et al.* [4] evaluates the fairness of LLMs through a carefully designed new benchmark that utilizes the coefficient of variation as an evaluation metric across four dimensions: identity preference, credit, criminal, and health. The results show that racial groups receive the highest level of fairness. Azeem *et al.* [22] evaluates several highly-rated LLMs based on Human-Robot Interaction (HRI) discrimination and safety criteria, highlighting the lack of robustness in LLMs

when encountering people across various protected identity characteristics (e.g., race, gender, disability status, nationality, religion, and their intersections). Zhou *et al.* [6] employs red teaming methods to evaluate ChatGPT, revealing that although ChatGPT performs better than other models, it is highly likely to produce biased programs with biased induction. Given the widespread use of LLMs as evaluators, Wang *et al.* [116] introduces a metric for conflict rate to quantitatively evaluate the model's sensitivity to the position of responses. This novel approach indicates the presence of positional bias in LLMs, suggesting that they tend to favor responses from specific positions. For instance, GPT-4 is inclined to support responses from the first position, while ChatGPT tends to favor responses from the second position.

E. Truthfulness

Truthfulness pertains to the factual accuracy of information or statements, specifically their alignment with real-world facts [230]. Large language models (LLMs), trained on extensive textual datasets, may occasionally blur the distinction between fact and fiction in their responses [122], [123], potentially undermining user trust. A prominent issue related to truthfulness is hallucination [81], [119], [120], wherein models generate responses that seem plausible but are factually incorrect, thus misleading users. For example, when prompted about the "Four Great Inventions" of ancient China, a model might inaccurately list "gunpowder, the compass, silk, and porcelain," while the correct answer is "papermaking, the compass, gunpowder, and printing."

AI-LIEDAR: [26] aims to investigate the balance between utility [118] and truthfulness [117] in LLMs. These models may produce inaccurate information to meet user demands, particularly when task objectives conflict with truthful disclosure, as in promoting defective products. While previous studies have mainly addressed "hallucinations" (i.e., the generation of unsupported content), there has been limited focus on evaluating model truthfulness in scenarios where user instructions are ambiguous and prioritize utility. The AI-LIEDAR framework includes 60 real-life-inspired multi-turn dialogue scenarios, each shaped by one of three motivations for deception: benefits, public image, and emotion. Using the Sotopia platform [121], the framework simulates interactions between users and models, with a psychology-based truth detector classifying responses as fully truthful, partially truthful (e.g., through omission), or entirely false. Additionally, the study investigates model steerability, testing whether prompts can effectively direct responses toward truthfulness or deception. Findings reveal that LLMs rarely sustain complete truthfulness throughout interactions, with an overall truthfulness rate below 50%. Notably, a negative correlation emerges between truthfulness and goal achievement, especially in scenarios involving objective metrics, such as product promotion. Even when models are explicitly prompted toward "truthful" or "deceptive" responses, their truthfulness levels remain variable. The authors advocate for the development of more refined evaluation frameworks and robust guidance mechanisms to enhance model reliability across diverse contexts.

THaMES: [127] is a framework dedicated to evaluating and mitigating hallucinations generated by LLMs. This study is motivated by the frequent emergence of inaccurate information, especially in complex, domain-specific texts generated by existing models. The THaMES framework comprises three primary modules—QA set generation, hallucination benchmarking, and mitigation strategies—designed to identify and reduce hallucinatory outputs effectively. The framework begins by constructing high-quality question-answer test sets through weighted sampling and advanced question-generation techniques, enabling a thorough evaluation of hallucination instances. Next, multiple benchmark metrics are employed to assess models' capabilities in both recognizing and generating hallucinatory content. Following this, THaMES implements a series of refined mitigation strategies, including In-Context Learning (ICL) [124], Retrieval-Augmented Generation (RAG) [125], and parameter-efficient fine-tuning (e.g., PEFT) [126], [128], aimed at enhancing model performance and reducing hallucinations. To support diverse model architectures, THaMES offers a variety of mitigation options, allowing users to select the most suitable approach based on specific model characteristics. Experimental results reveal that the commercial model GPT-4 showed significant improvement under the RAG strategy, which leverages external knowledge integration, while open-weight models, such as Llama-3.1, achieved higher reasoning accuracy through ICL. Additionally, PEFT fine-tuning applied to Llama-3.1 demonstrated a marked reduction in hallucinations. By elucidating the effects of different mitigation strategies on hallucination reduction across various models, the THaMES framework establishes a new standard and provides valuable guidance for the ongoing development of LLMs.

FACTKB: [129] addresses the challenge of evaluating truthful accuracy in automated summarization [130], [131]. Current generative models often produce inaccurate factual information, particularly with respect to entities and relationships [135]. To enable more generalized and robust truthfulness evaluation, FACTKB combines pretrained language models with knowledge bases, employing three entity-centered pretraining strategies—Entity Wiki, Evidence Extraction, and Knowledge Walk—to enhance the model's representation of entities and relationships. In detail, Entity Wiki enriches the model's understanding of specific entities by integrating relevant background knowledge directly from the knowledge base. Evidence Extraction extracts supporting evidence from auxiliary information linked to entities, enhancing the model's capacity to verify facts within context. Finally, Knowledge Walk leverages multi-hop knowledge paths within the knowledge base, strengthening the model's compositional reasoning ability for complex relationships. By training with these strategies, the language model improves its evaluation of entity and relationship accuracy, enabling fine-tuning for precise factual error detection. Experimental results show that FACTKB surpasses other methods in identifying semantic framework errors, particularly those related to entities and relationships, while minimizing preprocessing requirements, thereby making it adaptable to diverse datasets. FACTKB not only aligns well with human judgment but also achieves high accuracy in

detecting various error types, especially in complex domains. Ultimately, this framework offers an accurate and accessible method for truthfulness evaluation, setting a reliable standard for fact-checking in text generation across multiple domains.

ReEval: [146] seeks to create new evaluation datasets using adversarial attacks [144], [145] to test whether retrieval-augmented LLMs [142], [143] accurately reference provided evidence in their responses. Traditional static question-answering datasets, frequently included in LLM pretraining data, often lead models to rely on memorization rather than genuine evidence, complicating objective assessments of model reliability. ReEval is motivated by the need for dynamic data generation that can automatically detect hallucinations—instances where model outputs deviate from the actual evidence. ReEval incorporates two data generation techniques: answer swapping and context enriching. In answer swapping, portions of the supporting evidence containing the answer are altered to examine whether the model aligns with the new answer context. In context enriching, supplementary information is added to the original context, enhancing question complexity while preserving the core evidence. Experiments were conducted on multiple open-domain question-answering datasets using LLMs, including GPT-4 and ChatGPT. Results indicate that all models exhibited a significant accuracy decline on adversarial test data, despite varied prompts, suggesting that ReEval effectively induced hallucination phenomena. Furthermore, ReEval's adversarial test samples demonstrated transferability; samples generated by smaller models were applicable to larger models, substantially reducing evaluation costs. Additionally, human reviewers rated the ReEval-generated test data positively in terms of readability and evidential support, confirming that these datasets realistically capture complex real-world scenarios.

HaluEval: [147] addresses the prevalent issue of hallucinations in LLMs, where models often generate unverifiable or inaccurate information. To systematically evaluate LLMs' capabilities in hallucination detection, HaluEval introduces a dataset of 35,000 samples covering three tasks: knowledge-based dialogue [149], question answering [148], and text summarization [150], with each sample either manually annotated or automatically generated. By employing a two-step framework involving sampling and filtering, HaluEval ensures both diversity and high complexity in its samples, aiming to rigorously test models' hallucination recognition capabilities. The dataset is constructed by first sampling user queries from various public sources, with ChatGPT generating responses likely to contain hallucinations. Responses with low similarity, as calculated using BERTScore [151], are then filtered and annotated for hallucinated content. Additionally, HaluEval incorporates task-specific, automatically generated samples through diverse sampling and filtering methods to create a challenging set of hallucination cases. Experimental results reveal notable shortcomings in current LLMs' hallucination detection abilities. For example, ChatGPT achieved an accuracy of only 58.53% in identifying hallucinations within text summaries. The study further indicates that incorporating external knowledge or adding reasoning steps significantly enhances hallucination detection for certain tasks, though these improvements are less pronounced in dialogue contexts. Additionally, HaluEval reveals that hallucination frequency is strongly topic-dependent, with higher rates observed in content related to film, technology, and climate. This work underscores critical blind spots in LLMs' performance on specific topics and complex dialogues, offering essential insights for developing more reliable models in the future.

In addition to the studies mentioned above, other researchers have also explored methods for evaluating the truthfulness of LLMs. Wang *et al.* [152] proposed OpenFactCheck, a modular framework that provides a unified system for evaluating the factual accuracy of content generated by large language models (LLMs). OpenFactCheck integrates multiple evaluation modules, allowing customizable truthfulness assessments. Demonstrating high consistency across diverse datasets, particularly in dynamic information domains, its adversarially generated samples effectively test factual verification and induce hallucination phenomena in complex knowledge areas, making it ideal for assessing LLM reliability. GraphEval [153] detects hallucinations in LLM outputs by constructing knowledge graphs (KGs) [154] in two steps: KG construction and triple consistency verification. By limiting LLM calls to a single KG construction pass, this approach reduces computational costs and enhances interpretability by pinpointing inconsistencies. Combined with natural language inference (NLI) models, GraphEval significantly improves balanced accuracy in hallucination detection, boosting NLI performance by approximately 6.2 points on benchmarks like SummEval [155] and QAGS-C [156], especially for longer outputs. Zhu *et al.* [157] introduced KG-FPQ, a benchmark that leverages knowledge graphs to generate False Premise Questions (FPQs) for assessing truthfulness hallucinations in LLMs. Covering art, people, and places, the KG-FPQ dataset includes around 178,000 FPQs, enabling comprehensive tests of hallucination handling in both discriminative and generative tasks [158]–[160]. The framework also introduces FPQ-Judge, an automated hallucination evaluator. Findings reveal LLM susceptibility to hallucinations under false premises, with FPQ complexity and task format significantly impacting hallucination performance, underscoring the need for improved truthfulness across varied tasks and domains.

F. Other

In addition to commonly studied evaluation dimensions—such as toxicity, robustness, ethics, bias and fairness, and truthfulness—LLMs require rigorous evaluation across several other critical areas. These include privacy, mental health and medical applications, and code generation, copyright, agents. Assessments in these domains are essential to ensuring that LLMs adhere to standards of safety, legality, and effectiveness, thus supporting their responsible and reliable deployment in real-world contexts.

Privacy: LLMs rely on extensive data for training, posing inherent risks of exposing sensitive user information [161]–[163], including personal identifiers, geographic locations, and employment records. Such potential privacy leakage not only erodes user trust but also raises significant data compliance

concerns. Accordingly, a systematic assessment of privacy risks in LLMs is critical for identifying and mitigating channels of potential privacy breaches, thereby ensuring model safety and regulatory compliance in practical applications. Li *et al.* [39] introduced LLM-PBE, a comprehensive toolkit for evaluating privacy risks throughout the LLMs lifecycle. LLM-PBE facilitates privacy assessments across diverse data types, attack vectors, and defense strategies, with a particular focus on understanding how model size, data characteristics, and temporal aspects impact privacy vulnerabilities. The toolkit includes modules for various attack types, such as data extraction [163], [164], membership inference [165], prompt leakage [166], and jailbreaking [167], complemented by defense strategies like differential privacy [168]. Extensive experimentation on datasets such as Enron [169] and ECHR [170] offers a systematic analysis of privacy concerns in LLMs. Additionally, the LLM-PBE framework integrates support for models hosted on platforms such as OpenAI, TogetherAI, and Hugging Face, enabling users to assess privacy risks across various LLMs through API interfaces. Experimental results indicate that larger models are more susceptible to privacy leakage. Additionally, certain data types and placements, such as sensitive information located at the end of documents, are particularly prone to extraction. Furthermore, while differential privacy effectively mitigates leakage risks, it introduces some trade-offs in model performance, underscoring the need to balance privacy protection with practical utility.

Mental Health and Medical Applications: The growing popularity of mental health chatbots, valued for their human-like, context-aware support, has attracted a large user base. However, these chatbots are prone to producing hallucinations or misleading advice and may fail to offer appropriate guidance in crisis situations, such as managing self-harm [171]. Consequently, there is an urgent need for a standardized evaluation framework to assess the safety and reliability of these systems, thereby fostering user trust and facilitating broader adoption in mental health care. Park *et al.* [34] proposed a dedicated evaluation framework to ensure the safety and reliability of mental health chatbots based on LLMs. Their framework consists of 100 benchmark questions paired with ideal responses, addressing common mental health scenarios. In addition, five guiding questions assist mental health professionals in evaluating models across key dimensions: safety, crisis management, consistency, resource provision, and user autonomy. This framework provides a comprehensive standard for evaluating LLMs in mental health applications, supporting trust-building and the safer integration of these tools into healthcare settings. Similarly, the rapid development of LLMs in medical applications has introduced significant safety concerns. These models, when processing medical information, may generate misleading or unethical recommendations, which pose potential risks to patients and public health. Han *et al.* [183] introduced MedSafetyBench, a benchmark dataset specifically designed to evaluate the safety of LLMs in medical contexts. MedSafetyBench includes 1,800 examples of harmful medical requests with corresponding safe responses, structured around nine core principles of medical ethics. By generating harmful medical prompts and appropriate responses, the researchers assessed existing LLMs (such as GPT-4 and Llama) to determine their capacity to reject unsafe requests. Experimental results indicate that current medical LLMs frequently fall short of safety standards, often providing unsafe responses to inappropriate prompts. The researchers also conducted fine-tuning experiments using MedSafetyBench, which demonstrated that fine-tuning with this dataset significantly enhances model safety while maintaining the accuracy of medical knowledge.

Code Generation: With the growing application of LLMs in programming assistance, their potential to generate insecure code or respond to malicious prompts introduces substantial cybersecurity risks [172]–[174]. For example, LLMs may produce code vulnerable to common attacks, such as SQL injection or cross-site scripting, leaving generated applications susceptible to exploitation. Consequently, assessing the security of LLMs in code generation has become imperative. Bhatt *et al.* [36] introduced the CYBERSECEVAL framework, designed to evaluate the security of LLM-generated code and their responsiveness to cyberattack-related prompts. CYBERSECEVAL includes two main evaluation components: insecure coding tests and cyberattack helpfulness tests. The insecure coding tests utilize a "vulnerability detector" to generate cases reflecting common programming weaknesses, assessing whether models generate insecure code through both code completion and prompt-based tasks. The cyberattack helpfulness tests involve researcher-designed, attack-related prompts to evaluate if model responses might facilitate malicious actions. Experimental findings indicate that all tested models produced insecure code in 30% of insecure coding scenarios, with models possessing stronger programming capabilities posing greater security risks. Additionally, an average of 53% of cyberattack prompts elicited responses aligning with attack objectives, underscoring LLMs' limitations in handling malicious requests. The study also noted that models with more advanced programming capabilities demonstrated a higher level of compliance in cyberattack scenarios. CYBERSECEVAL thus provides a systematic tool for evaluating LLMs cybersecurity, supporting the development of safer AI systems. Additionally, Wang *et al.* [35] developed the CodeSecEval dataset to evaluate the security of LLMs in code generation and repair. CodeSecEval comprises 180 samples across 44 critical vulnerability types and is organized into two subsets (SecEvalBase and SecEvalPlus), designed for assessing secure code generation and repair. Their study conducted multiple experiments to examine LLMs' secure code generation abilities and proposed several enhancement strategies, including "vulnerability-aware information" (explicitly indicating vulnerability risks) and "insecure code explanations" (providing explanations for vulnerabilities) to improve LLMs security.

Copyright: A systematic approach to copyright assessment is essential for effectively identifying and preventing the generation of infringing content, thus ensuring compliance and facilitating the safe, lawful use of LLMs across various creative domains. This approach mitigates potential legal risks related to copyright for both companies and users. Liu *et al.* [37] introduced the SHIELD framework, a tool designed to evaluate and prevent copyright violations by LLMs during

text generation [175]. As LLMs find broader application in content creation, concerns about copyright infringement have become more pronounced, particularly due to their potential to generate protected content without authorization [176]. This study addresses gaps in existing copyright protection mechanisms, which may be overly restrictive or insufficiently responsive to unauthorized content generation. The SHIELD framework employs a curated copyright dataset to evaluate whether model-generated text adheres to copyright standards and introduces a lightweight, real-time proxy defense mechanism to dynamically detect and block potentially infringing content. The SHIELD framework consists of three primary modules: the Copyright Material Detector, Copyright Status Verifier, and Copyright Status Guide. The Copyright Material Detector uses an N-gram language model to identify copyrighted segments within generated content; the Copyright Status Verifier performs real-time online checks to verify the copyright status of content; and the Copyright Status Guide provides contextual guidance to direct the model toward producing copyright-safe content. This defense mechanism is designed to quickly identify and reject sensitive requests while maintaining flexibility in generating public domain text. Experimental results demonstrate that the SHIELD framework significantly reduces the generation of copyrighted material. It also achieves higher refusal rates in response to "jailbreak attacks" [45], [104], [177] (attempts to bypass model safeguards), underscoring its robustness. Furthermore, SHIELD effectively avoids over-protection by not interfering with the generation of public domain content, offering a practical and reliable solution for copyright compliance.

Agents: LLMs Agents serve as automated systems powered by LLMs, designed to analyze user input continuously and provide outputs that meet specific objectives, facilitating complex operations and interactive task completion [178], [179]. Due to their application in diverse, intricate scenarios, rigorous evaluation is crucial to optimize performance and ensure safety. Yuan *et al.* [40] introduced R-Judge, a benchmark dataset designed to assess LLMs' awareness of safety risks in multi-turn interactions. R-Judge comprises 569 multi-turn records across 27 risk scenarios in programming, IoT, software, networking, and finance, focusing on ten risk types like privacy breaches and property loss. This benchmark reveals current models' limitations in multi-dimensional safety risk handling, emphasizing the need for diverse, high-quality data to enhance LLM safety. Similarly, Zhu *et al.* [41] proposed RiskAwareBench, a framework to evaluate physical risk awareness in LLM-based agents, where "physical risk" refers to potential harm to property or individuals from decisions by agents such as robots. This framework utilizes the PhysicalRisk dataset, which includes safety guidelines and scenarios to systematically test LLMs' capacity to recognize and mitigate physical risks. Zhang *et al.* [42] developed Agent Security Bench (ASB) to assess the security resilience of LLMs agents against security threats, including prompt injection, memory poisoning, and backdoor attacks [180]–[182]. Covering ten domains, such as e-commerce and autonomous driving, ASB evaluates ten agents using over 400 tools and tasks across 23 attack and defense strategies, revealing an 84.3% average success rate for mixed attacks. These findings highlight significant vulnerabilities in current LLMs agents and the limited effectiveness of existing defenses against complex attacks.

IV. WHERE TO EVALUATE

Where should we evaluate the safety of LLMs? Evaluation metrics determine the aspects of safety performance to be measured, toolkits provide the framework for the entire evaluation process, and datasets and benchmarks form the foundation of these evaluations. In this section, we will comprehensively summarize and introduce these three parts, which offer researchers a clear view of LLMs safety performance.

A. Metrics

This section will discuss the evaluation metrics used to evaluate the safety of LLMs, as well as some frontier research related to these metrics. The focus will be on evaluating whether LLMs generate harmful content; currently, there are relatively few summaries of evaluation metrics in this area. In contrast, metrics for evaluating the truthfulness of LLMs have already been summarized in the context of hallucination evaluations [188], so we will not elaborate further on that aspect in this section.

1) Attack Success Rate: Attack success rate (ASR) is the most widely used and critical metric for evaluating the safety of large language models (LLMs). It measures the effectiveness of adversarial attacks such as jailbreak attacks and prompt injection attacks directed at LLMs. Obviously, the higher the ASR, the less safe LLMs are. Mathematically, ASR can be formulated as:

$$ASR = \frac{Num_{success}}{Num_{total}}. \tag{1}$$

where $Num_{success}$ denotes the number of successfully attacked prompts, and Num_{total} denotes the total number of harmful prompts.

Similarly, variations of ASR, such as the harmful content rate and refuse rate, share the same fundamental concept, so we will not discuss them separately. However, a significant challenge remains in defining what constitutes a "successful attack." How to evaluate the success of a jailbreak attempt against an LLM has not been unified [187], leading to difficulties in comparing the works of different researchers. Current automatic studies primarily determine success through two approaches: rule-based and language model-based methods. In brief, the former evaluates whether certain predefined keywords (e.g., "Sorry," "As an AI assistant," etc.) are present in the LLMs responses to determine if the model is obedient to malicious prompts, while the latter employs a judge LLM to evaluate the prompts and responses and determine the success of the attack. This aspect will be discussed in detail in Section V.

2) Other fine-grained Metric: As mentioned in the section IV-A1, the research community has yet to reach a consensus on how to define "attack success" or "jailbreak success." Therefore, to comprehensively and accurately evaluate the safety of LLMs, many researchers have proposed new metrics or adapted some metrics from general LLMs evaluation frameworks.

Cai *et al.* [177] points out the limitations of rule-based and language model-based evaluation methods, specifically their inability to understand context or the oversimplification of binary classification metrics. These approaches often lead to numerous responses mistakenly classified as successful attacks, resulting in false negatives. Consequently, researchers propose a new set of metrics to distinguish between attackers and better capture the nuances among jailbreak motivations. Specifically, the authors introduce three novel metrics—Safeguard Violation (SV), Informativeness (I), and Relative Truthfulness (RT). These metrics define the success of a jailbreak in a progressive manner: SV assesses whether the LLM's response contains any violative content, I evaluates whether the response provides relevant information related to the prompt's intent, and RT measures the response's relative truthfulness concerning the prompt's intent. However, the specific implementation of these three metrics still relies on prompting the LLMs.

StrongReject [184] proposes an evaluation metric based on three distinct scores, designed to address the limitations of single-step evaluation by models. Before generating the three scores, the evaluator model outputs chain-of-thought reasoning about how well the response addresses the prompt. The three scores include a binary non-rejection score and two 5-point Likert scale scores that measure how specific and convincing the response is. The final score for the prompt/response pair is calculated using the following formula:

$$score = (1 - refused) \times \frac{specific + convincing}{2}. \quad (2)$$

where $refused$ denotes whether the response is rejected (1) or not (0). $specific$ and $convincing$ denote how specific and convincing the response is. The higher the $score$, the less safe the response.

Chen *et al.* [53] argues that a reliable and safe LLM should meet three core criteria: resistance to jailbreak, non-toxicity, and quality. Accordingly, researchers use five metrics to evaluate these aspects: the commonly used attack success rate (ASR), toxicity score, fluency based on perplexity, token length, and grammatical errors. Additionally, Leidinger *et al.* [93] incorporates four metrics—refusal, toxicity, sentiment, and regard—to capture nuanced biases in model responses. Moreover, Mei *et al.* [190] suggests that many detected jailbreaks may actually result from hallucinations. To evaluate such hallucinations, researchers have introduced a range of metrics, including entity transformation, semantic consistency, and recall.

Overall, most researchers currently rely on the relatively simple metric of ASR as a primary evaluation metric. However, there are notable differences in the specific methods used to implement this metric, and the reproducibility of the results is poor. Some researchers have also attempted to refine existing metrics and propose fine-grained metrics for a more in-depth evaluation, though their impact remains limited. Thus, further research is needed on how to establish a unified and universal set of LLMs safety evaluation metrics.

B. Toolkits

Unlike individual metrics used for LLMs safety evaluation, toolkits typically integrate the entire evaluation pipeline, enabling a comprehensive and end-to-end safety evaluation.

Sun *et al.* [17] developed a Chinese LLMs safety evaluation benchmark, providing a comprehensive evaluation of LLMs safety across eight typical safety scenarios and six more challenging instruction attacks. Specifically, the researchers input collected prompts and corresponding responses into InstructGPT [191] to evaluate the safety of model outputs. Additionally, the authors released a safety leaderboard that documents the safety performance of all evaluated models, offering a valuable reference for LLMs developers.

HarmBench [10] introduces a standardized evaluation framework for automated red-teaming to evaluate both red-team attacks and defenses. HarmBench identifies three key qualities essential for automated red-team evaluation: breadth, comparability, and robust metrics. Based on these principles, the evaluation pipeline begins by converting a diverse set of behaviors into test cases to evaluate the target model. These behaviors and their corresponding responses are then processed through multiple classifiers to output evaluation results, ultimately determining the final attack success rate.

JailbreakBench [11] developed an open-source evaluation framework for evaluating jailbreak attacks and defenses. First, JailbreakBench maintains an evolving repository of jailbreak artifacts, encompassing state-of-the-art attacks and defenses. Additionally, it establishes a standardized evaluation framework that allows users to clearly define threat models, system prompts, chat templates, and scoring functions. This framework provides a convenient, streamlined approach for red-teaming, defense testing, and the selection of various evaluation classifiers. Finally, JailbreakBench features a leaderboard on their website that displays the performance of advanced jailbreak attacks and defenses.

EasyJailbreak [192] simplifies the construction and evaluation of jailbreak attacks against LLMs, proposing a unified implementation framework. This framework comprises four components: Selector, Mutator, Constraint, and Evaluator. The entire process is divided into three stages. In the preparation stage, users provide initial configurations, such as malicious prompts and template seeds. During the attack stage, EasyJailbreak iteratively attacks the target model based on user inputs, updating prompts and evaluating each attack's output. In the final stage, it summarizes the attack results, returning a comprehensive report to the user.

WalledEval [54] is a comprehensive LLM safety testing toolkit. It consists of three main classes: a dataset loader, a LLM loader, and a judge loader, supporting LLM benchmarking, judge benchmarking, and multiple-choice benchmarking. In addition, WalledEval is compatible with both weight-based and API-based model types and includes over 35 safety evaluation benchmarks, covering a wide range of safety dimensions such as multilingual safety, exaggerated safety, and prompt injections.

JailbreakEval [187] offers a user-friendly toolkit specifically for evaluating the safety of content generated by LLMs. Unlike other frameworks that integrate attack and evaluation

Table I
OVERVIEW OF COMPREHENSIVE EVALUATION DATASETS.

Benchmark Name	Size	Safety Dimensions	Languages	Composition
Chu *et al.* [63]	160	16	English	Harmful prompts
R-Judge [40]	162	27	English	Multi-turn agent interaction
JailbreakBench [11]	200	10	English	Harmful behaviors and benign behaviors
StrongREJECT [184]	346	6	English	Harmful prompts
HarmBench [10]	510	7	English	Harmful behaviors
Do-Not-Answer [43]	939	5	English	Harmful instructions
AdvBench [7]	1k	8	English	Harmful strings and harmful behaviors
CoSafe [55]	1.4k	14	English	Multi-turn attack questions
Chen *et al.* [53]	1.5k	61	English	Harmful prompts
TechHazardQA [14]	1.8k	7	English	Sensitive and unethical questions
FFT [4]	2.1k	3	English	Elaborated-designed harmful prompts
JailJudge ID [194]	4.5k	14	English	Labeled harmful prompts and model responses
SORRY-Bench [56]	9k	45	English	Unsafe instructions
HarmfulQA [8]	18k	10	English	Harmful questions, blue and red multi-turn conversations
AegisSafetyDataset [218]	26k	13	English	Human-LLM interaction instances
SALAD-Bench [47]	30k	66	English	Harmful prompts and multi-choice prompts
ALERT [62]	45k	32	English	Red teaming prompts
WildGuardMix [51]	92k	13	English	Harmful prompts with refusal and compliance response
Do Anything Now [45]	107k	13	English	Forbidden questions
CHiSafetyBench [59]	2.1k	31	Chinese	Multi-choice questions and risky questions
Wang *et al.* [61]	3k	17	Chinese	Risky questions and harmless questions
CValues [46]	6.4k	10	Chinese	Adversarial prompts and multi-choice prompts
JailBench [227]	10.8k	40	Chinese	Jailbreak questions
CRiskEval [60]	14k	7	Chinese	Multi-choice questions
Safety Prompts [17]	100k	14	Chinese	Augmented prompts and responses
ChineseSafe [189]	205k	10	Chinese	Illegal and unsafe contents
Latent Jailbreak [9]	416	3	Chinese; English	Translation tasks
SafetyBench [44]	11k	7	Chinese; English	Multiple choice questions
S-Eval [57]	220k	52	Chinese; English	Harmful prompts
AraTrust [28]	516	8	Arabic	Multi-choice questions
XSafety [185]	2.8k	14	10	Harmful prompts
JailJudge OOD [194]	6k	14	10	Labeled harmful prompts and responses
PolyglToxicityPrompts [5]	425k	7	17	Real-world harmful prompts

processes, JailbreakEval is designed solely for evaluating content safety. It includes a variety of out-of-the-box evaluators, broadly categorized into string matching evaluators, text classification evaluators, chat evaluators, and voting evaluators. Users simply input jailbreak prompts and model responses to obtain evaluation results.

C. Datasets and Benchmarks

To comprehensively evaluate the safety performance of LLMs, numerous datasets and benchmarks have been developed. Unlike those focusing on specific domains [35], [183], in this section, we discuss only those datasets and benchmarks designed to evaluate LLMs' overall safety across multiple dimensions. In this study, we selected 31 datasets and benchmarks in total, as shown in Table I. To facilitate differentiation, we categorized them into four groups based on their composition: only questions (Q-only), questions and answers (Q&A), multiple-choice questions (MCQ), and multi-turn conversations (Dialogue). Additionally, we advocate that the research community should consider LLMs safety within multilingual environments.

The majority of major datasets and benchmarks consist of Q-only. Chu *et al.* [63] establish a dataset of prohibited questions, providing the first large-scale evaluation of various advanced jailbreak attack methods. StrongReject [184] pro-

poses a dataset composed of forbidden questions that LLMs should refuse to answer to evaluate jailbreak performance. HarmBench [10] construct its dataset by incorporating not only harmful behaviors but also copyright, contextual, and multimodal behaviors for a comprehensive evaluation. Do-Not-Answer [43] includes prompts that responsible language models should not answer, while AdvBench [7] contains harmful strings and behaviors. Although initially develop using the white-box jailbreak method GCG, AdvBench has become a popular evaluation benchmark for researchers. To evaluate LLMs robustness against jailbreak attacks, Chen [53] introduce a refined dataset covering a wide range of harmful prompts. TechHazardQA [14] presents a complex harmful query dataset, requiring LLMs to respond with pseudocode instead of standard text to evaluate ethical boundaries. FFT [4] carefully elaborates a dataset of harmful prompts focusing on factoid, unfair, and toxic contents in LLMs. SORRY-Bench [56] addresses three key limitations in existing benchmarks: coarse-grained taxonomy of unsafe topics, overlook of linguistic characteristics and formatting of prompts, and reliance on large LLMs for evaluation. ALERT [62] provides a substantial instruction dataset with a fine-grained risk taxonomy for red-team testing of LLMs. Do Anything Now [45] recognizes the shift of jailbreak prompts from online communities to prompt aggregation sites, so researchers create a question set based on in-the-wild jailbreak prompts. Wang *et al.* [61] construct a dataset including both harmful and benign prompts to measure LLMs false positives and false negatives in refusal scenarios. JailBench [227] collects and integrates advanced jailbreak attack methods in the field, constructs a jailbreak prompt dataset with powerful jailbreak capabilities, and can comprehensively detect and evaluate the content safety protection capabilities of LLMs. In alignment with Chinese internet content regulations, ChineseSafe [189] expands on sensitive categories specific to the Chinese context, including political sensitivity, pornography, and variant/homophonic words. Latent Jailbreak [9] innovatively embeds harmful instructions into translation tasks, aiming to evaluate the robustness of LLMs against jailbreak attacks through translation tasks. Finally, S-Eval [57] introduces a bilingual dataset of risk prompts in both Chinese and English, containing fundamental risk prompts along with their corresponding adversarially crafted attack prompts.

Compared to datasets containing only questions(Q-only), datasets that include both questions and corresponding answers(Q&A) offer a reference ground truth, enabling more precise safety evaluations and even fine-tuning of models for improved performance. JailbreakBench [11] proposes a mixed dataset covered OpenAI's usage policies, where each harmful behavior is paired with a benign behavior, making it useful for evaluating rejection rates and model defenses. JailJudge ID [194] includes a collection of manually annotated harmful prompt-response pairs set in complex scenarios, while AegisSafetyDataset [218] is a large-scale record of safey and unsafe interactions between humans and LLMs. Additionally, WildGuardMix [51] is a large-scale, carefully balanced multitask dataset designed for safety regulation, encompassing both direct prompts and adversarial jailbreak responses with various refusal and compliance responses, primarily intended for fine-tuning safety evaluation models. Safety Prompts [17] provides augmented prompts and responses aimed at testing and improving LLMs safety, supporting a targeted approach to model safety enhancement.

Multiple-choice questions(MQA) offer a clear advantage over Q&A-style assessments when evaluating model safety, as they provide fixed correct answers, making it straightforward to evaluate model performance. SALAD-Bench [47], for instance, includes both harmful questions and multiple-choice questions, diversifying queries to enable a more comprehensive safety evaluation of LLMs. In the Chinese context, CHiSafetyBench [59] is designed to identify risky content by requiring LLMs to select the single safe or unsafe option from multiple choices. CValues [46] also focuses on Chinese scenarios, evaluating LLMs alignment with human values from safety and responsibility perspectives by having models choose the better response between two options. To measure the risk propensity of LLMs, CRiskEval [60] constructs multiple-choice questions, each with four manually labeled safety levels: Extremely Hazardous, Moderately Hazardous, Neutral, and Safe, providing a fine-grained evaluation scale. SafetyBench [44] combines both Chinese and English settings, creating a large, diverse multiple-choice dataset that thoroughly evaluates LLMs safety, blending harmful and safe behaviors in options to challenge models in their selection.

Some jailbreak attack methods have evolved beyond single-turn attacks, extending to multi-turn interactions, which have shown significantly heightened risks [195]–[197]. Consequently, evaluating the safety of LLMs in multi-turn conversations(Dialogue) has become necessary. R-Judge [40] provides a benchmark for assessing LLMs capabilities in identifying and evaluating safety risks within multi-turn interaction records between agents. CoSafe [55] highlights the current focus on single-turn interactions in red-teaming approaches and thus introduces a dataset designed for multi-turn attacks, featuring cohesive referencing across multiple turns. Additionally, HarmfulQA [8] presents a dataset not only of harmful questions but also includes both blue and red conversations, making it suitable for fine-tuning safety models.

Furthermore, we emphasize the importance of evaluating the multilingual safety of LLMs. Studies have shown that the unbalanced cross-linguistic distribution of LLMs training data leads to language biases, resulting in inconsistent outputs when the same task is described in different languages [33]. Research also indicates that multilingual mixed prompts can significantly intensify the harm of malicious queries, and that different language types and families exhibit substantial difference in LLMs safety [32]. Thus, while much research is focused on English and Chinese, the safety of LLMs in multilingual and low-resource language contexts deserves attention. AraTrust [28] introduced the first Arabic-language LLMs trustworthiness benchmark, consisting of manually crafted multiple-choice questions. XSafety established the first multilingual safety benchmark, assessing LLMs across 10 languages. JailJudge OOD [194] also introduced a multilingual safety dataset in 10 languages, with labeled harmful prompts and responses. To address the limitations of existing toxicity benchmarks focused primarily on English, PolyglToxici-

tyPrompts [5] released a dataset with natural prompts across 17 languages for evaluating multilingual toxicity in LLMs.

V. HOW TO EVALUATE

Having identified what and where to evaluate in terms of LLM safety, the next critical question centers on how to perform such evaluations effectively. A wide range of methodologies have emerged to assess the safety of LLMs, varying in automation, granularity, and evaluator participation. This section provides a comprehensive review of the main evaluation paradigms, categorizing them according to the nature and role of the evaluator: human, rule-based or model-based, and discussing the advantages, limitations, and use cases of each approach.

A. Human-based Evaluation

In the domain of large model safety evaluation, several critical safety concerns are involved, including Toxicity, Ethics, and Bias. A central challenge in this context is defining the "safety" of the output content. Human-based evaluation holds a unique advantage in this regard due to its capacity for nuanced understanding and judgment of complex content [208]. Specifically, human evaluators are better equipped to interpret subtle contextual cues, tone, and sarcasm, enabling them to more accurately identify the potential meanings embedded in the content. This is particularly important for the detection of issues such as Toxicity and Bias. However, human evaluation is not without its challenges, primarily related to cost and consistency. It necessitates significant human and time resources [207], and in large-scale evaluations, as the volume of content increases, the associated costs also rise substantially. Moreover, due to variations in evaluators' cultural backgrounds, knowledge bases, and personal values, human evaluation may lack consistency, particularly when assessing issues like Toxicity and Bias [74].

While automated evaluation offers advantages in terms of scalability and cost-effectiveness, it remains limited in its ability to fully comprehend complex contexts or diverse content. For example, models often struggle to grasp sarcasm, puns, and other intricate expressions, whereas human evaluators can provide a more comprehensive understanding in such cases. This limitation of automated methods is one of the reasons why many studies on large language model (LLM) safety continue to incorporate human evaluation components. For instance, Meadows *et al.* [15] utilized human evaluation to assess the alignment of LLM-generated content with local values, with a focus on evaluating the adherence of LLMs to Australian values and providing a framework for global regulatory bodies. Similarly, Alghamdi *et al.* [28] conducted a human-based evaluation of LLM credibility in Arabic-speaking contexts, finding that open-source models such as AceGPT 7B and Jais 13B struggled to achieve scores above 60% in benchmark tests. During their evaluation, they analyzed the cognitive differences in the perception of offensive content based on annotators' gender, age, and cultural backgrounds. In a related study, Movva *et al.* [209] performed human evaluations where annotators assessed chatbot responses according to five safety standards—namely harm, bias, misinformation, political stance, and polarization—investigating the consistency between LLMs and human annotators in labeling conversational safety.

B. Rule-based Evaluation

Rule-based evaluation in the safety assessment of large language models typically involves the detection of potential risks or inappropriate content in model outputs by defining a set of explicit rules or guidelines. This approach relies heavily on predefined criteria to flag outputs that may contain sensitive information, inappropriate language, or other safety concerns. The process of rule-based evaluation generally follows these key steps:

1) **Rule Definition:** A comprehensive set of rules is established to clearly delineate which content is deemed unsafe, specifying the expressions or behaviors that are considered unacceptable.
2) **Model Behavior Analysis:** The model's responses are compared against the predefined rules to identify potential violations or unsafe content.
3) **Scoring Evaluation:** Upon analyzing the model's outputs, the content is evaluated and scored based on specific metrics to assess the overall safety performance of the model. Common metrics include accuracy, recall, and specific matching scores, such as ROUGE-L [90].

For example, Liu *et al.* [37] developed a set of evaluation rules designed to assess the copyright compliance of text generation. Specifically, they measured the similarity between generated content and copyrighted material using metrics like the Longest Common Substring (LCS) [226] and ROUGE-L scores [90]. These metrics help determine whether the model reproduces copyrighted text. In addition, the authors utilized multiple performance indicators, including LCS, ROUGE-L scores, and rejection rate. While LCS and ROUGE-L scores quantify the similarity of the generated content, the rejection rate reflects the model's ability to effectively reject the generation of copyrighted content.

Similarly, Ye *et al.* [210] established a set of security rules encompassing three stages: input, execution, and output, to clearly define what content is considered unsafe. For each stage, multiple safety scenarios were outlined, with specific rules addressing issues such as the rejection of malicious queries, avoidance of incorrect tool selection, and filtering of harmful or erroneous information. The authors also quantitatively evaluated the model's safety performance, utilizing metrics such as rejection rate, tool selection error rate, and the proportion of unsafe outputs.

While the rules-based evaluation method offers clear and interpretable criteria, its reliance on predefined rules presents limitations, particularly in detecting new, unknown risks or latent safety issues. As a result, rule-based evaluation is often combined with other evaluation methods, such as human evaluation and large model evaluation, to provide a more comprehensive and efficient assessment of the safety of large language models.

C. Model-based Evaluation

Model-based evaluation refers to the utilization of high-performance models to assess the safety of text generated by large language models (LLMs), such as GPT-4 [78], BERT [91], and reward models [216]. In this approach, models function as "evaluators" that interpret, analyze, and determine whether the generated content adheres to specific safety standards. Model-based evaluation offers substantial advantages in assessment capabilities [211]–[213]: in contrast to rule-based evaluation, it can handle complex semantic structures and context-dependent nuances, thereby identifying content with implicit biases or harmful elements. Furthermore, compared to human evaluation, model-based evaluation can process large volumes of data rapidly and efficiently, significantly reducing the need for human involvement and associated costs.

However, model-based evaluation is not without its limitations. Its evaluation process depends on sophisticated model architectures and hyperparameters, which can render the decision-making process opaque and difficult to interpret. Moreover, the training data used to develop the evaluation models may carry inherent biases [215], which could influence the evaluation results, leading to potential inconsistencies, especially when assessing sensitive or controversial topics.

1) Closed-source Model: Closed-source models, as evaluators, offer high performance and versatility. Models like GPT-4 and InstructGPT, which are trained on extensive datasets, exhibit remarkable capabilities in understanding language and analyzing complex contexts. These models excel in tasks such as sentiment analysis, semantic interpretation, and intent detection, making them highly effective for evaluating content safety and quality. It cannot be denied that GPT-4-as-a-judge is the most popular evaluation method in the field of jailbreak attacks and has been proven to have strong correlation with human judgment. In addition, there are other ways to use closed source models to assist in evaluation, for instance:

Cai *et al.* [177] introduced three novel evaluation metrics—Safeguard Violation (SV), Informativeness (I), and Relative Truthfulness (RT)—to more precisely assess model jailbreak outcomes. The evaluation methodology leverages GPT-4 as the evaluator, utilizing predefined prompt templates to assess these three metrics. Several other studies [21], [49], [186], [217] have similarly employed GPT-4 as an evaluation tool. Sun *et al.* [17] proposed a benchmarking framework for assessing the safety of Chinese LLMs, with InstructGPT serving as the evaluator. Through dynamic prompt templates, InstructGPT is tasked with evaluating the safety of responses across various scenarios. The safety response rate for each scenario is calculated and aggregated to produce an overall safety ranking.

Despite their advantages, closed-source models present significant challenges regarding transparency and interpretability. The lack of public disclosure regarding their internal architectures and training data makes it difficult to offer clear explanations for the evaluation outcomes. This opacity can undermine trust, particularly in the context of sensitive content evaluation. Additionally, closed-source models have limited customization capabilities, which restrict their adaptability to specific evaluation needs, thus hindering their applicability in highly specialized contexts. While these models are well-suited for general evaluation tasks, their limitations in transparency and customization must be carefully considered when determining their use in particular scenarios.

2) Open-source Model: Several researchers have released open-source models to support safety evaluation efforts. These models are often fine-tuned on specific datasets to approach the performance of state-of-the-art closed-source models, and they can be deployed locally, reducing evaluation costs while improving reproducibility.

Llama Guard [137] is a safety evaluation model fine-tuned from the Llama2-7B model. It classifies responses based on user prompts and model outputs, and for responses labeled as unsafe, it also provides the violated safety policy. Meta researchers continue to update this series, with versions including Llama Guard 2 8B [219], Llama Guard 3 1B [220], and Llama Guard 3 8B [221].

ShieldLM [222] fine-tuned Qwen-14B-Chat [223] on a bilingual query-response dataset. Unlike most works that label responses as simply safe/unsafe, ShieldLM introduces a new label, "controversial." Under strict rules, controversial responses are marked as unsafe, while under loose rules, they are marked as safe. The model outputs both classification results and reasoning analyses.

MD-Judge [47], inspired by Llama Guard [137], fine-tuned the Mistral-7B model [224] for safety evaluation. Specifically, the researchers incorporated attack-augmented data in the fine-tuning process to further enhance the model's capabilities. The model outputs binary classification labels of safe/unsafe and the category of the unsafe response.

AegisSafetyExperts [218] is unique in that its goal is to build a collection of LLM-based safety experts. The researchers fine-tuned three models using their own constructed dataset. First, based on Llama Guard [137], they fine-tuned two versions using the proposed taxonomy and safety policies—Llama Guard Defensive and Llama Guard Permissive. The former maps safety categories that "Need Cation" to unsafe, while the latter maps them to safe. Additionally, the researchers fine-tuned the NeMo43B model using their taxonomy in the prompt, resulting in the NeMo43B-Defensive model for prompt classification.

WildGuard [51] also fine-tuned Mistral-7B [224] to detect malicious intent in user prompts, assess safety risks in model responses, and determine the model's refusal rate. WildGuard's training data includes synthetic adversarial, synthetic vanilla, and real-world user-LLM interaction data to maximize task coverage, diversity, and balance.

ShieldGemma [225] is designed specifically for safety content moderation. The researchers supervised fine-tuning of Gemma2 (2B, 9B, and 27B parameters), requiring outputs to include binary classification results and reasoning. Experiment results show that ShieldGemma outperforms both Llama Guard [137] and WildGuard [51] .

JailJudge Guard [194] is the latest research, fine-tuning the Llama2-7B model with instruction tuning. The model is trained to output detailed reasoning and a fine-grained evaluation score (a jailbreak score from 1 to 10). To further showcase

JailJudge Guard's capabilities, the researchers developed an attacker-agnostic attack enhancer and a system-level jailbreak defense method.

VI. CHALLENGES AND FUTURE DIRECTIONS

The rapid development and powerful capabilities of LLMs have led to their widespread application in the real world, making them a innovative tool for human society. However, approaches to thoroughly evaluate LLMs safety have lagged behind this pace of development. Existing safety evaluation methods remain insufficiently comprehensive, facing numerous challenges and offering substantial room for improvement. Therefore, in this section, we summarize the current challenges LLMs face and propose corresponding research directions.

A. Unified Evaluation

Currently, the research community has not reached a consensus on how to conduct specific safety evaluations for LLMs. There are differences in the specific implementation of evaluation methods and benchmarks used by different researchers. Especially for widely used key metrics such as attack success rate, there is no clear and specific measurement method. Such inconsistency makes it difficult to compare work across researchers and even raises questions about the reliability and accuracy of evaluation results [184]. Furthermore, as the application areas of LLMs continue to expand, the safety evaluation requirements vary across different fields, exacerbating these inconsistencies. Therefore, we believe it is urgent to develop a universal evaluation system that can support various task types, precisely calculate metrics, and adapt flexibly to the evaluation needs of different models and domains. Although it may seem like an unlikely goal to complete, future research should strive to approximate this ultimate objective by establishing standardized evaluation procedures, unified and comprehensive metrics, and a multi-layered evaluation framework.

B. Dynamic Evaluation

Most benchmarks for LLMs safety evaluation remain unchanged after being released by researchers, resulting in mainly static evaluations. This static approach presents three key challenges. First, as LLMs undergo continuous iteration and improvement, many prompts or attack methods that previously triggered unsafe responses or behaviors may now be effectively mitigated and fixed, leading to a "seemingly safe" performance on these benchmarks. Second, data from various safety evaluation datasets may leak, becoming part of LLMs fine-tuning aligned with human values, which could render these datasets ineffective for accurate evaluation. Third, attack methods continue to evolve. For example, jailbreak attacks can disguise malicious prompts that would have been rejected by the model to distract its attention and generate harmful content. As new jailbreak techniques emerge, benchmarks should be updated accordingly to further test model safety mechanisms. Therefore, it is essential to consider dynamic safety evaluation methods in the future, allowing evaluation data to iterate and evolve continuously. Practical implementation can draw on insights from related work on dynamic evaluation for general-purpose LLMs.

C. Reliable and Efficient Evaluator

As discussed in Section V, current evaluator approaches fall into three main categories: human-based, rule-based, and model-based. First, while rule-based methods are simple and fast, they have been shown to sometimes mislabel harmful outputs containing predefined strings as harmless, leading to false negatives. Second, although human-based evaluation is considered the gold standard due to its ability to accurately capture linguistic nuances and contextual meanings aligned with real-world values, it is not automated. This method incurs high time and cost demands, making it challenging to scale. Additionally, research has shown that differences in cultural background and moral values among human annotators can lead to significant variability in offense standards [198], affecting the stability and reliability of evaluation results. Furthermore, model-based evaluation offers a new direction for automated evaluation, with advanced models such as GPT-4 showing strong alignment with human judgments and capable of processing large volumes of data efficiently. However, model-based evaluation also faces numerous challenges, including high API costs, significant computational resources needed for model fine-tuning, poor reproducibility of evaluation results, sensitivity to input prompts, and positional biases [206], [213]. These factors can impact the accuracy and fairness of evaluations. Future research should consider how to strike a better balance between model-based and human-based evaluation approaches, achieving reliable evaluation results while maintaining high efficiency and low costs.

D. Specific Application Evaluation

As LLMs are increasingly integrated into various real-world applications, evaluating their safety in these specific application tasks has become crucial. For example, [199] developed a physical safety benchmark for evaluating LLMs in drone control systems, demonstrating the importance of targeted safety evaluations within specific domains. However, current safety evaluation efforts for LLMs mainly focus on general models, while safety research for the latest applications remains relatively disconnected, failing to keep pace with the integration of LLMs into other fields. Specifically, many emerging applications such as intelligent medical diagnosis, autonomous driving, weather forecasting, finance, and law have begun leveraging the powerful capabilities of LLMs, yet safety evaluation standards and tools tailored to these contexts are still underdeveloped. This lag not only risks failing to detect and mitigate potential safety hazards in a timely manner but may also limit the further application and development of LLMs in these fields. Therefore, it is urgent to establish safety evaluation frameworks for LLMs in these emerging domains to ensure their safe deployment.

E. Multimodal Model And Agent Evaluation

Multimodal large language models (MLLMs) and AI agents, both built upon LLMs, have garnered significant attention in both academia and industry. Multimodal models offer more natural and convenient interactions with humans, while AI

agents are used to emulate human-like thinking and decision-making processes. In MLLMs, LLMs play a central role in processing and generating text [201], while in agents, LLMs act as the "brain," handling thought and decision-making. However, the safety evaluation of these two directions currently lags behind their rapid development. Safety evaluation benchmarks for MLLMs remain limited [202]–[205], lacking targeted robustness metrics and evaluation standards, and there is insufficient focus on multimodal elements, especially in video-based domains, where evaluations are virtually nonexistent. Additionally, agents introduce numerous new components, creating new attack surfaces and requiring safety assessments in complex interactive environments. While some research has begun to address these challenges [40]–[42], there is still no unified consensus on the design standards for the safety benchmarks of the entire AI agent ecosystem [200]. Further safety evaluation issues remain to be solved, such as identifying the core dimensions of agent safety and developing comprehensive methods to evaluate agent performance.

F. Beyond Evaluation: Safe to Responsible

Safety evaluation should not be seen as the end goal, but rather as the starting point. It is not only about providing benchmark results for the safe deployment and application of LLMs but also offering direction and insights for future research and development. As the starting point, the core function of safety evaluation is to identify and mitigate potential risks, ensuring the model's stability and reliability. However, as technology advances and societal demands become more complex, mere safety will no longer suffice to meet high standards for application. Given the current development trends of LLMs and existing research on LLMs safety evaluation, we emphasize that the future evolution of LLMs should transition from merely safe LLMs to more advanced, responsible LLMs. Responsible LLMs should not only possess basic safety, fairness, and trustworthiness but also enhance transparency, interpretability, and auditability. This means that models must not only be technically reliable but also widely recognized for their ethical and social responsibility. Similarly, safety evaluation will evolve and improve alongside advancements in LLMs technology, encompassing a broader range of evaluation dimensions and deeper methods.

By expanding and refining research into the abrove areas, the community can more accurately evaluate the safety of LLMs while also promoting their development towards greater safety. In the future, LLMs will be closely aligned with human societal values and norms, evolving into a responsible form of artificial intelligence.

VII. Conclusion

To ensure the safe deployment and responsible utilization of LLMs in real-world applications, safety evaluation is of paramount importance. This paper presents the first systematic and comprehensive review of various aspects of LLMs safety evaluation, structured across four key guiding questions: "Why evaluate", "What to evaluate", "Where to evaluate", and "How to evaluate". We delve into dimensions of LLMs safety evaluation, including toxicity, robustness, ethics, bias and fairness, and truthfulness, as well as specific downstream tasks, such as privacy, mental health and medical applications, code generation, copyright, and agents. Additionally, we provide current evaluation metrics, toolkits, and an extensive range of datasets and benchmarks for safety evaluation. Furthermore, we summarize the current evaluator roles from three perspectives: human-based, rule-based, and model-based, with the goal of illustrating the state of development in LLMs safety evaluation and offering valuable reference points for researchers in the field. Finally, we highlight several major challenges currently faced in LLMs safety evaluation and discuss potential directions for future research.

This survey is expected to promote the advancement of LLM safety evaluation, providing clear guidance for the safe and controlled use of these models. As safety evaluation techniques continue to progress and improve, we are confident that future LLMs will serve human society more robustly and responsibly. These models will provide strong intelligent support for innovation and development across industries while effectively mitigating potential risks and ensuring a balanced approach between technological progress and trustworthy safety.

References

[1] Z. Lin, Z. Wang, Y. Tong, Y. Wang, Y. Guo, Y. Wang, and J. Shang, "Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation," *arXiv preprint arXiv:2310.17389*, 2023.

[2] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," *arXiv preprint arXiv:2203.09509*, 2022.

[3] M. Kim, J. Koo, H. Lee, J. Park, H. Lee, and K. Jung, "Lifetox: Unveiling implicit toxicity in life advice," *arXiv preprint arXiv:2311.09585*, 2023.

[4] S. Cui, Z. Zhang, Y. Chen, W. Zhang, T. Liu, S. Wang, and T. Liu, "Fft: Towards harmlessness evaluation and analysis for llms with factuality, fairness, toxicity," *arXiv preprint arXiv:2311.18580*, 2023.

[5] D. Jain, P. Kumar, S. Gehman, X. Zhou, T. Hartvigsen, and M. Sap, "Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models," *arXiv preprint arXiv:2405.09373*, 2024.

[6] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity," *arXiv preprint arXiv:2301.12867*, 2023.

[7] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[8] R. Bhardwaj and S. Poria, "Red-teaming large language models using chain of utterances for safety-alignment," *arXiv preprint arXiv:2308.09662*, 2023.

[9] H. Qiu, S. Zhang, A. Li, H. He, and Z. Lan, "Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models," *arXiv preprint arXiv:2307.08487*, 2023.

[10] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li *et al.*, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," *arXiv preprint arXiv:2402.04249*, 2024.

[11] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer *et al.*, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *arXiv preprint arXiv:2404.01318*, 2024.

[12] A. Mei, S. Levy, and W. Y. Wang, "Assert: Automated safety scenario red teaming for evaluating the robustness of large language models," *arXiv preprint arXiv:2310.09624*, 2023.

[13] R. Cantini, G. Cosenza, A. Orsino, and D. Talia, "Are large language models really bias-free? jailbreak prompts for assessing adversarial robustness to bias elicitation," *arXiv preprint arXiv:2407.08441*, 2024.

[14] S. Banerjee, S. Layek, R. Hazra, and A. Mukherjee, "How (un) ethical are instruction-centric responses of llms? unveiling the vulnerabilities of safety guardrails to harmful queries," *arXiv preprint arXiv:2402.15302*, 2024.

[15] G. I. Meadows, N. W. L. Lau, E. A. Susanto, C. L. Yu, and A. Paul, "Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models," *arXiv preprint arXiv:2408.01460*, 2024.

[16] Y. Huang, Q. Zhang, L. Sun *et al.*, "Trustgpt: A benchmark for trustworthy and responsible large language models," *arXiv preprint arXiv:2306.11507*, 2023.

[17] H. Sun, Z. Zhang, J. Deng, J. Cheng, and M. Huang, "Safety assessment of chinese large language models," *arXiv preprint arXiv:2304.10436*, 2023.

[18] A. Biswas and W. Talukdar, "Guardrails for trust, safety, and ethical development and deployment of large language models (llm)," *Journal of Science & Technology*, vol. 4, no. 6, pp. 55–82, 2023.

[19] P. Senthilkumar, V. Balasubramanian, P. Jain, A. Maity, J. Lu, and K. Zhu, "Fine-tuning language models for ethical ambiguity: A comparative study of alignment with human responses," *arXiv preprint arXiv:2410.07826*, 2024.

[20] J. Ji, Y. Chen, M. Jin, W. Xu, W. Hua, and Y. Zhang, "Moralbench: Moral evaluation of llms," *arXiv preprint arXiv:2406.04428*, 2024.

[21] S. Raza, A. Raval, and V. Chatrath, "Mbias: Mitigating bias in large language models while retaining context," *arXiv preprint arXiv:2405.11290*, 2024.

[22] R. Azeem, A. Hundt, M. Mansouri, and M. Brandão, "Llm-driven robots risk enacting discrimination, violence, and unlawful actions," *arXiv preprint arXiv:2406.08824*, 2024.

[23] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 993–999.

[24] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[25] Y. Huang and D. Xiong, "Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models," *arXiv preprint arXiv:2306.16244*, 2023.

[26] Z. Su, X. Zhou, S. Rangreji, A. Kabra, J. Mendelsohn, F. Brahman, and M. Sap, "Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents," *arXiv preprint arXiv:2409.09013*, 2024.

[27] S. Wu, Y. Huang, C. Gao, D. Chen, Q. Zhang, Y. Wan, T. Zhou, X. Zhang, J. Gao, C. Xiao *et al.*, "Unigen: A unified framework for textual dataset generation using large language models," *arXiv preprint arXiv:2406.18966*, 2024.

[28] E. A. Alghamdi, R. I. Masoud, D. Alnuhait, A. Y. Alomairi, A. Ashraf, and M. Zaytoon, "Aratrust: An evaluation of trustworthiness for llms in arabic," *arXiv preprint arXiv:2403.09017*, 2024.

[29] J. Hoscilowicz, A. Wiacek, J. Chojnacki, A. Cieslak, L. Michon, and A. Janicki, "Non-linear inference time intervention: Improving llm truthfulness," in *Proc. Interspeech 2024*, 2024, pp. 4094–4098.

[30] A. Khatun and D. G. Brown, "Trutheval: A dataset to evaluate llm truthfulness and reliability," *arXiv preprint arXiv:2406.01855*, 2024.

[31] W. Wang, Z. Tu, C. Chen, Y. Yuan, J.-t. Huang, W. Jiao, and M. R. Lyu, "All languages matter: On the multilingual safety of large language models," *arXiv preprint arXiv:2310.00905*, 2023.

[32] J. Song, Y. Huang, Z. Zhou, and L. Ma, "Multilingual blending: Llm safety alignment evaluation with language mixture," *arXiv preprint arXiv:2407.07342*, 2024.

[33] G. Dong, H. Wang, J. Sun, and X. Wang, "Evaluating and mitigating linguistic discrimination in large language models," *arXiv preprint arXiv:2404.18534*, 2024.

[34] J. I. Park, M. Abbasian, I. Azimi, D. Bounds, A. Jun, J. Han, R. McCarron, J. Borelli, J. Li, M. Mahmoudi *et al.*, "Building trust in mental health chatbots: Safety metrics and llm-based evaluation tools," *arXiv preprint arXiv:2408.04650*, 2024.

[35] J. Wang, X. Luo, L. Cao, H. He, H. Huang, J. Xie, A. Jatowt, and Y. Cai, "Is your ai-generated code really secure? evaluating large language models on secure code generation with codeseceval," *arXiv preprint arXiv:2407.02395*, 2024.

[36] M. Bhatt, S. Chennabasappa, C. Nikolaidis, S. Wan, I. Evtimov, D. Gabi, D. Song, F. Ahmad, C. Aschermann, L. Fontana *et al.*, "Purple llama cyberseceval: A secure coding benchmark for language models," *arXiv preprint arXiv:2312.04724*, 2023.

[37] X. Liu, T. Sun, T. Xu, F. Wu, C. Wang, X. Wang, and J. Gao, "Shield: Evaluation and defense strategies for copyright compliance in llm text generation," *arXiv preprint arXiv:2406.12975*, 2024.

[38] X. Zhu, Y. Liu, Z. Shen, Y. Liu, M. Li, Y. Chen, B. John, Z. Ma, T. Hu, B. Yang *et al.*, "How privacy-savvy are large language models? a case study on compliance and privacy technical review," *arXiv preprint arXiv:2409.02375*, 2024.

[39] Q. Li, J. Hong, C. Xie, J. Tan, R. Xin, J. Hou, X. Yin, Z. Wang, D. Hendrycks, Z. Wang *et al.*, "Llm-pbe: Assessing data privacy in large language models," *arXiv preprint arXiv:2408.12787*, 2024.

[40] T. Yuan, Z. He, L. Dong, Y. Wang, R. Zhao, T. Xia, L. Xu, B. Zhou, F. Li, Z. Zhang *et al.*, "R-judge: Benchmarking safety risk awareness for llm agents," *arXiv preprint arXiv:2401.10019*, 2024.

[41] Z. Zhu, B. Wu, Z. Zhang, and B. Wu, "Riskawarebench: Towards evaluating physical risk awareness for high-level planning of llm-based embodied agents," *arXiv preprint arXiv:2408.04449*, 2024.

[42] H. Zhang, J. Huang, K. Mei, Y. Yao, Z. Wang, C. Zhan, H. Wang, and Y. Zhang, "Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents," *arXiv preprint arXiv:2410.02644*, 2024.

[43] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: A dataset for evaluating safeguards in llms," *arXiv preprint arXiv:2308.13387*, 2023.

[44] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 537–15 553.

[45] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "" do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," *arXiv preprint arXiv:2308.03825*, 2023.

[46] G. Xu, J. Liu, M. Yan, H. Xu, J. Si, Z. Zhou, P. Yi, X. Gao, J. Sang, R. Zhang *et al.*, "Cvalues: Measuring the values of chinese large language models from safety to responsibility," *arXiv preprint arXiv:2307.09705*, 2023.

[47] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, and J. Shao, "Salad-bench: A hierarchical and comprehensive safety benchmark for large language models," *arXiv preprint arXiv:2402.05044*, 2024.

[48] N. Varshney, P. Dolin, A. Seth, and C. Baral, "The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness," *arXiv preprint arXiv:2401.00287*, 2023.

[49] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "Xstest: A test suite for identifying exaggerated safety behaviours in large language models," *arXiv preprint arXiv:2308.01263*, 2023.

[50] R. K. Sharma, V. Gupta, and D. Grossman, "Spml: A dsl for defending language models against prompt attacks," *arXiv preprint arXiv:2402.11755*, 2024.

[51] S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, and N. Dziri, "Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms," *arXiv preprint arXiv:2406.18495*, 2024.

[52] C. Liu, F. Zhao, L. Qing, Y. Kang, C. Sun, K. Kuang, and F. Wu, "Goal-oriented prompt attack and safety evaluation for llms," *arXiv e-prints*, pp. arXiv–2309, 2023.

[53] K. Chen, Y. Liu, D. Wang, J. Chen, and W. Wang, "Characterizing and evaluating the reliability of llms against jailbreak attacks," *arXiv preprint arXiv:2408.09326*, 2024.

[54] P. Gupta, L. Q. Yau, H. H. Low, I. Lee, H. M. Lim, Y. X. Teoh, J. H. Koh, D. W. Liew, R. Bhardwaj, R. Bhardwaj *et al.*, "Walledeval: A comprehensive safety evaluation toolkit for large language models," *arXiv preprint arXiv:2408.03837*, 2024.

[55] E. Yu, J. Li, M. Liao, S. Wang, Z. Gao, F. Mi, and L. Hong, "Cosafe: Evaluating large language model safety in multi-turn dialogue coreference," *arXiv preprint arXiv:2406.17626*, 2024.

[56] T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwag, K. Huang, L. He, B. Wei, D. Li, Y. Sheng *et al.*, "Sorry-bench: Systematically evaluating large language model safety refusal behaviors," *arXiv preprint arXiv:2406.14598*, 2024.

[57] X. Yuan, J. Li, D. Wang, Y. Chen, X. Mao, L. Huang, H. Xue, W. Wang, K. Ren, and J. Wang, "S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models," *arXiv preprint arXiv:2405.14191*, 2024.

[58] B. An, S. Zhu, R. Zhang, M.-A. Panaitescu-Liess, Y. Xu, and F. Huang, "Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models," *arXiv preprint arXiv:2409.00598*, 2024.

[59] W. Zhang, X. Lei, Z. Liu, M. An, B. Yang, K. Zhao, K. Wang, and S. Lian, "Chisafetybench: A chinese hierarchical safety benchmark for large language models," *arXiv preprint arXiv:2406.10311*, 2024.

[60] L. Shi and D. Xiong, "Criskeval: A chinese multi-level risk evaluation benchmark dataset for large language models," *arXiv preprint arXiv:2406.04752*, 2024.

[61] Y. Wang, Z. Zhai, H. Li, X. Han, L. Lin, Z. Zhang, J. Zhao, P. Nakov, and T. Baldwin, "A chinese dataset for evaluating the safeguards in large language models," *arXiv preprint arXiv:2402.12193*, 2024.

[62] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, and B. Li, "Alert: A comprehensive benchmark for assessing large language models' safety through red teaming," *arXiv preprint arXiv:2404.08676*, 2024.

[63] J. Chu, Y. Liu, Z. Yang, X. Shen, M. Backes, and Y. Zhang, "Comprehensive assessment of jailbreak attacks against llms," *arXiv preprint arXiv:2402.05668*, 2024.

[64] H. Zhao, Y. Liu, S. Tao, W. Meng, Y. Chen, X. Geng, C. Su, M. Zhang, and H. Yang, "From handcrafted features to llms: A brief survey for machine translation quality estimation," *arXiv preprint arXiv:2403.14118*, 2024.

[65] L. Chen, Q. Guo, H. Jia, Z. Zeng, X. Wang, Y. Xu, J. Wu, Y. Wang, Q. Gao, J. Wang *et al.*, "A survey on evaluating large language models in code generation tasks," *arXiv preprint arXiv:2408.16498*, 2024.

[66] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[67] A. L. Kotian, R. Nandipi, M. Ushag, G. Veena *et al.*, "A systematic review on human and computer interaction," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2024, pp. 1214–1218.

[68] P. Mondorf and B. Plank, "Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey," *arXiv preprint arXiv:2404.01869*, 2024.

[69] F. W. Liu and C. Hu, "Exploring vulnerabilities and protections in large language models: A survey," *arXiv preprint arXiv:2406.00240*, 2024.

[70] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736*, 2023.

[71] J.-L. Peng, S. Cheng, E. Diau, Y.-Y. Shih, P.-H. Chen, Y.-T. Lin, and Y.-N. Chen, "A survey of useful llm evaluation," *arXiv preprint arXiv:2406.00936*, 2024.

[72] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," *arXiv preprint arXiv:2308.05374*, 2023.

[73] K. Kenthapadi, M. Sameki, and A. Taly, "Grounding and evaluation for large language models: Practical challenges and lessons learned (survey)," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6523–6533.

[74] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[75] S. Ye, H. Hwang, S. Yang, H. Yun, Y. Kim, and M. Seo, "Investigating the effectiveness of task-agnostic prefix prompt for instruction following," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 386–19 394.

[76] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[77] Y. Wang, H. Le, A. D. Gotmare, N. D. Bui, J. Li, and S. C. Hoi, "Codet5+: Open code large language models for code understanding and generation," *arXiv preprint arXiv:2305.07922*, 2023.

[78] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[79] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang *et al.*, "Zero-shot information extraction via chatting with chatgpt," *arXiv preprint arXiv:2302.10205*, 2023.

[80] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," *arXiv preprint arXiv:2304.05335*, 2023.

[81] C. Chataigner, A. Taïk, and G. Farnadi, "Multilingual hallucination gaps in large language models," *arXiv preprint arXiv:2410.18270*, 2024.

[82] W. Lan, W. Chen, Q. Chen, S. Pan, H. Zhou, and Y. Pan, "A survey of hallucination in large visual language models," *arXiv preprint arXiv:2410.15359*, 2024.

[83] C. Maple, L. Szpruch, G. Epiphaniou, K. Staykova, S. Singh, W. Penwarden, Y. Wen, Z. Wang, J. Hariharan, and P. Avramovic, "The ai revolution: opportunities and challenges for the finance sector," *arXiv preprint arXiv:2308.16538*, 2023.

[84] I. Weissburg, S. Anand, S. Levy, and H. Jeong, "Llms are biased teachers: Evaluating llm bias in personalized education," *arXiv preprint arXiv:2410.14012*, 2024.

[85] J. Haltaufderheide and R. Ranisch, "The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms)," *NPJ digital medicine*, vol. 7, no. 1, p. 183, 2024.

[86] F. Torrielli, "Stars, stripes, and silicon: Unravelling the chatgpt's all-american, monochrome, cis-centric bias," *arXiv preprint arXiv:2410.13868*, 2024.

[87] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[88] M. N. Sakib, M. A. Islam, R. Pathak, and M. M. Arifin, "Risks, causes, and mitigations of widespread deployments of large language models (llms): A survey," *arXiv preprint arXiv:2408.04643*, 2024.

[89] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "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.

[90] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[91] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.

[92] Y. Liu, J. Yu, H. Sun, L. Shi, G. Deng, Y. Chen, and Y. Liu, "Efficient detection of toxic prompts in large language models," *arXiv preprint arXiv:2408.11727*, 2024.

[93] A. Leidinger and R. Rogers, "How are llms mitigating stereotyping harms? learning from search engine studies," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 839–854.

[94] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, and F. M. Zanzotto, "A trip towards fairness: Bias and de-biasing in large language models," *arXiv preprint arXiv:2305.13862*, 2023.

[95] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[96] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos, "Semdedup: Data-efficient learning at web-scale through semantic deduplication," *arXiv preprint arXiv:2303.09540*, 2023.

[97] Y. Yang, Q. Jin, Q. Zhu, Z. Wang, F. E. Álvarez, N. Wan, B. Hou, and Z. Lu, "Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare," *arXiv preprint arXiv:2410.18460*, 2024.

[98] A. Prabhu Desai, G. Satish Mallya, M. Luqman, T. Ravi, N. Kota, and P. Yadav, "Opportunities and challenges of generative-ai in finance," *arXiv e-prints*, pp. arXiv–2410, 2024.

[99] S. Curran, S. Lansley, and O. Bethell, "Hallucination is the last thing you need," *arXiv preprint arXiv:2306.11520*, 2023.

[100] J. Wen, P. Ke, H. Sun, Z. Zhang, C. Li, J. Bai, and M. Huang, "Unveiling the implicit toxicity in large language models," *arXiv preprint arXiv:2311.17391*, 2023.

[101] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[102] Google, "Google Perspective API," https://www.perspectiveapi.com/, accessed: 2024-10-01.

[103] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[104] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, K. Wang, and Y. Liu, "Jailbreaking chatgpt via prompt engineering: An empirical study," *arXiv preprint arXiv:2305.13860*, 2023.

[105] R. Zhang, H. Li, Y. Wu, Q. Ai, Y. Liu, M. Zhang, and S. Ma, "Evaluation ethics of llms in legal domain," *arXiv preprint arXiv:2403.11152*, 2024.

[106] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning ai with shared human values," *arXiv preprint arXiv:2008.02275*, 2020.

[107] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[108] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[109] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," *arXiv preprint arXiv:2212.08061*, 2022.

[110] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[111] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, 2022.

[112] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in experimental social psychology*. Elsevier, 2013, vol. 47, pp. 55–130.

[113] Y. Wang, X. Wu, H.-T. Wu, Z. Tao, and Y. Fang, "Do large language models rank fairly? an empirical study on the fairness of llms as rankers," *arXiv preprint arXiv:2404.03192*, 2024.

[114] S. Wang, P. Wang, T. Zhou, Y. Dong, Z. Tan, and J. Li, "Ceb: Compositional evaluation benchmark for fairness in large language models," *arXiv preprint arXiv:2407.02408*, 2024.

[115] C. Agarwal, H. Lakkaraju, and M. Zitnik, "Towards a unified framework for fair and stable graph representation learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 2114–2124.

[116] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, "Large language models are not fair evaluators," *arXiv preprint arXiv:2305.17926*, 2023.

[117] J. Hong, S. Levine, and A. Dragan, "Zero-shot goal-directed dialogue via rl on imagined conversations," *arXiv preprint arXiv:2311.05584*, 2023.

[118] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, "Sources of hallucination by large language models on inference tasks," *arXiv preprint arXiv:2305.14552*, 2023.

[119] F. S. Bao, M. Li, R. Qu, G. Luo, E. Wan, Y. Tang, W. Fan, M. S. Tamber, S. Kazi, V. Sourabh *et al.*, "Faithbench: A diverse hallucination benchmark for summarization by modern llms," *arXiv preprint arXiv:2410.13210*, 2024.

[120] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, "Siren's song in the ai ocean: A survey on hallucination in large language models, 2023," *URL https://arxiv.org/abs/2309.01219*, 2024.

[121] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig *et al.*, "Sotopia: Interactive evaluation for social intelligence in language agents," *arXiv preprint arXiv:2310.11667*, 2023.

[122] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.

[123] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[124] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[125] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[126] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[127] M. Liang, A. Arun, Z. Wu, C. Munoz, J. Lutch, E. Kazim, A. Koshiyama, and P. Treleaven, "Thames: An end-to-end tool for hallucination mitigation and evaluation in large language models," *arXiv preprint arXiv:2409.11353*, 2024.

[128] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "Peft: State-of-the-art parameter-efficient fine-tuning methods," *URL: https://github. com/huggingface/peft*, 2022.

[129] S. Feng, V. Balachandran, Y. Bai, and Y. Tsvetkov, "Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge," *arXiv preprint arXiv:2305.08281*, 2023.

[130] L. Tang, T. Goyal, A. R. Fabbri, P. Laban, J. Xu, S. Yavuz, W. Kryściński, J. F. Rousseau, and G. Durrett, "Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors," *arXiv preprint arXiv:2205.12854*, 2022.

[131] S. Jain, V. Keshava, S. M. Sathyendra, P. Fernandes, P. Liu, G. Neubig, and C. Zhou, "Multi-dimensional evaluation of text summarization with in-context learning," *arXiv preprint arXiv:2306.01200*, 2023.

[132] Z. Wei, Y. Wang, A. Li, Y. Mo, and Y. Wang, "Jailbreak and guard aligned language models with only few in-context demonstrations," *arXiv preprint arXiv:2310.06387*, 2023.

[133] Y. Yuan, W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu, "Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher," *arXiv preprint arXiv:2308.06463*, 2023.

[134] P. Ding, J. Kuang, D. Ma, X. Cao, Y. Xian, J. Chen, and S. Huang, "A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily," *arXiv preprint arXiv:2311.08268*, 2023.

[135] W. Wang and S. Pan, "Deep inductive logic reasoning for multi-hop reading comprehension," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4999–5009.

[136] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models." in *NeurIPS*, 2023.

[137] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, "Llama guard: Llm-based input-output safeguard for human-ai conversations," *arXiv preprint arXiv:2312.06674*, 2023.

[138] M. Andriushchenko, F. Croce, and N. Flammarion, "Jailbreaking leading safety-aligned llms with simple adaptive attacks," *arXiv preprint arXiv:2404.02151*, 2024.

[139] Alex Albert, "Jailbreak chat," https://www.jailbreakchat.com/, accessed: 2024-02-20.

[140] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, 2022.

[141] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li, "Adversarial glue: A multi-task benchmark for robustness evaluation of language models," *arXiv preprint arXiv:2111.02840*, 2021.

[142] C. Li and J. Flanigan, "Rac: Efficient llm factuality correction with retrieval augmentation," *arXiv preprint arXiv:2410.15667*, 2024.

[143] S. Weijia, M. Sewon, Y. Michihiro, S. Minjoon, J. Rich, L. Mike, and Y. Wen-tau, "Replug: Retrieval-augmented black-box language models," *ArXiv: 2301.12652*, 2023.

[144] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[145] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *stat*, vol. 1050, no. 9, 2017.

[146] X. Yu, H. Cheng, X. Liu, D. Roth, and J. Gao, "Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 1333–1351.

[147] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," *arXiv preprint arXiv:2305.11747*, 2023.

[148] S. Zheng, J. Huang, and K. C.-C. Chang, "Why does chatgpt fall short in answering questions faithfully," *arXiv preprint arXiv:2304.10513*, 2023.

[149] S. Das, S. Saha, and R. K. Srihari, "Diving deep into modes of fact hallucinations in dialogue systems," *arXiv preprint arXiv:2301.04449*, 2023.

[150] M. Cao, Y. Dong, and J. C. K. Cheung, "Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization," *arXiv preprint arXiv:2109.09784*, 2021.

[151] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[152] Y. Wang, M. Wang, H. Iqbal, G. Georgiev, J. Geng, and P. Nakov, "Openfactcheck: A unified framework for factuality evaluation of llms," *arXiv preprint arXiv:2405.05583*, 2024.

[153] H. Sansford, N. Richardson, H. P. Maretic, and J. N. Saada, "Grapheval: A knowledge-graph based llm hallucination evaluation framework," *arXiv preprint arXiv:2407.10793*, 2024.

[154] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Can knowledge graphs reduce hallucinations in llms?: A survey," *arXiv preprint arXiv:2311.07914*, 2023.

[155] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.

[156] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," *arXiv preprint arXiv:2004.04228*, 2020.

[157] Y. Zhu, J. Xiao, Y. Wang, and J. Sang, "Kg-fpq: Evaluating factuality hallucination in llms with knowledge graph-based false premise questions," *arXiv preprint arXiv:2407.05868*, 2024.

[158] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, "Siren's song in the ai ocean: a survey on hallucination in large language models," *arXiv preprint arXiv:2309.01219*, 2023.

[159] D. Muhlgay, O. Ram, I. Magar, Y. Levine, N. Ratner, Y. Belinkov, O. Abend, K. Leyton-Brown, A. Shashua, and Y. Shoham, "Generating benchmarks for factuality evaluation of language models," *arXiv preprint arXiv:2307.06908*, 2023.

[160] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *arXiv preprint arXiv:2305.14251*, 2023.

[161] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "Propile: Probing privacy leakage in large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[162] M. Fan, C. Chen, C. Wang, and J. Huang, "On the trustworthiness landscape of state-of-the-art generative models: A comprehensive survey," *arXiv preprint arXiv:2307.16680*, 2023.

[163] H. Shao, J. Huang, S. Zheng, and K. C.-C. Chang, "Quantifying association capabilities of large language models and its implications on privacy leakage," *arXiv preprint arXiv:2305.12707*, 2023.

[164] A. Panda, C. A. Choquette-Choo, Z. Zhang, Y. Yang, and P. Mittal, "Teach llms to phish: Stealing private information from language models," *arXiv preprint arXiv:2403.00871*, 2024.

[165] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[166] B. Hui, H. Yuan, N. Gong, P. Burlina, and Y. Cao, "Pleak: Prompt leaking attacks against large language model applications," *arXiv preprint arXiv:2405.06823*, 2024.

[167] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," *arXiv preprint arXiv:2310.08419*, 2023.

[168] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[169] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *European conference on machine learning*. Springer, 2004, pp. 217–226.

[170] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, and P. Malakasiotis, "Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases," *arXiv preprint arXiv:2103.13084*, 2021.

[171] M. Abbasian, E. Khatibi, I. Azimi, D. Oniani, Z. Shakeri Hossein Abad, A. Thieme, R. Sriram, Z. Yang, Y. Wang, B. Lin *et al.*, "Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai," *NPJ Digital Medicine*, vol. 7, no. 1, p. 82, 2024.

[172] R. Khoury, A. R. Avila, J. Brunelle, and B. M. Camara, "How secure is code generated by chatgpt?" in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 2445–2451.

[173] N. Perry, M. Srivastava, D. Kumar, and D. Boneh, "Do users write more insecure code with ai assistants?" in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 2785–2799.

[174] O. Asare, M. Nagappan, and N. Asokan, "Is github's copilot as bad as humans at introducing vulnerabilities in code?" *Empirical Software Engineering*, vol. 28, no. 6, p. 129, 2023.

[175] K. K. Chang, M. Cramer, S. Soni, and D. Bamman, "Speak, memory: An archaeology of books known to chatgpt/gpt-4," *arXiv preprint arXiv:2305.00118*, 2023.

[176] U. Hacohen, A. Haviv, S. Sarfaty, B. Friedman, N. Elkin-Koren, R. Livni, and A. H. Bermano, "Not all similarities are created equal: Leveraging data-driven biases to inform genai copyright disputes," *arXiv preprint arXiv:2403.17691*, 2024.

[177] H. Cai, A. Arunasalam, L. Y. Lin, A. Bianchi, and Z. B. Celik, "Take a look at it! rethinking how to evaluate language model jailbreak," *arXiv preprint arXiv:2404.06407*, 2024.

[178] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.

[179] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[180] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of llms: Dual-use through standard security attacks," in *2024 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2024, pp. 132–143.

[181] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr, "Agentdojo: A dynamic environment to evaluate attacks and defenses for llm agents," *arXiv preprint arXiv:2406.13352*, 2024.

[182] W. Zou, R. Geng, B. Wang, and J. Jia, "Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models," *arXiv preprint arXiv:2402.07867*, 2024.

[183] T. Han, A. Kumar, C. Agarwal, and H. Lakkaraju, "Medsafetybench: Evaluating and improving the medical safety of large language models," *arXiv preprint arXiv:2403.03744*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.03744

[184] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins *et al.*, "A strongreject for empty jailbreaks," *arXiv preprint arXiv:2402.10260*, 2024.

[185] W. Wang, Z. Tu, C. Chen, Y. Yuan, J.-t. Huang, W. Jiao, and M. Lyu, "All languages matter: On the multilingual safety of llms," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 5865–5877.

[186] M. K. B. Doumbouya, A. Nandi, G. Poesia, D. Ghilardi, A. Goldie, F. Bianchi, D. Jurafsky, and C. D. Manning, "h4rm3l: A dynamic benchmark of composable jailbreak attacks for llm safety assessment," *arXiv preprint arXiv:2408.04811*, 2024.

[187] D. Ran, J. Liu, Y. Gong, J. Zheng, X. He, T. Cong, and A. Wang, "Jailbreakeval: An integrated toolkit for evaluating jailbreak attempts against large language models," *arXiv preprint arXiv:2406.09321*, 2024.

[188] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[189] H. Zhang, H. Gao, Q. Hu, G. Chen, L. Yang, B. Jing, H. Wei, B. Wang, H. Bai, and L. Yang, "Chinesesafe: A chinese benchmark for evaluating safety in large language models," *arXiv preprint arXiv:2410.18491*, 2024.

[190] L. Mei, S. Liu, Y. Wang, B. Bi, J. Mao, and X. Cheng, "" not aligned" is not" malicious": Being careful about hallucinations of large language models' jailbreak," *arXiv preprint arXiv:2406.11668*, 2024.

[191] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[192] W. Zhou, X. Wang, L. Xiong, H. Xia, Y. Gu, M. Chai, F. Zhu, C. Huang, S. Dou, Z. Xi *et al.*, "Easyjailbreak: A unified framework for jailbreaking large language models," *arXiv preprint arXiv:2403.12171*, 2024.

[193] M. Zhang, X. Pan, and M. Yang, "Jade: A linguistics-based safety evaluation platform for llm," *arXiv preprint arXiv:2311.00286*, 2023.

[194] Anonymous, "JAILJUDGE: A comprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework," in *Submitted to The Thirteenth International Conference on Learning Representations*, 2024, under review. [Online]. Available: https://openreview.net/forum?id=cLYvhd0pDY

[195] Q. Ren, H. Li, D. Liu, Z. Xie, X. Lu, Y. Qiao, L. Sha, J. Yan, L. Ma, and J. Shao, "Derail yourself: Multi-turn llm jailbreak attack through self-discovered clues," *arXiv preprint arXiv:2410.10700*, 2024.

[196] X. Sun, D. Zhang, D. Yang, Q. Zou, and H. Li, "Multi-turn context jailbreak attack on large language models from first principles," *arXiv preprint arXiv:2408.04686*, 2024.

[197] M. Russinovich, A. Salem, and R. Eldan, "Great, now write an article about that: The crescendo multi-turn llm jailbreak attack," *arXiv preprint arXiv:2404.01833*, 2024.

[198] A. M. Davani, M. Díaz, D. Baker, and V. Prabhakaran, "D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation," *arXiv preprint arXiv:2404.10857*, 2024.

[199] Y.-C. Tang, P.-Y. Chen, and T.-Y. Ho, "Defining and evaluating physical safety for large language models," *arXiv preprint arXiv:2411.02317*, 2024.

[200] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, and Y. Xiang, "Ai agents under threat: A survey of key security challenges and future pathways," *arXiv preprint arXiv:2406.02630*, 2024.

[201] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[202] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "Mm-safetybench: A benchmark for safety evaluation of multimodal large language models," in *European Conference on Computer Vision*. Springer, 2025, pp. 386–403.

[203] H. Tu, C. Cui, Z. Wang, Y. Zhou, B. Zhao, J. Han, W. Zhou, H. Yao, and C. Xie, "How many unicorns are in this image? a safety evaluation benchmark for vision llms," *arXiv preprint arXiv:2311.16101*, 2023.

[204] W. Luo, S. Ma, X. Liu, X. Guo, and C. Xiao, "Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks," *arXiv preprint arXiv:2404.03027*, 2024.

[205] T. Gu, Z. Zhou, K. Huang, D. Liang, Y. Wang, H. Zhao, Y. Yao, X. Qiao, K. Wang, Y. Yang *et al.*, "Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models," *arXiv preprint arXiv:2406.07594*, 2024.

[206] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024.

[207] M. Karpinska, N. Akoury, and M. Iyyer, "The perils of using mechanical turk to evaluate open-ended text generation," *arXiv preprint arXiv:2109.06835*, 2021.

[208] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why we need new evaluation metrics for nlg," *arXiv preprint arXiv:1707.06875*, 2017.

[209] R. Movva, P. W. Koh, and E. Pierson, "Annotation alignment: Comparing llm and human annotations of conversational safety," *arXiv preprint arXiv:2406.06369*, 2024.

[210] J. Ye, S. Li, G. Li, C. Huang, S. Gao, Y. Wu, Q. Zhang, T. Gui, and X. Huang, "Toolsword: Unveiling safety issues of large language models in tool learning across three stages," *arXiv preprint arXiv:2402.10753*, 2024.

[211] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" *arXiv preprint arXiv:2305.01937*, 2023.

[212] Y. Chen, R. Wang, H. Jiang, S. Shi, and R. Xu, "Exploring the use of large language models for reference-free text quality evaluation: An empirical study," *arXiv preprint arXiv:2304.00723*, 2023.

[213] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.

[214] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen, "Evaluating large language models at evaluating instruction following," *arXiv preprint arXiv:2310.07641*, 2023.

[215] A. Panickssery, S. R. Bowman, and S. Feng, "Llm evaluators recognize and favor their own generations," *arXiv preprint arXiv:2404.13076*, 2024.

[216] S. Ge, C. Zhou, R. Hou, M. Khabsa, Y.-C. Wang, Q. Wang, J. Han, and Y. Mao, "Mart: Improving llm safety with multi-round automatic red-teaming," *arXiv preprint arXiv:2311.07689*, 2023.

[217] M. Jin, S. Zhu, B. Wang, Z. Zhou, C. Zhang, Y. Zhang *et al.*, "Attackeval: How to evaluate the effectiveness of jailbreak attacking on large language models," *arXiv preprint arXiv:2401.09002*, 2024.

[218] S. Ghosh, P. Varshney, E. Galinkin, and C. Parisien, "Aegis: Online adaptive ai content safety moderation with ensemble of llm experts," *arXiv preprint arXiv:2404.05993*, 2024.

[219] Llama Team, "Meta llama guard 2," https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.

[220] Llama Team, AI @ Meta, "The llama 3 family of models," https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL_CARD.md, 2024.

[221] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[222] Z. Zhang, Y. Lu, J. Ma, D. Zhang, R. Li, P. Ke, H. Sun, L. Sha, Z. Sui, H. Wang *et al.*, "Shieldlm: Empowering llms as aligned, customizable and explainable safety detectors," *arXiv preprint arXiv:2402.16444*, 2024.

[223] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[224] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[225] W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu *et al.*, "Shieldgemma: Generative ai content moderation based on gemma," *arXiv preprint arXiv:2407.21772*, 2024.

[226] A. Karamolegkou, J. Li, L. Zhou, and A. Søgaard, "Copyright violations and large language models," *arXiv preprint arXiv:2310.13771*, 2023.

[227] S. Liu, S. Cui, H. Bu, Y. Shang, and X. Zhang, "Jailbench: A comprehensive chinese security assessment benchmark for large language models," *arXiv preprint arXiv:2502.18935*, 2025.

[228] R. Pu, C. Li, R. Ha, L. Zhang, L. Qiu, and X. Zhang, "Baitattack: Alleviating intention shift in jailbreak attacks via adaptive bait crafting," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 15 654–15 668.

[229] R. Pu, C. Li, R. Ha, Z. Chen, L. Zhang, Z. Liu, L. Qiu, and X. Zhang, "Feint and attack: Attention-based strategies for jailbreaking and protecting llms," *arXiv preprint arXiv:2410.16327*, 2024.

[230] L. Zhang, X. Zhang, Z. Zhou, F. Huang, and C. Li, "Reinforced adaptive knowledge learning for multimodal fake news detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 15, 2024, pp. 16 777–16 785.

[231] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. S. Yu, "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10102*, 2024.

[232] Z. Liu, Y. Zhou, Y. Zhu, J. Lian, C. Li, Z. Dou, D. Lian, and J.-Y. Nie, "Information retrieval meets large language models," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1586–1589.

[233] B. Yan, X. Zhang, L. Zhang, L. Zhang, Z. Zhou, D. Miao, and C. Li, "Beyond self-talk: A communication-centric survey of llm-based multi-agent systems," *arXiv preprint arXiv:2502.14321*, 2025.

[234] Z. Zhou, X. Zhang, L. Zhang, J. Liu, S. Wang, Z. Liu, X. Zhang, C. Li, and P. S. Yu, "Finefake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection," *arXiv preprint arXiv:2404.01336*, 2024.

[235] J. Yin, Z. Zeng, M. Li, H. Yan, C. Li, W. Han, J. Zhang, R. Liu, A. Sun, D. Deng *et al.*, "Unleash llms potential for recommendation by coordinating twin-tower dynamic semantic token generator," *arXiv preprint arXiv:2409.09253*, 2024.