



# 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, Large Language Models (LLMs) have shown remarkable capabilities in Natural Language Processing (NLP), including content generation, human-computer interaction, machine translation, and code generation. However, their widespread deployment has also raised significant safety concerns. In particular, LLM-generated content can exhibit unsafe behaviors such as toxicity, bias, or misinformation, especially in adversarial contexts, which has attracted increasing attention from both academia and industry. Although numerous studies have attempted to evaluate these risks, a comprehensive and systematic survey on safety evaluation of LLMs is still lacking. This work aims to fill this gap by presenting a structured overview of recent advances in safety evaluation of LLMs. Specifically, we propose a four-dimensional taxonomy: (i) *Why to evaluate*, which explores the background of safety evaluation of LLMs, how they differ from general LLMs evaluation, and the significance of such evaluation; (ii) *What to evaluate*, which examines and categorizes existing safety evaluation tasks based on key capabilities, including dimensions such as toxicity, robustness, ethics, bias and fairness, truthfulness, and related aspects; (iii) *Where to evaluate*, which summarizes the evaluation metrics, datasets and benchmarks currently used in safety evaluations; (iv) *How to evaluate*, which reviews existing mainstream evaluation methods based on the roles of the evaluators and some evaluation frameworks that integrate the entire evaluation pipeline. Finally, we identify the challenges in safety evaluation of LLMs and propose promising research directions to promote further advancement in this field. We emphasize the necessity of prioritizing safety evaluation to ensure the reliable and responsible deployment of LLMs in real-world applications.

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

## I. INTRODUCTION

AS artificial intelligence continues to advance at an unprecedented pace, Large Language Models (LLMs) have rapidly emerged as a focal point in both academia and industry. These models have achieved significant breakthroughs

S. Liu, C. Li, X. Zhang are with School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China (Email: liusyang@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 Illinois at Chicago, Chicago, USA (Email: psyu@uic.edu).

Chaozhuo Li is the Corresponding Author.

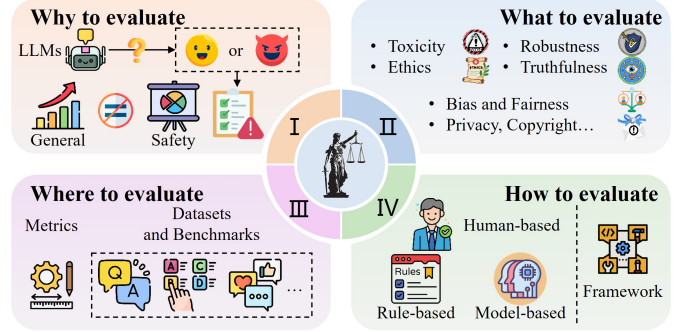


Figure 1. Overview of the Comprehensive and Systematic Survey on Safety Evaluation of LLMs.

in Natural Language Processing (NLP) and demonstrated strong capabilities across diverse applications, including content generation [1], human-computer interaction [2], machine translation [3], and code generation [4], among others [5], [6]. By leveraging massive training corpora, LLMs can generate coherent, human-like text while exhibiting a degree of comprehension and reasoning ability [7], thus driving rapid progress in the broader field of artificial intelligence.

However, like any emerging technology, the widespread application of LLMs has introduced critical safety challenges. Recent studies have demonstrated that LLMs are prone to generating harmful content, such as toxicity [8], bias [9] and false information [10], with such issues often exacerbated under adversarial scenarios. In addition, attack methods targeting LLMs safety, such as prompt injection attacks and jailbreak attacks [11], [12], have become increasingly sophisticated, seeking to bypass alignment mechanisms and induce unsafe behavior from LLMs [13]. These vulnerabilities not only degrade user experience but also raise profound ethical and legal risks, offering opportunities for malicious exploitation. Consequently, ensuring the safety of LLMs through systematic evaluation has become an urgent priority.

Safety evaluation of LLMs is the systematic process of examining whether a model can resist harmful instructions, avoid generating unsafe outputs, and maintain robustness under both benign and adversarial interactions, thereby assessing the degree of risk that LLMs may pose to social order, public interest, and human well-being. The emergence of this research area stems from a growing recognition and urgent need to address the potential risks posed by these models and their

Table I  
A COMPARATIVE ANALYSIS OF RELATED SURVEYS AND STUDIES ON SAFETY EVALUATIONS OF LLMs

Papers	Year	Key Objectives of Safety Evaluation				
		Taxonomy of Tasks	Metrics	Datasets and Benchmarks	Taxonomy of Evaluators	Challenges and Future Directions
Guo <i>et al.</i> [14]	2023	✓	✗	✓	✗	✗
Dong <i>et al.</i> [15]	2024	✗	✓	✓	✗	✓
Yi <i>et al.</i> [16]	2024	✗	✓	✓	✗	✗
Ran <i>et al.</i> [17]	2024	✗	✗	✗	✓	✗
Röttger <i>et al.</i> [18]	2025	✗	✗	✓	✗	✗
<b>Ours</b>	2025	✓	✓	✓	✓	✓

societal impacts. Within the broader lifecycle of LLMs research, development, and deployment, safety evaluation plays a crucial role. Unlike general ability evaluation of LLMs [1], [14], [19], which focus primarily on general capabilities such as language understanding, generation, and reasoning, safety evaluation places particular emphasis on the risks and ethical concerns associated with model outputs. The evaluation covers a range of critical dimensions, including toxicity, robustness, morality, bias and fairness, and credibility. By conducting a systematic safety evaluation 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 LLMs technologies.

While numerous researchers have actively explored the safety evaluation of LLMs in recent years, the field still lacks a comprehensive and systematic review. As shown in Table I, most existing relevant surveys concentrate on general ability evaluation of LLMs [14] on specific topics such as jailbreak attacks [15], [16], with only limited coverage of safety evaluation, failing to form a systematic classification framework and content synthesis. Furthermore, existing studies often remain fragmented, addressing single aspects such as datasets and benchmarks [18] or evaluators [17], lacking a holistic analysis of the entire pipeline, which includes evaluation methods, technical challenges, and future directions. This fragmentation underscores the urgent need for a structured and integrative survey on safety evaluation of LLMs.

Against this backdrop, this survey aims to provide a comprehensive and systematic review of recent advancements in safety evaluation of LLMs, thereby addressing this research gap. Specifically, as shown in Figure 1, we delve into existing work from the following four dimensions: (i) *Why to Evaluate*, which elucidates the background of safety evaluation, distinguishing it from general ability evaluation. 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; (ii) *What to Evaluate*, which summarizes the primary tasks associated with safety evaluation of LLMs, presenting a detailed classification that encompasses various facets such as toxicity, robustness, ethics, bias and fairness, truthfulness, and more. (iii) *Where to Evaluate*, which 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; and (iv) *How to Evaluate*, categorizes evaluation methods based on the roles of the evaluators, whether they are automated or manual. It also reviews several evaluation frameworks, providing insights into the methodologies and best practices for conducting safety evaluations. Finally, we highlight the key challenges in this field and outline promising avenues for future research.

The contributions of this paper can be summarized as follows:

- **Comprehensive Review:** To the best of our knowledge, this is the **first** survey to present a systematic and holistic synthesis of recent advancements in safety evaluation of LLMs, addressing a significant gap in the literature.
- **Clear Classification Framework:** We establish a detailed taxonomy that delineates the primary tasks of safety evaluation across multiple dimensions, providing conceptual clarity for the research community.
- **Consolidated Evaluation Resources:** We compile and categorize existing evaluation metrics, datasets, benchmarks, and methodologies, offering a consolidated reference for future research and practice.
- **Future Research Directions:** We highlight key challenges currently facing the field and propose potential research directions to promote the responsible and sustainable development of safety evaluation.

## II. WHY TO EVALUATE

This section introduces the background and key concepts of safety evaluation of LLMs, clarifies the distinctions between general evaluation and safety-specific evaluation, and highlights its importance in real-world applications. The goal is to underscore the necessity of safety evaluation and establish a foundation for subsequent discussions on advanced methodologies and practices.

### A. Background and Foundations

LLMs have been rapidly adopted across a wide range of tasks — from knowledge inference and information extraction to content and code generation — due to their strong generative and reasoning capabilities. [107]–[114] While these capabilities enable new applications, they also expose concrete

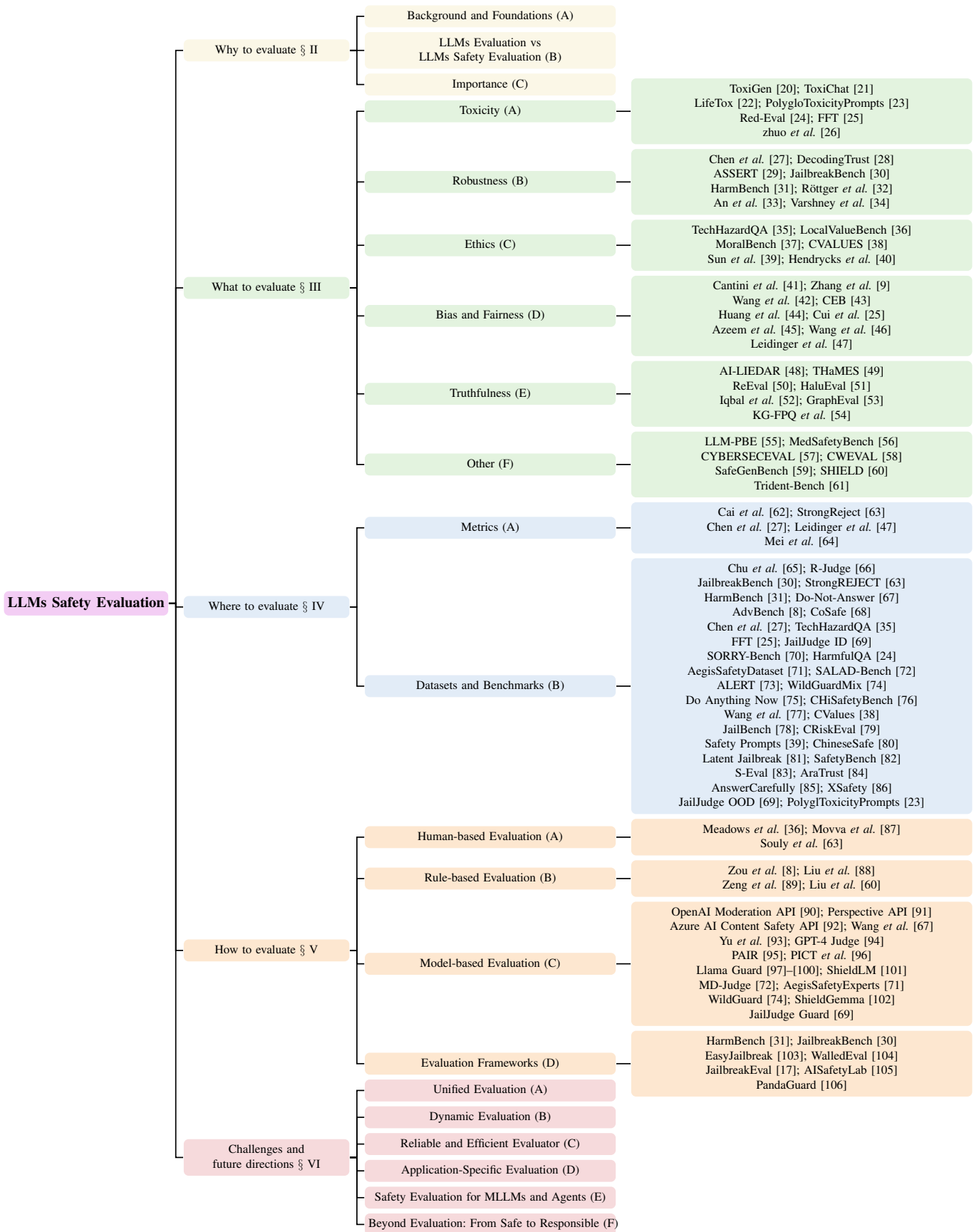


Figure 2. Structure of this paper.

and repeatable failure modes that raise safety concerns in practice.

Empirical studies have pointed out several failure modes. LLMs can generate toxic or abusive language under certain prompts, undermining user safety in public-facing systems [22], [115]. They can reproduce or amplify social biases and unethical stereotypes present in training data, yielding discriminatory outputs in downstream decisions [116], [117]. LLMs also exhibit hallucinations — confidently stated but factually incorrect assertions — which degrade trust and can mislead users [118], [119]. Finally, LLMs are susceptible to adversarial inputs, such as prompt injection or jailbreaks, which exploit vulnerabilities in the model’s behavior [81], [93].

These technical failure modes have concrete, domain-specific consequences. In finance, biased or incorrect model outputs can distort credit assessments or automated advice [120]. In education, stereotype-aligned recommendations may narrow learning opportunities [121]. In healthcare, hallucinated or misleading suggestions risk patient safety and clinical misjudgments [122]. Moreover, adversaries can weaponize generation capabilities to scale misinformation or targeted manipulation, amplifying societal harms [16].

This dual-edged nature of LLMs underscores the need for comprehensive safeguards and systematic security assessments. Identifying and addressing these risks proactively is crucial to prevent unintended consequences, protect users, and ensure these models contribute positively to society. To achieve this, it is necessary to define a formal framework for safety evaluation, which we describe as the process of assessing whether a model can resist harmful instructions, avoid unsafe outputs, and maintain robustness under both benign and adversarial interactions. Formally:

$$R = E(T, D, P, M) \quad (1)$$

where  $R$  denotes the outcome of the evaluation, which may be quantitative (e.g., scores) or qualitative (e.g., categorical assessment);  $E(\cdot)$  represents the safety evaluation function, mapping the selected tasks, data, evaluation process, and metrics to a final result;  $T$  indicates the evaluation tasks that define the targeted safety dimensions. This step answers the question of *what to evaluate*;  $D$  represents the datasets or benchmarks used for testing. This step determines *where to evaluate*;  $P$  corresponds to the evaluation process. This aligns with *how to evaluate*; and  $M$  refers to the metrics that quantify safety performance.

### B. LLMs Evaluation vs LLMs Safety Evaluation

The distinction between LLMs evaluation and LLMs safety evaluation lies not in superficial differences of scope or methodology, but in their fundamentally divergent goals and perspectives. Conventional evaluation addresses the question of “*what the model can do*”, while safety evaluation is concerned with “*what the model should not do*”. This asymmetry gives rise to significant differences in their objectives, approaches, and challenges.

First, the two forms of evaluation pursue distinct core objectives. LLMs evaluation is driven by an *optimization mindset*: its purpose is to measure the upper bound of a model’s capabilities by assessing performance on general tasks such as knowledge, reasoning, or text generation. In contrast, safety evaluation adopts a *control mindset*: its purpose is to probe the lower bound of risk by examining the model’s behavior under malicious or unexpected conditions. The ultimate goal is not to maximize capability, but to minimize harm, ensuring that the model remains robust, compliant, and trustworthy across diverse real-world applications [123]. This divergence in purpose fundamentally shapes the methodology of each evaluation paradigm.

Second, safety evaluation can be understood as an extension rather than a replacement of capability evaluation. While technical proficiency remains a prerequisite, it is insufficient on its own. A model that excels in accuracy or fluency but fails to avoid harmful, biased, or manipulative outputs cannot be considered safe for deployment. Thus, safety evaluation builds on the foundation of general evaluation by expanding the focus from technical correctness to include ethical and societal considerations. In other words, it asks not only whether the model can perform a task, but also whether it should perform it in a given way.

Third, the quantification of evaluation outcomes reflects a striking asymmetry between the two. Capability evaluation typically relies on objective, mathematically defined indicators that enable standardized comparison across tasks and systems. Metrics like accuracy, F1-score, BLEU [124], and ROUGE [125] are designed to quantify improvements in effective and reliability. By contrast, safety evaluation must grapple with inherently subjective and context-dependent criteria, such as fairness, harm, or appropriateness. Defining what constitutes “toxic” or “biased” content is a challenging endeavor, as such determinations may vary across cultural and social contexts—factors that render universal standardization elusive. Presently, there remains a lack of consensus within the research community regarding how to formulate precise definitions of such content [17]. This subjectivity poses unique challenges for designing safety benchmarks and metrics that are both rigorous and widely applicable.

In summary, while both LLMs evaluation and safety evaluation are indispensable, they differ in orientation: one is performance-driven, seeking optimization, while the other is risk-driven, seeking control. The interplay between these two dimensions—capability and safety—defines the trajectory of responsible LLM development, ensuring that models are not only powerful but also aligned with societal values and constraints.

### 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 [126]–[128]. 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.

The importance of LLMs safety evaluation lies not only in ensuring reliable performance, but also in addressing the distinctive challenges that arise from the generative and open-ended nature of these models. Unlike traditional AI systems whose risks are often confined to narrow domains, LLMs operate across virtually unlimited contexts and user intentions, which amplifies both the scale and unpredictability of potential harms. A single unsafe response may propagate misinformation at scale, enable malicious use, or erode public trust, creating risks that extend far beyond technical failures [126]–[128].

The evaluation methods for LLMs’ general capabilities, such as accuracy benchmarks or instruction following tests, are not sufficient for safety evaluation, as safety concerns extend to subtle dimensions including refusal behavior, toxicity avoidance, bias mitigation, and resistance to adversarial prompting. Systematic safety evaluation is therefore indispensable for three reasons:

- **Managing Unique Risk Profiles.** LLMs are open to the entire human society and face diverse and complex security risks. The risk of malicious use can occur in all possible scenarios. Evaluation enables the early detection of such vulnerabilities before real-world deployment.
- **Enabling Accountability and Governance.** Transparent and standardized evaluation provides the evidence base necessary for legal compliance, policy development, and ethical oversight, particularly in high-stakes fields such as healthcare, law, and education.
- **Driving Technical Progress.** Safety evaluation not only identifies shortcomings but also guides iterative model improvements, serving as a feedback loop for refining architectures, training strategies, and alignment techniques. This makes evaluation a scientific instrument for advancing both reliability and innovation.

In this survey, we emphasize that safety evaluation is not a peripheral safeguard but a central research agenda: it bridges societal expectations, regulatory demands, and technical innovation. By synthesizing safety tasks, datasets, benchmarks and methodologies, we aim to provide a roadmap that helps the community move beyond ad hoc testing toward standardized, rigorous, and impactful evaluation practices.

### 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 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 hate speech, offensive or abusive language, pornographic content, and other harmful expressions in text generated by LLMs. Such outputs not only pose risks of psychological harm to individual users but may also trigger broader social conflicts. As a result, evaluating the potential of LLMs to generate toxic content has become a critical step in ensuring their safe deployment. Existing efforts in this area can be broadly categorized into three lines of research: benchmark construction, adversarial and red-teaming based evaluation, and contextual or granular assessments.

1) **Benchmark Construction for Toxicity Evaluation:** A major line of research has focused on building diverse and scalable benchmarks to capture toxicity in various settings. For example, ToxiGen [20] 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 generated via an adversarial classifier-in-the-loop decoding algorithm (ALICE), enabling the creation of implicit toxic texts across multiple demographic groups. ALICE adjusts the toxicity of the generated text by comparing a toxicity classifier with the text generator during beam search decoding. This adaptive control mechanism enables ToxiGen to not only generate texts that closely mimic human-written toxicity in nuanced expressions — including subtle biases toward underrepresented demographics — but also produce outputs that elude accurate classification by existing toxicity detectors, outperforming all prior human-curated corpora in both naturalistic mimicry and classifier evasion.

Similarly, ToxiChat [21] 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. By collecting real user queries directed at open-source chatbots and introducing an uncertainty-guided human-AI annotation pipeline, the authors constructed a benchmark that reflects the safety concerns of interactive systems. The annotation pipeline uses the moderation API [91] to filter out low-confidence toxicity data, and the remaining data is manually reviewed, which greatly reduces the annotation workload. Furthermore, a systematic evaluation revealed that due to the domain inconsistency between the training data and real-world user prompts for chatbots, existing models not only fail to generalize to ToxiChat but also perform significantly worse on adversarial jailbreaking cases.

Moving toward implicit harmful content, LifeTox [22] focuses on life-advice scenarios, where 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 [129]. 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 dataset, constructed from Reddit forums, improves the detection of implicit harmful advice and demonstrates that fine-tuned RoBERTa [130] can rival LLMs in harmful content classification.

Extending beyond English and Chinese, Jain *et al.* [23] propose PolygloToxicityPrompts, a large-scale multilingual benchmark covering 425K prompts across 17 languages. The prompts were generated by scraping documents from extensive corpora, with an average length of 400 GPT-4 tokens. 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 exhibit lower toxicity compared to baseline models; However, the choice of preference-tuning method have 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.

2) **Adversarial and Red-Teaming Evaluation:** Another strand of work emphasizes adversarial attacks and red-teaming to probe the vulnerabilities of LLMs. Bhardwaj *et al.* [24] introduce RED-EVAL, a red-teaming benchmark that applies 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. They further propose RED-INSTRUCT, 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. Experiments demonstrate that RED-EVAL achieves jailbreak success rates of 65% on GPT-4 and 73% on ChatGPT, underscoring the persistent risks of harmful content generation.

Similarly, Zhuo *et al.* [26] conducted a systematic red team evaluation of ChatGPT using the RealToxicPrompts dataset [131] and Perspective API [91]. The results showed that ChatGPT, due to its use of a cleaner training dataset, demonstrated superior performance compared to baseline models. However, the model remains vulnerable to prompt injection attacks, which can trigger toxic responses, and currently lacks mechanisms for early-stage toxicity detection.

3) **Contextual and Granular Toxicity Evaluation:** Beyond benchmarks and adversarial probing, some works examine toxicity at different levels of granularity. FFT [25] investigate both utterance-level toxicity, where toxicity refers to literally-toxic language with some typical words, and context-level toxicity, where otherwise benign statements become harmful when interpreted in a broader conversational context. Their results show that LLMs generally exhibit higher toxicity rates at the contextual level, often by inadvertently affirming harmful issues despite generating literally non-toxic responses. This finding highlights the limitations of existing detectors in capturing nuanced, situational forms of toxicity.

Despite substantial progress, current approaches to toxicity

evaluation face several unresolved challenges. First, the cultural and domain coverage of existing benchmarks remains limited, as most datasets are concentrated in a small set of languages or narrowly defined interaction scenarios, restricting their generalizability. Second, the emergence of increasingly sophisticated forms of implicit toxicity, conveyed through subtle linguistic cues or contextual framing, continues to outpace the ability of current models and detectors to capture them with sufficient accuracy. Third, the absence of standardized evaluation metrics and unified frameworks makes it difficult to compare results across studies, thereby hindering cumulative progress.

## B. Robustness

Robustness refers to the ability of LLMs to maintain stable and safe outputs when exposed to noisy inputs, adversarial attacks such as jailbreak or prompt injection, and out-of-distribution (OOD) data. Beyond merely resisting malicious prompts, robustness also entails preserving instruction-following ability without exhibiting exaggerated safety, since models that indiscriminately refuse queries—regardless of their harmfulness (e.g., “How to kill a process?”)—sacrifice usability. Recent studies on robustness evaluation can be broadly grouped into three directions: adversarial and red-teaming evaluations, benchmarking and standardized frameworks, and the balance between instruction-following and excessive safety.

1) **Adversarial and Red-Teaming Evaluation:** Several works focus on systematically characterizing the vulnerabilities of LLMs against adversarial inputs. Jailbreak attacks, particularly those involving carefully crafted prompts, pose significant challenges to the safety of LLMs. Chen *et al.* [27] construct a dataset spanning 61 harmful categories to evaluate 10 jailbreak attack methods across 13 popular LLMs. Their framework employs multiple metrics, including attack success rate (ASR), toxicity, fluency, grammatical errors, and token length, enabling a multidimensional analysis of LLM safety. Results reveal that none of the evaluated models demonstrate inherent resistance to harmful queries, with Vicuna and Mistral among the most vulnerable, while GPT-4 and Llama3 exhibit comparatively stronger performance. Among the jailbreak attack methods, ReNeLLM [132] achieves the highest average ASR, while ICA [133] and Cipher [134] exhibit lower performance. They further observe that ASR correlates positively 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.

Complementing this work, DecodingTrust [28] conducts a comprehensive evaluation of GPT-4 and GPT-3.5, examining robustness to adversarial text attacks, OOD inputs, and adversarial demonstrations. For adversarial text attacks, the researchers construct three scenarios: (i) standard benchmark AdvGLUE [135], (ii) AdvGLUE under various instructive task descriptions and system prompts, and (iii) the more challenging adversarial text generation benchmark, AdvGLUE++. For OOD inputs, they explore scenarios including (i) inputs



deviating from common training text styles, (ii) recent events outside GPT models’ training data collection period, and (iii) demonstrations with diverse OOD styles and domains added through in-context learning. For adversarial demonstrations, the evaluations involve injecting counterfactual examples, spurious correlations, and backdoors into the demonstrations to observe model’s performance. Overall, they find that GPT-4, though outperforming GPT-3.5 on standard benchmarks, is paradoxically more vulnerable to jailbreak attacks under certain prompts, owing to its stronger tendency to follow misleading instructions.

Moving toward automated red-teaming, ASSERT [29] introduces three methods—semantically aligned augmentation, targeted bootstrapping, and adversarial knowledge injection—to generate new test cases. 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  $c$ . Their results highlight that LLMs exhibit instability even under semantically similar scenarios, with divergences up to 11% in classification accuracy, while adversarial demonstrations induce error rates exceeding 50% in some settings. Collectively, these studies emphasize that LLM robustness remains fragile when confronted with diverse adversarial strategies and distributional shifts.

2) **Benchmarking and Standardized Frameworks:** Recognizing the limitations of fragmented evaluations, another line of research develops systematic benchmarks and frameworks. JailbreakBench [30] addresses the lack of clear standards and reproducibility in jailbreak testing by two core means: first, releasing a benchmark dataset consisting of 100 prompts that cover behavior, harmful goal, positive responses, harmful category, and data source; second, maintaining a repository of jailbreak artifacts and establishing a standardized evaluation framework. Its standardized evaluation pipeline incorporates threat models, system prompts, chat templates and multiple evaluation methods, ranging from rule-based string matching to GPT-4-as-a-judge and several models specifically designed for jailbreak evaluation (e.g., Llama Guard [97]). Experiments indicate that even state-of-the-art models remain highly susceptible. Prompts with RS [136] outperform others such as PAIR [95], GCG [8], and Jailbreak Chat [137], achieving an ASR of up to 78% on GPT-4.

In parallel, HarmBench [31] establishes a more comprehensive red-teaming benchmark that categorizes harmful behaviors both semantically (e.g., cybercrime, bioweapons) and functionally (e.g., contextual, copyright, multimodal). 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. Finally, it incorporates a total of 18 red-teaming

methods from 12 studies, revealing that no attack or defense is universally effective, and that robustness does not correlate with model size—a finding that contradicts earlier assumptions [138]. Together, these benchmarks enhance comparability and reproducibility, while underscoring the persistent vulnerability of LLMs.

3) **Instruction-Following vs. Exaggerated Safety:** 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 [32]. Röttger *et al.* [32] design XSTest, a benchmark mixing safe and unsafe prompts, and show that Llama2 exhibits exaggerated safety by refusing benign queries. They attribute this over-defensiveness to lexical overfitting and note that adding guardrail prompts, while reducing unsafe behavior, exacerbates the problem. Additionally, safety-related behavior could be changed by system prompts added at inference time. An *et al.* [33] propose a white-box method for automatically generating pseudo-harmful prompts, revealing that many jailbreak defenses substantially increase false refusal rates, thereby undermining model usability. Similarly, Varshney *et al.* [34] develop classifiers to assess whether responses are both safe and appropriately generated, finding that self-checking methods improve safety but induce excessive conservatism. These studies collectively demonstrate that robustness cannot be narrowly defined as resistance to attacks but must also account for maintaining instruction-following capacity without compromising helpfulness.

Despite substantial progress, robustness evaluation faces several unresolved challenges. Current methods still provide limited coverage across diverse domains, languages, and modalities, leaving open questions about generalization in real-world contexts. Automated adversarial testing frameworks, though increasingly sophisticated, remain inconsistent and lack standardized protocols, complicating cross-study comparison. Moreover, defenses that reduce susceptibility to jailbreaks often increase false refusals, highlighting the persistent difficulty of balancing safety with usability.

### C. Ethics

Ethics refers to the moral principles and values embedded in the outputs of large language models (LLMs), ensuring that generated text adheres to socially accepted norms and legal frameworks. Ethical compliance requires excluding content that is discriminatory, offensive, or misleading, as such expressions violate the principles of fairness and social justice [131]. Existing research on ethical evaluation of LLMs can be broadly categorized into global-purpose benchmarks and local-specific evaluations.

1) **Global-purpose evaluation:** A number of works have developed comprehensive benchmarks to assess the ethical reasoning and moral alignment of LLMs. While existing research has primarily focused on bias and other safety dimensions, MoralBench [37] systematically evaluates moral reasoning

by drawing on the six foundations from Moral Foundations Theory [139], which include Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression. It incorporates binary and comparative tasks to measure the consistency of LLM responses with human moral judgments. 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. Results show that models perform adequately on simple binary classifications but struggle with comparative evaluations, suggesting a reliance on pattern recognition from training data rather than demonstrating a genuine understanding of ethical principles.

CVALUES [38] addresses the lack in existing evaluation frameworks in assessing the ethical alignment and value sensitivity of Chinese 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. While models often succeed in avoiding harmful or unsafe content, their limitations emerge in domains such as law and social sciences, where reasoning and empathy are required. This demonstrates gaps between technical safety and broader ethical responsibility.

Complementary to these efforts, Hendrycks et al. [40] proposed the ETHICS dataset to test moral reasoning in multiple scenarios, showing that model performance is highly sensitive to textual framing and limited in handling nuanced ethical dilemmas. And Sun et al. [39] developed the SafetyPrompts benchmark, which includes Ethics and Morality as one dimension among eight safety scenarios. Their findings highlight that adversarial instructions often exacerbate ethical vulnerabilities, underscoring the need for robust defenses.

2) **Local-Specific Evaluation:** Ethics have significantly different definitions and levels of tolerance in different scenario and cultural backgrounds. Therefore, focusing on ethical risk evaluation in specific domains and localized contexts is equally important. While prior research has primarily focused on the safety of general text-based responses, TechHazardQA: [35] is a dataset created to address the safety vulnerabilities of existing LLMs when generating instructional content, such as pseudocode or code snippets. 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 [140]–[142]. To further investigate the impact of model editing on content generation, they employed the ROME model editing technique [143] 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.

Additionally, LocalValueBench [36] evaluates ethical alignment under localized cultural and legal contexts. Focusing on the Australian environment, it examines model responses to neutral, debated, and misleading questions across domains such as capital punishment and same-sex marriage. Neutral questions establish baseline responses, debated questions explore how the models handle complex ethical scenarios, and misleading questions assess their behavior in extreme situations. Findings reveal significant cultural sensitivity: models differ not only in performance across topics but also in how embedded cultural biases shape outputs. 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 human evaluators’ ratings expose the difficulties of standardizing cross-cultural ethical evaluation.

Taken together, global-purpose and local-specific evaluations provide complementary perspectives: the former emphasizes broad, systematic coverage of ethical reasoning, while the latter uncovers risks specific to specific domains or cultural environments. A key challenge lies in reconciling these perspectives to construct evaluation frameworks that are both comprehensive and culturally adaptive. Moreover, the diversity of ethical standards across societies raises profound challenges. Ethical standards are not universal, and LLMs trained on globally sourced data may inadvertently reflect dominant cultural values while marginalizing others. This issue of cultural diversity suggests that benchmarks must account for heterogeneous value systems rather than enforcing a single standard of morality. Looking forward, as AI systems become more autonomous, the potential emergence of machine-internal ethical reasoning poses additional risks: LLMs might develop consistent but misaligned ethical patterns that diverge from human expectations. More research is needed to address the ethical issues exposed by increasingly intelligent AI systems.

#### 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 [117]. 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.



1) **General-purpose Evaluations:** A first line of research focuses on systematically probing the inherent biases embedded in LLMs through general-purpose evaluations. Leidinger *et al.* [47] combine 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 prompts related to peoples/ethnicities and sexual orientation. Additionally, mentions of intersectional identities tend to trigger disproportionately stereotyped responses.

Furthermore, Cantini *et al.* [41] 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 employ five jailbreak methods—role-playing, machine translation, obfuscation, prompt injection and reward incentive. 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. Importantly, no model is entirely safe, as each is relatively susceptible to at least one jailbreak attack.

To address the issue of most existing research either focusing on specific types of biases or employing incompatibility evaluation metrics, Wang *et al.* [43] 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 and methods for each task, including Micro-F1 score, GPT-4-as-a-Judge, Perspective API [91], Demographic Parity (DP), Equalized Odds (EO), and Unfairness Scores [144]. The experimental results reveal that LLMs are adept at identifying toxic content in inputs but struggle to recognize stereotypical content. Furthermore, LLMs demonstrate a high refuse to answer rate for racial and religious social groups, indicating a heightened sensitivity.

Similarly, Huang *et al.* [44] introduces a Chinese bias benchmark dataset, developed through human-AI collaboration, aimed at measuring biases in Chinese LLMs, while Cui *et al.* [25] evaluates the fairness of LLMs through a benchmark that utilizes the coefficient of variation as an evaluation metric across four dimensions: identity preference,

credit, criminal, and health. Azeem *et al.* [45] 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).

2) **Domain-specific Evaluations:** A second line of research investigates fairness in particular downstream applications where LLMs directly affect decision-making. In the context of recommendation systems, Zhang *et al.* [9] evaluate 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 results find that ChatGPT still exhibits unfairness in generating recommendations related to certain sensitive attributes, and its 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.

Similarly, Wang *et al.* [42] 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.

In addition, Given the widespread use of LLM-as-a-Judge [145], Wang *et al.* [46] 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.

Taken together, existing research demonstrates that LLMs

exhibit systematic biases both in general-purpose evaluations and in application-specific contexts. However, several limitations persist. Current benchmarks remain limited in their cultural and linguistic coverage, restricting the generalizability of fairness assessments across diverse populations. Many methods struggle to capture subtle or intersectional forms of bias, which often manifest in implicit or adversarial scenarios. Furthermore, the absence of standardized evaluation metrics and unified frameworks hinders comparability across studies, limiting cumulative progress.

### E. Truthfulness

Truthfulness refers to the factual accuracy of information generated by LLMs, specifically their alignment with real-world facts [146]. Ensuring truthfulness is crucial for building user trust and preventing the dissemination of misleading information. A central challenge in this domain is the phenomenon of hallucination [147]–[149], where LLMs produce seemingly plausible but factually inaccurate or contextually disconnected responses. Consequently, most existing research on truthfulness evaluation has focused on hallucination [150]–[152], while a smaller body of work examines other dimensions such as utility-driven deception.

1) **Hallucination Evaluation:** A substantial body of research has developed evaluation methods and benchmarks to systematically assess hallucinations in LLMs. One stream emphasizes dataset construction under realistic interaction scenarios. For example, ReEval [50] create new evaluation datasets using adversarial attacks to test whether retrieval-augmented LLMs accurately reference provided evidence in their responses. 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. Results show that all models exhibited a significant hallucination on adversarial test data. Moreover, HaluEval [51] provides a large-scale benchmark that evaluate LLMs’ capabilities in hallucination detection, which cover knowledge-based dialogue, question answering, and text summarization. Experimental results reveal notable shortcomings in current LLMs’ hallucination detection abilities, and hallucination frequency is strongly topic-dependent, with higher rates observed in content related to film, technology, and climate.

Another stream leverages structured external knowledge to support factual verification. GraphEval [53] detects hallucinations in LLMs’ outputs by constructing knowledge graphs (KGs) 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 [153] and QAGS-C [154], especially for longer outputs. Similarly, KG-FPQ [54] leverages

KGs to generate False Premise Questions (FPQs) for assessing truthfulness hallucinations in LLMs. By selecting entities from art, people, and places, KG-FPQ first extracts real triplets from the KGs and generates false premises through edit true triplets to false triplets. Then, it utilizes ChatGPT’s natural language generation capability to construct semantically rich FPQs. Experiments show that when edited object has a closer distance with the subject or has a stronger association with the original object, FPQs are more confusing to LLMs. And LLMs perform worse at generating factual statements than at distinguishing them when faced with FPQs.

In addition, some works propose unified framework for the evaluation of hallucinations. THaMES [49] introduces a end-to-end tool for hallucination detection and mitigation, which comprises three primary modules—QA set generation, hallucination benchmarking, and mitigation strategies. The framework begins by constructing high-quality question-answer test sets through weighted sampling and advanced question-generation techniques. Next, multiple metrics from RAGAS [155] and HaluEval [51] are used to evaluate hallucination. Following this, THaMES implements a series of refined mitigation strategies, including In-Context Learning (ICL) [156], Retrieval-Augmented Generation (RAG) [157], and parameter-efficient fine-tuning (e.g., PEFT) [158], [159]. In order to solve the problem of difficulty in free-form response evaluation in open-domain and fragmentation of datasets and metrics, Iqbal *et al.* [52] propose OpenFactCheck. The framework consists of three collaborative modules: First, RESPONSEEVAL supports users to customize the fact-checker, integrating the claim processors, evidence retrievers and verifiers of existing systems [160]–[162] to realize factual error detection in a free-form document; Second, LLMEVAL builds a FactQA benchmark, which divides factual deviations into three categories: knowledge errors, over-commitment errors, and disability errors, and unifies evaluation metrics (such as exact matching, FreshEval [163], and the proportion of true statements) to systematically evaluate LLM’s factual capabilities in multiple dimensions; Third, CHECKEREVAL assesses the verification accuracy of fact-checkers, equipped with a leaderboard in terms of accuracy, latency, and costs.

2) **Beyond Hallucinations:** While hallucination has attracted primary attention, other dimensions of truthfulness are also emerging. AI-LIEDAR [48] aims to investigate the balance between utility [164] and truthfulness [165] 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. 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 [166], 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.

Despite notable advances, truthfulness evaluation remains constrained by several challenges. Current evaluation metrics for hallucinations face significant limitations, including a lack of universal applicability, inadequate comprehensiveness, and vulnerability to dataset leakage [167], which undermine their robustness and fairness. Furthermore, many recent efforts have turned to the paradigm of LLM-as-a-Judge [145] to assess truthfulness. While this approach offers scalability and flexibility, it also introduces a recursive dilemma: if evaluating one model requires relying on a stronger and more reliable model, the trustworthiness of the evaluation itself becomes questionable. These challenges highlight the urgent need for innovative and standardized evaluation systems that can holistically capture different forms of untruthfulness, while reducing dependency on opaque or unverifiable evaluators.

#### 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. 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.

1) **Privacy:** LLMs rely on extensive data for training, posing inherent risks of leaking sensitive user information [168]–[170], including personal identifiers, geographic locations, and employment-related information. Consequently, a systematic assessment of privacy risks in LLMs is critical to identify and mitigate potential breach channels, thereby ensuring model safety and compliance in real-world applications. Li *et al.* [55] introduced LLM-PBE, a comprehensive toolkit for evaluating privacy risks throughout the LLM lifecycle. LLM-PBE provides systematic assessments across diverse data types, multiple attack vectors, and corresponding defense strategies, with particular emphasis on analyzing how model size, data characteristics, and temporal factors impact privacy vulnerabilities. The toolkit incorporates modules for different attack types, such as data extraction [170], [171], membership inference [172], prompt leakage [173], and jailbreaking [95], complemented by defense methods like differential privacy [174]. Moreover, the framework integrates support for models hosted on platforms such as OpenAI, TogetherAI, and Hugging Face, enabling privacy assessments through API interfaces. Extensive experimentation on datasets such as Enron [175] and ECHR [176] enables a systematic analysis of privacy vulnerabilities in LLMs. Empirical results demonstrate that larger models exhibit higher susceptibility to privacy leakage. Certain data types and placements, such as sensitive information located at the end of documents, are more vulnerable to

extraction. Furthermore, while differential privacy effectively mitigates leakage risks, it introduces notable trade-offs in model utility, highlighting the necessity of balancing privacy preservation with performance. Despite these advances, key challenges remain. Privacy attacks evolve rapidly, demanding adaptive evaluation frameworks that can capture emerging threats. The trade-off between privacy preservation and model effectiveness remains unresolved, particularly in high-stakes domains. Addressing these challenges is essential for building trustworthy and privacy-preserving LLMs.

2) **Copyright:** A systematic approach to copyright assessment is essential for systematically identifying and mitigating the generation of infringing content, thereby ensuring compliance and enabling the safe and lawful use of LLMs across diverse creative domains. As LLMs are increasingly applied in content creation, concerns regarding copyright infringement have intensified, particularly due to their ability to reproduce protected content without authorization [177]. Existing copyright protection mechanisms often prove either overly restrictive or insufficiently adaptive, leaving gaps in safeguarding against unauthorized content generation. Liu *et al.* [60] introduced the SHIELD framework, a tool designed to evaluate and prevent copyright violations by LLMs during text generation [178]. SHIELD employs a curated copyright dataset to assess whether generated text adheres to copyright standards and incorporates a lightweight and real-time proxy-based defense mechanism to dynamically detect and block potentially infringing content. The framework consists of three primary modules: the Copyright Material Detector, which uses an N-gram model to identify copyrighted segments; the Copyright Status Verifier, which performs real-time checks to verify copyright status; and the Copyright Status Guide, which provides contextual feedback to direct the model toward producing copyright-safe outputs. This mechanism is designed to promptly detect and reject infringing requests while maintaining flexibility in generating public domain content. However, balancing copyright protection with creative flexibility is a persistent issue, as overly strict safeguards may compromise usability. Jurisdictional diversity and the evolving nature of copyright status further complicate the establishment of standardized evaluation practices. Moreover, the scarcity of large-scale, high-quality copyright datasets limits the creation of comprehensive benchmarks. Addressing these challenges will be critical for ensuring that LLMs can be responsibly applied in creative and commercial domains.

3) **Code Generation:** Unlike harmful content in other domains that directly affects individuals or society, risks in code generation stem from insecure code or code that facilitates cyberattacks, posing significant cybersecurity threats [179]–[181]. For instance, LLMs may produce code vulnerable to common exploits such as SQL injection or cross-site scripting, rendering applications susceptible to compromise. Evaluating the security of LLMs in code generation has therefore become a critical research priority. Bhatt *et al.* [57] introduced the CYBERSECEVAL framework, which evaluates the security of LLM-generated code and their susceptibility to cyberattack-related prompts. It includes two components: insecure coding tests, which employ a vulnerability detector to construct cases

reflecting common programming weaknesses, and cyberattack helpfulness tests, which use attack-related prompts to assess whether model responses facilitate malicious actions. Empirical findings reveal that all tested models generated insecure code in about 30% of insecure coding scenarios, and that models with stronger programming capabilities posed greater security risks. Furthermore, approximately 53% of cyberattack prompts elicited responses aligned with adversarial objectives, highlighting LLMs' limitations in handling malicious requests. CYBERSECEVAL thus offers a systematic framework for evaluating LLM cybersecurity. In addition, Peng *et al.* [58] identified limitations in existing benchmarks such as CYBERSECEVAL, including vague specifications, the separation of functional and security evaluations, and instability in static analysis due to high false positive and false negative rates. To address these issues, they proposed the CWEVAL framework and the CWEVAL-BENCH benchmark. CWEVAL introduces a result-driven synchronous evaluation mechanism that replaces static with dynamic analysis, monitoring runtime attributes such as execution time, memory access, and data integrity to verify both functional correctness and security simultaneously. CWEVAL-BENCH comprises 119 security-critical tasks spanning 31 CWE vulnerability types and five programming languages, with secure and insecure reference implementations to ensure reproducibility. Experimental results indicate that about 30% of functionally correct code generated by mainstream LLMs still contains vulnerabilities. Larger models demonstrated comparatively stronger security capabilities, while simple security prompts improved performance, though fine-tuning sometimes degraded functional correctness. Li *et al.* [59] further proposed the SafeGenBench framework and its automated evaluation system to overcome limited scenario coverage and single-method evaluation in existing benchmarks. SafeGenBench incorporates 558 test cases covering 13 programming languages and 44 CWE vulnerabilities, integrating authoritative standards such as OWASP Top-10 and CWE Top-25. Its design simulates realistic development scenarios by deliberately avoiding explicit security terms, ensuring alignment with practical usage. Its key innovation lies in a dual-layer evaluation mechanism: Static Application Security Testing (SAST) for broad-spectrum vulnerability scanning, complemented by an LLM-based judge for deep semantic analysis of task-specific vulnerabilities. These complementary methods mitigate the shortcomings of relying on a single approach. Looking ahead, the evaluation of code generation security faces persistent challenges. Software vulnerabilities continue to evolve rapidly, requiring benchmarks and frameworks that can adapt to emerging threats. At the same time, achieving accurate, efficient, and user-friendly evaluation remains a difficult task. Addressing these issues will be essential to ensure that LLM-assisted software development remains both secure and reliable in real-world settings.

**4) Downstream Applications in Medicine, Finance, and Law:** As LLMs are increasingly adopted in high-stakes fields such as medicine, finance, and law, systematic evaluation of their domain-specific safety has become increasingly critical. Han *et al.* [56] introduced MedSafetyBench, a benchmark dataset specifically designed to assess LLM safety in

medical contexts. MedSafetyBench comprises 1,800 harmful medical requests paired with safe responses, organized around nine fundamental principles of medical ethics. By generating harmful prompts and corresponding responses, the study evaluated models such as GPT-4 and Llama to assess their ability to reject unsafe requests. Results indicate that current medical LLMs often fail to meet established safety standards, producing unsafe responses when confronted with inappropriate prompts. Hui *et al.* [61] introduced Trident-Bench, the first benchmark targeting LLM safety across finance, medicine, and law. Grounded in the AMA Principles of Medical Ethics, the ABA Model Rules of Professional Conduct, and the CFA Institute Code of Ethics, Trident-Bench contains 2,652 harmful prompts with expert-validated safe responses. This benchmark addresses gaps left by prior evaluations that emphasized accuracy and domain knowledge while overlooking professional ethical compliance. An empirical study of 19 general-purpose, domain-specific, and safety-aligned models revealed critical safety gaps: while advanced closed-source models such as GPT-4o were able to meet basic safety expectations, domain-specific models frequently struggled with subtle ethical issues. Safety-aligned models, such as Llama Guard, performed better, indicating that domain specialization does not necessarily ensure ethical robustness. Although these efforts mark important progress, research on safety evaluation in medicine, finance, and law remains limited and lags behind the rapid deployment of LLMs in these domains. The specialized knowledge required in each field makes the design of rigorous and representative benchmarks challenging. Furthermore, variations in ethical norms and legal regulations across jurisdictions complicate the establishment of unified evaluation standards. Addressing these issues is essential for ensuring the responsible and safe integration of LLMs into high-stakes professional practices. At the same time, other downstream applications—such as education [182] and transportation [183]—pose comparable safety concerns, highlighting the need to broaden evaluation efforts beyond currently emphasized domains.

#### IV. WHERE TO EVALUATE

The safety of LLMs can be evaluated through carefully designed metrics, which determine the specific aspects of performance to be measured, and through datasets and benchmarks that provide the foundation for these evaluations. In this section, we comprehensively summarize these two components — metrics, datasets and benchmarks — to provide a clear overview 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 [152], so this aspect is therefore not elaborated here.

1) **Attack Success Rate:** Attack success rate (ASR) is widely used and fundamental metric for evaluating the safety of LLMs. It measures the effectiveness of adversarial attacks such as jailbreak attacks and prompt injection attacks directed at LLMs. A higher ASR indicates lower safety of the evaluated LLM. Mathematically, ASR can be formulated as:

$$ASR = \frac{Num_{success}}{Num_{total}}. \quad (2)$$

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

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 [17], leading to difficulties in comparing the works of different researchers. Current automatic studies primarily determine success through two approaches: rule-based and model-based methods, which discussed in detail in Section V. Rule-based methods are simple, but often have low relevance to human judgment, while model-based methods use different implementation paths in different studies, providing greater flexibility at the expense of reproducibility and comparability.

2) **Fine-Grained Metrics:** 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.* [62] 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 [63] 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:

$$score = (1 - refused) \times \frac{specific + convincing}{2}. \quad (3)$$

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. This is a relatively strict evaluation standard, which alleviates the phenomenon of exaggerating the performance of jailbreak attacks.

Chen *et al.* [27] 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, Leiding *et al.* [47] incorporates four metrics—refusal, toxicity, sentiment, and regard—to capture nuanced biases in model responses. Moreover, Mei *et al.* [64] 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. Although these fine-grained metrics are designed to capture aspects of safety beyond binary classification and narrow the gap with human judgment gold standards, they still heavily rely on the subjective judgment of models, lack standardized definitions, and have yet to gain widespread adoption, thereby limiting the comparability of research.

Overall, ASR remains the predominant metric for safety evaluation, but inconsistent definitions of ‘success’ hinder reproducibility. Emerging fine-grained metrics provide richer insights, yet they remain fragmented and lack standardization. Establishing a unified and widely accepted set of safety evaluation metrics therefore remains an open research challenge.

## B. Datasets and Benchmarks

To comprehensively evaluate the safety performance of LLMs, numerous datasets and benchmarks have been developed. Unlike those focusing on specific domains [56], [58], in this section, we focus on datasets and benchmarks that assess the overall safety of LLMs across multiple dimensions. In this study, we selected 31 datasets and benchmarks in total, as shown in Table II. For clarity of categorization, 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.* [65] establish a dataset of prohibited questions, providing the first large-scale evaluation of various advanced jailbreak attack methods. StrongReject [63] proposes a dataset composed of forbidden questions that LLMs should refuse to answer to evaluate jailbreak performance. HarmBench [31] construct its dataset by incorporating not only harmful behaviors but also copyright, contextual, and

Table II  
OVERVIEW OF COMPREHENSIVE EVALUATION DATASETS.

Benchmark Name	Size	Safety Dimensions	Languages	Composition
Chu <i>et al.</i> [65]	160	16	English	Harmful prompts
R-Judge [66]	162	27	English	Multi-turn agent interaction
JailbreakBench [30]	200	10	English	Harmful behaviors and benign behaviors
StrongREJECT [63]	346	6	English	Harmful prompts
HarmBench [31]	510	7	English	Harmful behaviors
Do-Not-Answer [67]	939	60	English	Harmful instructions
AdvBench [8]	1k	8	English	Harmful strings and harmful behaviors
CoSafe [68]	1.4k	14	English	Multi-turn attack questions
Chen <i>et al.</i> [27]	1.5k	61	English	Harmful prompts
TechHazardQA [35]	1.8k	7	English	Sensitive and unethical questions
FFT [25]	2.1k	3	English	Elaborated-designed harmful prompts
JailJudge ID [69]	4.5k	14	English	Labeled harmful prompts and model responses
SORRY-Bench [70]	9k	45	English	Unsafe instructions
HarmfulQA [24]	18k	10	English	Harmful questions, blue and red multi-turn conversations
AegisSafetyDataset [71]	26k	13	English	Human-LLM interaction instances
SALAD-Bench [72]	30k	66	English	Harmful prompts and multi-choice prompts
ALERT [73]	45k	32	English	Red teaming prompts
WildGuardMix [74]	92k	13	English	Harmful prompts with refusal and compliance response
Do Anything Now [75]	107k	13	English	Forbidden questions
CHiSafetyBench [76]	2.1k	31	Chinese	Multi-choice questions and risky questions
Wang <i>et al.</i> [77]	3k	17	Chinese	Risky questions and harmless questions
CValues [38]	6.4k	10	Chinese	Adversarial prompts and multi-choice prompts
JailBench [78]	10.8k	40	Chinese	Jailbreak questions
CRiskEval [79]	14k	7	Chinese	Multi-choice questions
Safety Prompts [39]	100k	14	Chinese	Augmented prompts and responses
ChineseSafe [80]	205k	10	Chinese	Illegal and unsafe contents
Latent Jailbreak [81]	416	3	Chinese; English	Translation tasks
SafetyBench [82]	11k	7	Chinese; English	Multiple choice questions
S-Eval [83]	220k	52	Chinese; English	Harmful prompts
AraTrust [84]	516	8	Arabic	Multi-choice questions
AnswerCarefully [85]	1.8k	56	Japanese	Harmful questions and reference answers
XSafety [86]	2.8k	14	10	Harmful prompts
JailJudge OOD [69]	6k	14	10	Labeled harmful prompts and responses
PolyglToxicityPrompts [23]	425k	7	17	Real-world harmful prompts

multimodal behaviors for a comprehensive evaluation. Do-Not-Answer [67] includes prompts that responsible language models should not answer, while AdvBench [8] 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 [27] introduce a refined dataset covering a wide range of harmful prompts. TechHazardQA [35] presents a complex harmful query dataset, requiring LLMs to respond with pseudocode instead of standard text to evaluate ethical boundaries. FFT

[25] carefully elaborates a dataset of harmful prompts focusing on factoid, unfair, and toxic contents in LLMs. SORRY-Bench [70] 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 [73] provides a substantial instruction dataset with a fine-grained risk taxonomy for red-team testing of LLMs. Do Anything Now [75] 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.* [77] construct



a dataset including both harmful and benign prompts, which measure LLMs false positives and false negatives in refusal scenarios. JailBench [78] 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 [80] expands on sensitive categories specific to the Chinese context, including political sensitivity, pornography, and variant/homophonic words. Latent Jailbreak [81] innovatively embeds harmful instructions into translation tasks, aiming to evaluate the robustness of LLMs against jailbreak attacks through translation tasks. Finally, S-Eval [83] introduces a bilingual dataset of risk prompts in both Chinese and English, containing fundamental risk prompts along with their corresponding adversarially crafted attack prompts. In summary, Q-only datasets provide broad coverage for testing refusal and robustness, but they lack reference responses for fine-grained evaluation.

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 [30] 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 [69] includes a collection of manually annotated harmful prompt-response pairs set in complex scenarios, while AegisSafetyDataset [71] is a large-scale record of safety and unsafe interactions between humans and LLMs. Additionally, WildGuardMix [74] 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 [39] provides augmented prompts and responses aimed at testing and improving LLMs safety, supporting a targeted approach to model safety enhancement. Overall, Q&A datasets enable accurate assessment and model training but require extensive annotation efforts.

Both Q-only and Q&A datasets are used to evaluate whether the content generated by LLMs in response to harmful queries complies with relevant policies. In contrast to this generative approach, multiple-choice questions (MCQ), as a form of discriminative evaluation, place a greater emphasis on the LLMs’ ability to identify harmful content. Furthermore, a major advantage of MCQ over generative evaluation is that they can provide a highly intuitive and convenient measure of a model’s safety performance, as the answers are unique and the evaluation results can be determined straightforwardly. SALAD-Bench [72], 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 [76] is designed to identify risky content by requiring LLMs to select the single safe or unsafe option from multiple choices. CValues [38] 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 [79] constructs multiple-choice questions, each with four manually labeled safety levels: Extremely Hazardous, Moderately Hazardous, Neutral, and Safe, enabling fine-grained evaluation. SafetyBench [82] 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. MCQ datasets offer intuitive, reproducible evaluation metrics but may oversimplify complex generative behaviors.

Some jailbreak attack methods have evolved beyond single-turn attacks, extending to multi-turn interactions, which have shown significantly heightened risks [184]–[186]. Consequently, evaluating the safety of LLMs in multi-turn conversations(Dialogue) has become necessary. R-Judge [66] provides a benchmark for assessing LLMs capabilities in identifying and evaluating safety risks within multi-turn interaction records between agents. CoSafe [68] 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 [24] presents a dataset not only of harmful questions but also includes both blue and red conversations, making it suitable for fine-tuning safety models. Multi-turn datasets are essential for assessing real-world conversational safety, though they are resource-intensive to construct and analyze.

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 [187]. 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 [188]. Thus, while much research is focused on English and Chinese, the safety of LLMs in multilingual and low-resource language contexts deserves attention. AraTrust [84] and AnswerCarefully [85] extend benchmarks to Arabic and Japanese, respectively, while XSafety [86], JailJudge OOD [69], and PolyglToxicityPrompts [23] offer multilingual coverage across 10–17 languages. These datasets highlight the importance of assessing LLMs safety in diverse linguistic contexts.

## V. HOW TO EVALUATE

Having identified what and where to evaluate in LLM safety, the next critical issue centers on how such evaluations should be conducted. A variety of methodologies have emerged, differing in levels of automation, granularity, and evaluator involvement. This section reviews the major evaluation paradigms, classifying them according to the nature and role of the evaluator—human, rule-based, or model-based—and discusses their respective advantages, limitations, and use cases. In addition, some evaluation frameworks that integrate the entire evaluation pipeline are also summarized.

### A. Human-based Evaluation

Since the birth of the field of machine learning, human judgment has been regarded as the gold standard. In the safety evaluation of LLMs, central concerns include toxicity, ethics, and bias. A persistent challenge in this area is defining what “safety” means in relation to model outputs. Human-based evaluation holds a unique advantage here, as it allows for nuanced interpretation and context-sensitive judgment that automated approaches struggle to replicate [189]. Human evaluators can recognize subtle cues such as tone, sarcasm, or implicit meanings, making them particularly well-suited to detecting toxicity and bias. Nevertheless, human evaluation faces significant challenges, especially in terms of cost and consistency. It necessitates significant human and time resources [190], and in large-scale evaluations, as the volume of content increases, the associated costs also rise substantially. For instance, Yuan *et al.* [134] compensate annotators at a rate of \$27.5 per hour according to the standard hourly income for U.S. annotators. Moreover, variations in cultural background, knowledge, and personal values can reduce consistency across annotators, particularly in sensitive domains such as toxicity and bias [109].

Despite these limitations, human evaluation remains indispensable because automated methods are still constrained in their ability to grasp complex or culturally embedded content. Models often fail to interpret sarcasm, puns, or layered meanings, whereas human annotators can provide a more comprehensive understanding. This explains why many LLM safety studies continue to integrate human judgment into their evaluation pipelines. For instance, Meadows *et al.* [36] relied on human evaluation to examine whether LLM outputs aligned with Australian values, offering insights for global regulatory bodies. Similarly, Movva *et al.* [87] used annotators to evaluate chatbot responses across five safety dimensions—harm, bias, misinformation, political stance, and polarization—analyzing the agreement between human and model-based assessments. Human judgment is also a key reference point in jailbreak attack research, where it serves as the benchmark for evaluating both the success of jailbreaks and the reliability of automated evaluators. For example, Souly *et al.* [63] showed strong alignment between their proposed method and human judgment by comparing bias levels and mean absolute error (MAE) against human evaluation results.

### B. Rule-based Evaluation

Rule-based evaluation in the safety assessment of LLMs typically involves the detection of potential risks or inappropriate content in model outputs by defining a set of explicit rules or guidelines. Such methods rely on well-specified criteria to flag responses that may contain sensitive information, harmful language, or other safety concerns. The typical process involves defining rule sets, applying them to model outputs, and categorizing responses accordingly.

Among these methods, string matching remains the most widely adopted, particularly in jailbreak attack research. In this setting, evaluators probe models with harmful queries and determine the success of the jailbreak attempt by checking

for the presence of “rejection strings” in the output (e.g., “I am sorry” or “As an AI”). Each study typically constructs its own phrase set, often based on heuristic observations from empirical experiments. String matching is favored for its interpretability and cost-effectiveness. However, it suffers from significant drawbacks. Most notably, it exhibits a high false positive rate and tends to overestimate unsafe behavior, with low correlation to human judgment. This occurs because rejection phrase lists cannot exhaustively capture all valid refusal expressions. For instance, the list employed by Zou *et al.* [8] excludes common phrases such as “not legal” and “not ethical”, which may lead to misclassification of safe outputs as harmful. Furthermore, jailbreak prompts often embed benign content to obscure malicious intent. In such cases, model responses may be irrelevant or unhelpful to the attack but are nevertheless flagged as unsafe under string-matching rules. Despite these limitations, string matching remains widely used, largely due to its negligible implementation cost and the need for comparability with prior studies that have adopted the same methodology [133], [191]. In addition, string matching has been observed to be combined with other security evaluation methods. For example, Liu *et al.* [88] combined string matching with text classifiers, while Zeng *et al.* [89] used string matching as a preliminary filter before conducting a more precise evaluation with GPT-4.

Beyond string matching, rule-based evaluation can also integrate traditional machine learning metrics to assess more specialized risks. For example, Liu *et al.* [60] proposed a framework to evaluate copyright compliance in text generation. Their method combined multiple rule-based indicators, including similarity metrics such as the Longest Common Substring (LCS) [192] and ROUGE-L scores [125], to quantify the overlap between generated text and copyrighted material. They further incorporated rejection rates as an additional measure of the model’s capacity to deny requests for infringing outputs. This integration of rule-based and metric-based evaluation provides a more comprehensive perspective than string matching alone.

In summary, rule-based evaluation offers clear, interpretable, and cost-efficient criteria for identifying unsafe or noncompliant content in LLM outputs. Nonetheless, its dependence on predefined rules makes it inherently limited in detecting novel risks or subtle safety issues. As such, rule-based methods are most valuable when employed as baseline or complementary tools, often in combination with human or model-based evaluation, to ensure more robust and scalable safety assessments.

### C. Model-based Evaluation

Model-based evaluation refers to the use of high-performance models to assess the safety of text generated by LLMs, such as GPT-4 [111], BERT [193], and reward models [194]. In this paradigm, models function as “evaluators” that interpret, analyze, and determine whether generated content adheres to predefined safety standards. Compared with rule-based evaluation, model-based methods can capture complex semantic structures and context-dependent nuances, thereby identifying subtle biases and harmful contents [195]–[197].

Moreover, in contrast to human evaluation, model-based approaches scale efficiently to large volumes of data, reducing human labor and associated costs. Broadly, model-based evaluation techniques can be divided into three categories: (i) specialized text classifiers, (ii) closed-source chat models, and (iii) open-source chat models.

1) **Text Classifiers:** Text classifiers generate structured outputs such as labels or confidence scores, which makes them particularly well-suited for safety evaluation tasks.

Several companies provide proprietary content moderation services for detecting harmful text. Widely used tools include OpenAI’s Moderation Endpoint [90], Microsoft’s Azure AI Content Safety API [92], and Google’s Perspective API [91]. For instance, OpenAI’s Moderation Endpoint is designed to determine whether content complies with its usage policies. It evaluates 11 categories—including hate, harassment, self-harm, sexual content, and violence—returning both binary labels and confidence scores between 0 and 1. Google’s Perspective API adopts a similar approach, focusing on “toxicity,” which refers to rude, disrespectful, or otherwise inappropriate text, but leaves the choice of decision threshold to users. While such closed-source classifiers are convenient, their utility for research purposes is limited. They primarily focus on surface-level harmful content, which can result in false positives in scenarios where unsafe behavior occurs implicitly (e.g., jailbreak demonstrations without overtly harmful text) [93].

In contrast, open-source classifiers are more widely used in academic evaluations. These models are typically fine-tuned from established sequence classification architectures such as BERT, RoBERTa, or DeBERTa. With parameter counts often below one billion, they are significantly smaller than contemporary open-source chat models. For example, Wang *et al.* [67] fine-tuned a Longformer [198] for harmful response detection, while Yu *et al.* [93] fine-tuned a RoBERTa-large model [130] for evaluation tasks. Nevertheless, the effectiveness of such classifiers heavily depends on the quality of their training datasets, and their performance often degrades on out-of-distribution data.

2) **Closed-source Chat Model:** Closed-source chat models, when employed as evaluators, provide both high performance and versatility. Models such as GPT-4 and InstructGPT, trained on massive datasets, demonstrate strong abilities in contextual understanding, semantic interpretation, and intent recognition. As a result, they are increasingly adopted as evaluators for assessing the safety and quality of model outputs. In particular, GPT-4-as-a-judge has become the most widely used evaluation method in this domain, showing strong correlation with human judgments.

The implementation details of the evaluation process are all based on prompt engineering. By designing an evaluation prompt word, the model is prompted to evaluate the input-output pair according to the evaluation rules. For example, the GPT-4 Judge prompt [94] uses a Likert-style prompt to assess the harmfulness of a response. The model is informed of OpenAI’s relevant safety policies as a reference for the evaluation process. PAIR prompt [95] assesses the degree to which the model’s response violates ethical, legal, or safety guidelines, outputting a score out of 10. Similarly, PICT [96]

assesses whether the victim model’s response contains harmful information related to a given category of prohibited behavior.

Despite their advantages, closed-source models face challenges in reliability. Prior studies highlight issues such as limited reproducibility, sensitivity to prompt wording, and a lack of robustness [197], [199], [200]. These shortcomings undermine consistency across evaluations and raise concerns regarding reproducibility and transparency in academic or regulatory contexts. While closed-source models remain powerful and cost-efficient alternatives to human evaluation, further research is needed to enhance their robustness and reliability.

3) **Open-source Chat Model:** Several researchers have released open-source chat 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. Existing works can be broadly divided into two categories: models focusing on binary results with classification labels, and models that emphasize diversified labels and detailed reasoning.

**Binary results and classification labels.** Llama Guard [97] 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 [98], Llama Guard 3 1B [99], and Llama Guard 3 8B [100]. Similarly, MD-Judge [72], inspired by Llama Guard, fine-tuned the Mistral-7B model [201] 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.

**Diversified results and reasoning.** Other works go beyond binary classification and focus on providing richer evaluation outcomes. ShieldLM [101], fine-tuned from Qwen-14B-Chat [202], 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. AegisSafetyExperts [71] 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, 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 for prompt classification. WildGuard [74] also fine-tuned Mistral-7B to detect malicious intent in user prompts, assess safety risks in model responses, and determine the model’s refusal rate. Its training data includes synthetic adversarial, synthetic vanilla, and real-world user-LLM interaction data to maximize task coverage, diversity, and balance. ShieldGemma [102] 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 and WildGuard. Finally, JailJudge Guard [69] represents the latest research in this line. By fine-tuning the Llama2-7B model with instruction tuning, it produces detailed reasoning together with a fine-grained evaluation score (a jailbreak score from 1 to 10).

Overall, open-source chat models present a flexible alternative for automated safety evaluation. The primary differences among existing models lie in the composition of training data and the form of evaluation outputs. However, their deployment requires access to large-scale, high-quality labeled datasets and significant computational resources for fine-tuning. Moreover, their generalization to out-of-distribution data remains an open challenge, raising questions about their reliability in real-world applications.

#### D. Evaluation Frameworks

Evaluation frameworks integrate the entire evaluation pipeline, enabling convenient end-to-end assessment and simplifying the process. These frameworks take evaluation datasets as input and produce evaluation results as output. In the intermediate process, in addition to interacting with target models through interfaces (such as HF, VLLM, or OpenAI) to obtain outputs and conduct evaluation, some frameworks also include components for attack augmentation on data or defense enhancement on models. These are designed to test the robustness of target models against attacks and the effectiveness of defense methods. As a result, most evaluation frameworks are typically composed of four parts, or a subset of them: Attacks, Defenses, Judges, and Interfaces.

As shown in Table III, we summarize several representative evaluation frameworks and provide a brief comparison. JailbreakEval [17] is a toolkit dedicated to evaluating jailbreak attempts. It does not include a preliminary data processing stage but instead focuses only on evaluating results. Its input is “question–answer” pairs, and it provides a set of out-of-the-box evaluators. In contrast, EasyJailbreak [103] is designed specifically for jailbreak attacks and consists of four components: Selector, Mutator, Constraint, and Evaluator. The process has three stages: in the preparation stage, users provide malicious prompts, prompt templates, or other configurations; in the attack stage, EasyJailbreak iteratively attacks the target model and updates the prompts while evaluating the outcome of each attack; in the final stage, the system aggregates results and generates a report for the user. Moving beyond general jailbreak-specific settings, WalledEval [104] introduces innovation by supporting a wide range of safety benchmarks, with more than 35 benchmarks built in. These cover multilingual safety, exaggerated safety, prompt injection, and red-teaming datasets. Finally, HarmBench [31] is a standardized framework for automated red-teaming. It emphasizes three qualities of automated red-team evaluation: breadth, comparability, and robustness. Based on these principles, the pipeline begins by converting a diverse set of behaviors into test cases to probe target models, and then applies multiple classifiers to evaluate the responses and calculate the attack success rate.

The above frameworks do not consider defenses. In contrast, JailbreakBench [30] incorporates several defense methods

Table III  
COMPARISON OF DIFFERENT EVALUATION FRAMEWORKS.

Frameworks	Attacks	Defenses	Judges	Interfaces
JailbreakEval [17]	-	-	22	-
EasyJailbreak [103]	12	-	7	4
WalledEval [104]	9	-	6	6
HarmBench [31]	18	-	3	5
JailbreakBench [30]	4	5	6	3
AI SafetyLab [105]	13	16	7	3
PandaGuard [106]	19	12	4	7

compared with EasyJailbreak. This toolkit allows users to configure threat models, system prompts, chat templates, and scoring functions, and it maintains a public leaderboard that compares the performance of state-of-the-art jailbreak attacks and defenses. Building on a broader perspective, AISafetyLab [105] is organized into three core modules — attack, defense, and evaluation — along with four shared auxiliary modules: models, dataset, utils, and logging. It provides researchers with a highly extensible framework. Further, PandasGuard [106] unifies attackers, defenders, target models, and judges into a modular system. Its main contribution is conducting a large-scale benchmark of about 3B tokens across 49 models.

In terms of evaluators, these frameworks generally adopt rule-based string matching and model-based evaluation methods. There is no clear preference between closed-source and open-source chat models, though text classifiers are used less frequently. While existing frameworks continue to expand in attacks, defenses, judges, and interfaces, there is still no consensus on evaluation methods. The divergence in evaluation method choices further highlights the lack of standardized practices in the field. This inconsistency not only weakens the comparability of results across frameworks but also limits the accumulation of research progress. Overall, current evaluation frameworks focus more on integrating and balancing existing methods than on advancing evaluation methodology. For higher-level evaluation, it is crucial to develop mechanisms that combine efficiency, accuracy, and robustness, and to establish standardized benchmarks and evaluation protocols.

## 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 an 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

Despite growing efforts in safety evaluation of LLMs, the research community has yet to reach consensus on standardized procedures and methodologies for conducting evaluations. Prior work has highlighted that that LLM safety evaluation lacks robustness due to noise stemming from factors such as

small datasets, methodological inconsistencies, and unreliable evaluation setups [203]. 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 concerns about reproducibility and reliability [63]. Furthermore, as LLMs are increasingly applied in diverse domains, ranging from education to healthcare and law, domain-specific safety requirements further exacerbate these inconsistencies. Therefore, we argue that future research should aim to establish a more unified evaluation paradigm. While achieving complete unification may be difficult, progress can be made by (i) developing standardized evaluation procedures and reporting protocols, (ii) defining unified and comprehensive metrics with clear computational formulations, and (iii) designing a multi-layered, modular evaluation framework that can flexibly adapt to different models, domains, and application scenarios. Such efforts would not only improve the comparability and reproducibility of safety evaluations but also foster a shared foundation for cumulative progress in this rapidly evolving field.

### B. Dynamic Evaluation

Most benchmarks for LLMs safety evaluation are released as static resources, with limited subsequent updates. This static approach poses several challenges. First, as LLMs evolve through continuous iteration, prompts or attack methods that once exposed vulnerabilities may no longer be effective, resulting in artificially inflated safety performance. Second, evaluation datasets risk contamination, as their contents may be incorporated into model training or fine-tuning for alignment, thereby reducing their validity as independent evaluation resources. Third, adversarial techniques continue to advance; for instance, jailbreak attacks often disguise malicious prompts to bypass safeguards and induce harmful outputs. As new jailbreak strategies emerge, static benchmarks quickly become outdated and insufficient for rigorous safety assessment. Although some initial studies have begun to explore dynamic evaluation [204], [205], this line of research is still nascent. Future efforts should aim to design adaptive safety evaluation methods that enable benchmarks and protocols to evolve in tandem with both model capabilities and adversarial tactics. Promising directions include automated pipelines for generating novel adversarial prompts, community-driven platforms for continuously updating benchmarks, and integrating dynamic evaluation into red-teaming frameworks. Such approaches would help ensure that safety evaluations remain relevant, robust, and reflective of real-world risks.

### C. Reliable and Efficient Evaluator

As discussed in Section V, current evaluation methods for LLMs safety can be broadly categorized into rule-based, human-based, and model-based approaches. First, rule-based methods are simple and efficient but often fail to capture implicit harms beyond predefined patterns, 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 costly, time-intensive, and difficult to scale. Moreover, cross-cultural and moral value differences among annotators can introduce variability in harm perception and compromise the consistency of results [206]. Third, 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, they face significant challenges, including high inference costs, substantial computational requirements for fine-tuning, limited reproducibility, high sensitivity to evaluation prompts, and lack of robustness [197], [199], [200]. 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. Future research should prioritize the development of evaluators that are both reliable and efficient. Promising directions include hybrid approaches that combine automated filters with lightweight model-based scoring and selective human oversight, methods to enhance the consistency and reproducibility of model-based evaluators (e.g., prompt calibration and majority vote).

### D. Application-Specific Evaluation

As LLMs are increasingly integrated into various real-world applications, evaluating their safety in these specific application tasks has become crucial. For example, [207] 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. General benchmarks, which primarily test for risks such as toxicity or bias, are often insufficient in high-stakes domains, where unique and domain-dependent hazards must be rigorously addressed. 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 gap underscores the urgent need to establish safety evaluation frameworks tailored to emerging domains. Importantly, these frameworks should not only extend beyond generic benchmarks but also incorporate domain expertise and align with regulatory and ethical standards relevant to each field (e.g., FDA regulations in medicine, or financial compliance requirements). Practical implementations could include the creation of domain-specific benchmark datasets, the use of simulation-based environments to test LLM behavior in high-risk scenarios, and the integration of interdisciplinary perspectives to ensure robustness and applicability.

### E. Safety Evaluation for MLLMs and Agents

Multimodal large language models (MLLMs) and agents, both grounded in LLM technology, have attracted growing attention across academia and industry. MLLMs enable more natural and seamless human-machine interaction by integrating text with modalities such as images, audio, and video, while AI agents emulate human-like reasoning and decision-making processes. In MLLMs, LLMs serve as the core engine for text understanding and generation [208], whereas in agents, LLMs function as the brain, orchestrating planning and decision-making. Despite rapid progress, the safety evaluation of both directions significantly lags behind their development. For MLLMs, existing benchmarks remain limited [209]–[212], lacking robustness metrics and evaluation standards tailored to multimodal contexts. In particular, video-based domains such as video understanding and generation remain underexplored. For agents, safety risks are amplified by the integration of multiple new components, which create additional attack surfaces and require evaluation in complex, interactive environments. Examples include emerging communication protocols such as the Model Context Protocol (MCP) and agent-to-agent (A2A) communication, which introduce new possibilities for coordination but also novel vulnerabilities. While some studies have begun addressing these challenges [66], [213], [214], no consensus has yet been reached on unified benchmark design standards for the broader agent ecosystem [215]. Key open challenges include identifying the core dimensions of agent safety, designing evaluation methods capable of capturing risks in dynamic, multi-agent environments, and developing comprehensive frameworks that can evolve alongside these rapidly advancing systems.

### F. Beyond Evaluation: From Safe to Responsible

Safety evaluation should not be regarded as the ultimate goal, but as the foundation for further progress. It is not only about producing benchmark results to guide the safe deployment of LLMs, but also about shaping directions for future research and development. The core role of safety evaluation is to identify and mitigate risks, thereby ensuring model stability and reliability. Yet, as technology advances and societal demands grow more complex, safety alone will no longer be sufficient to meet the expectations for real-world applications. Building on current development trends and research in LLM safety evaluation, we argue that the future trajectory should move from merely safe LLMs toward truly responsible LLMs. Responsible LLMs should extend beyond basic safety, fairness, and trustworthiness to also embody transparency, interpretability, and auditability. In other words, models must not only be technically reliable, but also broadly acknowledged for their ethical integrity and social responsibility. Correspondingly, safety evaluation itself will continue to evolve, encompassing a wider set of dimensions and employing deeper, more sophisticated methods in order to support this transition.

By expanding and refining research into the above areas, the community can more accurately evaluate the safety of LLMs while also promoting their development towards higher

safety standards. In the future, LLMs will be closely aligned with societal values and norms, evolving into a responsible paradigm of artificial intelligence.

## VII. CONCLUSION

Ensuring the safe deployment and responsible use of LLMs in real-world applications requires systematic safety evaluation. This paper provides the first comprehensive review of LLMs safety evaluation, organized around four guiding questions: “Why evaluate,” “What to evaluate,” “Where to evaluate,” and “How to evaluate.” We systematically survey key dimensions of LLMs safety, including toxicity, robustness, ethics, bias and fairness, and truthfulness, as well as specific downstream tasks such as privacy, mental health, medical applications, code generation and copyright. In addition, we review current evaluation metrics, and a broad set of datasets and benchmarks, highlighting their characteristics, strengths, and limitations. The roles of different evaluators — human-based, rule-based, and model-based — are also examined to illustrate the current state of LLM safety evaluation. Moreover, we review several representative evaluation frameworks, highlighting how they integrate attacks, defenses, target models, judges in practical assessments. This structured overview aims to provide researchers with a clear understanding of available resources and methodologies, supporting systematic and reproducible safety evaluations. Finally, we highlight several major challenges currently faced in safety evaluation of LLMs and discuss potential directions for future research.

Looking forward, we anticipate that continued research in safety evaluation will facilitate the development of LLMs that are more reliable, responsible, and aligned with human values. By combining robust evaluation frameworks with improved model design, future LLMs can provide intelligent support across diverse applications while effectively mitigating potential risks, ultimately fostering a balanced approach between technological innovation and trustworthy safety.

## REFERENCES

- [1] 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.
- [2] A. L. Kotian, R. Nandipati, 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] P. Mondorf and B. Plank, “Beyond accuracy: Evaluating the reasoning behavior of large language models—a survey,” *arXiv preprint arXiv:2404.01869*, 2024.



- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] F. W. Liu and C. Hu, "Exploring vulnerabilities and protections in large language models: A survey," *arXiv preprint arXiv:2406.00240*, 2024.
- [14] 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.
- [15] Z. Dong, Z. Zhou, C. Yang, J. Shao, and Y. Qiao, "Attacks, defenses and evaluations for llm conversation safety: A survey," *arXiv preprint arXiv:2402.09283*, 2024.
- [16] S. Yi, Y. Liu, Z. Sun, T. Cong, X. He, J. Song, K. Xu, and Q. Li, "Jailbreak attacks and defenses against large language models: A survey," *arXiv preprint arXiv:2407.04295*, 2024.
- [17] 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.
- [18] P. Röttger, F. Pernisi, B. Vidgen, and D. Hovy, "Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 26, 2025, pp. 27 617–27 627.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] D. Jain, P. Kumar, S. Gehman, X. Zhou, T. Hartvigsen, and M. Sap, "Polyglototoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models," *arXiv preprint arXiv:2405.09373*, 2024.
- [24] R. Bhardwaj and S. Poria, "Red-teaming large language models using chain of utterances for safety-alignment," *arXiv preprint arXiv:2308.09662*, 2023.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] P. Chao, E. DeBenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwal, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer *et al.*, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 005–55 029, 2024.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] J. Ji, Y. Chen, M. Jin, W. Xu, W. Hua, and Y. Zhang, "Moralbench: Moral evaluation of llms," *arXiv preprint arXiv:2406.04428*, 2024.
- [38] 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.
- [39] H. Sun, Z. Zhang, J. Deng, J. Cheng, and M. Huang, "Safety assessment of chinese large language models," *arXiv preprint arXiv:2304.10436*, 2023.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] Z. Su, X. Zhou, S. Rangrejji, 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] H. Iqbal, Y. Wang, M. Wang, G. Georgiev, J. Geng, I. Gurevych, and P. Nakov, "Openfactcheck: A unified framework for factuality evaluation of llms," *arXiv preprint arXiv:2408.11832*, 2024.
- [53] 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.

- [54] 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.
- [55] 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.
- [56] 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>
- [57] M. Bhatt, S. Chennabasappa, C. Nikolaidis, S. Wan, I. Evtimov, D. Gabi, D. Song, F. Ahmad, C. Aschermann, L. Fontana *et al.*, "Purple llama cyberseval: A secure coding benchmark for language models," *arXiv preprint arXiv:2312.04724*, 2023.
- [58] J. Peng, L. Cui, K. Huang, J. Yang, and B. Ray, "Cweval: Outcome-driven evaluation on functionality and security of llm code generation," in *2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code)*. IEEE, 2025, pp. 33–40.
- [59] X. Li, J. Ding, C. Peng, B. Zhao, X. Gao, H. Gao, and X. Gu, "Safegenbench: A benchmark framework for security vulnerability detection in llm-generated code," *arXiv preprint arXiv:2506.05692*, 2025.
- [60] 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.
- [61] Z. Hui, Y. R. Dong, E. Shareghi, and N. Collier, "Trident: Benchmarking llm safety in finance, medicine, and law," *arXiv preprint arXiv:2507.21134*, 2025.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: Evaluating safeguards in llms," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 896–911.
- [68] 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.
- [69] F. Liu, Y. Feng, Z. Xu, L. Su, X. Ma, D. Yin, and H. Liu, "Jailjudge: A comprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework," *arXiv preprint arXiv:2410.12855*, 2024.
- [70] T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwal, 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.
- [71] 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.
- [72] 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.
- [73] 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] L. Shi and D. Xiong, "Criskeval: A chinese multi-level risk evaluation benchmark dataset for large language models," *arXiv preprint arXiv:2406.04752*, 2024.
- [80] 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.
- [81] 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.
- [82] 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.
- [83] 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.
- [84] 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.
- [85] H. Suzuki, S. Katsumata, T. Kodama, T. Takahashi, K. Nakayama, and S. Sekine, "Answercarefully: A dataset for improving the safety of japanese llm output," *arXiv preprint arXiv:2506.02372*, 2025.
- [86] 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.
- [87] R. Movva, P. W. Koh, and E. Pierson, "Annotation alignment: Comparing llm and human annotations of conversational safety," *arXiv preprint arXiv:2406.06369*, 2024.
- [88] T. Liu, Y. Zhang, Z. Zhao, Y. Dong, G. Meng, and K. Chen, "Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 4711–4728.
- [89] Y. Zeng, Y. Wu, X. Zhang, H. Wang, and Q. Wu, "Autodefense: Multi-agent llm defense against jailbreak attacks," *arXiv preprint arXiv:2403.04783*, 2024.
- [90] OpenAI, "OpenAI Moderation Endpoint," <https://platform.openai.com/docs/guides/moderation/>, accessed: 2024-12-12.
- [91] Google, "Google Perspective API," <https://www.perspectiveapi.com/>, accessed: 2024-12-12.
- [92] Microsoft, "Microsoft Azure AI Content Safety API," <https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety>, accessed: 2024-12-12.
- [93] J. Yu, X. Lin, Z. Yu, and X. Xing, "Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts," *arXiv preprint arXiv:2309.10253*, 2023.
- [94] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" *arXiv preprint arXiv:2310.03693*, 2023.
- [95] 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.
- [96] R. Shah, S. Pour, A. Tagade, S. Casper, J. Rando *et al.*, "Scalable and transferable black-box jailbreaks for language models via persona modulation," *arXiv preprint arXiv:2311.03348*, 2023.
- [97] 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.
- [98] Llama Team, "Meta llama guard 2," [https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL\\_CARD.md](https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md), 2024.
- [99] Llama Team, AI @ Meta, "The llama 3 family of models," [https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL\\_CARD.md](https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL_CARD.md), 2024.
- [100] 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.

- [101] 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.
- [102] W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu *et al.*, “Shield-gemma: Generative ai content moderation based on gemma,” *arXiv preprint arXiv:2407.21772*, 2024.
- [103] 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.
- [104] 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.
- [105] Z. Zhang, L. Lei, J. Yang, X. Huang, Y. Lu, S. Cui, R. Chen, C. Zhang, X. Wang, H. Wang *et al.*, “Aisafetyslab: A comprehensive framework for ai safety evaluation and improvement,” *arXiv preprint arXiv:2502.16776*, 2025.
- [106] G. Shen, D. Zhao, L. Feng, X. He, J. Wang, S. Shen, H. Tong, Y. Dong, J. Li, X. Zheng *et al.*, “Pandaguard: Systematic evaluation of llm safety against jailbreaking attacks,” *arXiv preprint arXiv:2505.13862*, 2025.
- [107] 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.
- [108] S. Yao, D. Yu, J. Zhao, I. Shafraan, 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.
- [109] 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.
- [110] 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.
- [111] 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.
- [112] 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.
- [113] 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.
- [114] 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.
- [115] 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.
- [116] S. Raza, A. Raval, and V. Chatrath, “Mbias: Mitigating bias in large language models while retaining context,” *arXiv preprint arXiv:2405.11290*, 2024.
- [117] E. Ferrara, “Should chatgpt be biased? challenges and risks of bias in large language models,” *arXiv preprint arXiv:2304.03738*, 2023.
- [118] 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.
- [119] 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.
- [120] C. Maple, L. Szpruch, G. Epiphanou, 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.
- [121] 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.
- [122] 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.
- [123] 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.
- [124] 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.
- [125] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [126] 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.
- [127] 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.
- [128] S. Curran, S. Lansley, and O. Bethell, “Hallucination is the last thing you need,” *arXiv preprint arXiv:2306.11520*, 2023.
- [129] 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.
- [130] Y. Liu, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [131] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “Realtotoxicityprompts: Evaluating neural toxic degeneration in language models,” *arXiv preprint arXiv:2009.11462*, 2020.
- [132] 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.
- [133] 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.
- [134] 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.
- [135] 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.
- [136] M. Andriushchenko, F. Croce, and N. Flammarion, “Jailbreaking leading safety-aligned llms with simple adaptive attacks,” *arXiv preprint arXiv:2404.02151*, 2024.
- [137] Alex Albert, “Jailbreak chat,” <https://www.jailbreakchat.com/>, accessed: 2024-02-20.
- [138] 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.
- [139] 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.
- [140] 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.
- [141] 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.
- [142] 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.
- [143] 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.
- [144] 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.
- [145] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, “A survey on llm-as-a-judge,” *arXiv preprint arXiv:2411.15594*, 2024.
- [146] 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.
- [147] C. Chataigner, A. Taïk, and G. Farnadi, “Multilingual hallucination gaps in large language models,” *arXiv preprint arXiv:2410.18270*, 2024.

- [148] 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.
- [149] 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.
- [150] C. Li, P. Wang, C. Wang, L. Zhang, Z. Liu, Q. Ye, Y. Xu, F. Huang, X. Zhang, and P. S. Yu, “Loki’s dance of illusions: A comprehensive survey of hallucination in large language models,” *arXiv preprint arXiv:2507.02870*, 2025.
- [151] 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,” *Computational Linguistics*, pp. 1–46, 2025.
- [152] 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,” *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.
- [153] 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.
- [154] A. Wang, K. Cho, and M. Lewis, “Asking and answering questions to evaluate the factual consistency of summaries,” *arXiv preprint arXiv:2004.04228*, 2020.
- [155] S. Es, J. James, L. E. Anke, and S. Schockaert, “Ragas: Automated evaluation of retrieval augmented generation,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.
- [156] 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.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan *et al.*, “Attributed text generation via post-hoc research and revision,” *arXiv preprint arXiv:2210.08726*, 2022.
- [161] I. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, P. Liu *et al.*, “Factool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios,” *arXiv preprint arXiv:2307.13528*, 2023.
- [162] Y. Wang, R. G. Reddy, Z. M. Mujahid, A. Arora, A. Rubashevskii, J. Geng, O. M. Afzal, L. Pan, N. Borenstein, A. Pillai *et al.*, “Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output,” *arXiv preprint arXiv:2311.09000*, 2023.
- [163] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le *et al.*, “Freshllms: Refreshing large language models with search engine augmentation,” *arXiv preprint arXiv:2310.03214*, 2023.
- [164] 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.
- [165] J. Hong, S. Levine, and A. Dragan, “Zero-shot goal-directed dialogue via rl on imagined conversations,” *arXiv preprint arXiv:2311.05584*, 2023.
- [166] 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.
- [167] B. Yang, M. A. Al Mamun, J. M. Zhang, and G. Uddin, “Hallucination detection in large language models with metamorphic relations,” *Proceedings of the ACM on Software Engineering*, vol. 2, no. FSE, pp. 425–445, 2025.
- [168] 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.
- [169] 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.
- [170] 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.
- [171] 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.
- [172] 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.
- [173] 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.
- [174] C. Dwork, “Differential privacy,” in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.
- [175] 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.
- [176] 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.
- [177] U. Hacoheh, 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.
- [178] 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.
- [179] 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.
- [180] 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.
- [181] 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.
- [182] Z. Chu, S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, X. Hu, J. Liang, P. S. Yu *et al.*, “Llm agents for education: Advances and applications,” *arXiv preprint arXiv:2503.11733*, 2025.
- [183] D. Zhang, H. Zheng, W. Yue, and X. Wang, “Advancing its applications with llms: A survey on traffic management, transportation safety, and autonomous driving,” in *International Joint Conference on Rough Sets*. Springer, 2024, pp. 295–309.
- [184] 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.
- [185] 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.
- [186] 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.
- [187] G. Dong, H. Wang, J. Sun, and X. Wang, “Evaluating and mitigating linguistic discrimination in large language models,” *arXiv preprint arXiv:2404.18534*, 2024.
- [188] J. Song, Y. Huang, Z. Zhou, and L. Ma, “Multilingual blending: Llm safety alignment evaluation with language mixture,” *arXiv preprint arXiv:2407.07342*, 2024.
- [189] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, “Why we need new evaluation metrics for nlg,” *arXiv preprint arXiv:1707.06875*, 2017.
- [190] 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.
- [191] X. Liu, N. Xu, M. Chen, and C. Xiao, “Autodan: Generating stealthy jailbreak prompts on aligned large language models,” *arXiv preprint arXiv:2310.04451*, 2023.
- [192] A. Karamolegkou, J. Li, L. Zhou, and A. Søgaard, “Copyright violations and large language models,” *arXiv preprint arXiv:2310.13771*, 2023.

- [193] 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.
- [194] 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.
- [195] C.-H. Chiang and H.-y. Lee, “Can large language models be an alternative to human evaluations?” *arXiv preprint arXiv:2305.01937*, 2023.
- [196] 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.
- [197] 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.
- [198] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.
- [199] 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.
- [200] F. Eiras, E. Zemor, E. Lin, and V. Mugunthan, “Know thy judge: On the robustness meta-evaluation of llm safety judges,” *arXiv preprint arXiv:2503.04474*, 2025.
- [201] 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.
- [202] 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.
- [203] T. Beyer, S. Xhonneux, S. Geisler, G. Gidel, L. Schwinn, and S. Günemann, “Llm-safety evaluations lack robustness,” *arXiv preprint arXiv:2503.02574*, 2025.
- [204] P. Zhang, H. Jin, L. Kang, and H. Wang, “Guardval: Dynamic large language model jailbreak evaluation for comprehensive safety testing,” *arXiv preprint arXiv:2507.07735*, 2025.
- [205] M. Jindal, H. Shrawgi, P. Agrawal, and S. Dandapat, “SAGE: A generic framework for llm safety evaluation,” *arXiv preprint arXiv:2504.19674*, 2025.
- [206] 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.
- [207] 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.
- [208] 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.
- [209] 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.
- [210] 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.
- [211] 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.
- [212] T. Gu, Z. Zhou, K. Huang, D. Liang, Y. Wang, H. Zhao, Y. Yao, X. Qiao, K. Wang, Y. Yang *et al.*, “Mllmgaurd: A multi-dimensional safety evaluation suite for multimodal large language models,” *arXiv preprint arXiv:2406.07594*, 2024.
- [213] 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.
- [214] 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.
- [215] 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.