Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086229

REVIEW

Safety, Alignment, and Robustness of Large Language Models: A Review

Milad Moradi*

AI Research Lab, Tricentis, Vienna, Austria

*Corresponding Author: Milad Moradi. Email: m.moradi-vastegani@tricentis.com

Received: 26 May 2026; Accepted: 29 June 2026; Published: 13 August 2026

ABSTRACT: Large Language Models (LLMs) have rapidly evolved into general-purpose systems with broad applicability across information access, reasoning, decision support, and human-computer interaction. Their growing deployment, however, has intensified concerns regarding safety, alignment, and robustness, especially as these models become integrated with external tools, retrieval systems, and increasingly agentic workflows. This review provides an analytical overview of the principal risks, technical advances, evaluation practices, and future directions in this area. It first clarifies the conceptual foundations of safety, alignment, robustness, and reliability in the context of LLMs. It then examines the major risk categories associated with LLM deployment, including harmful content generation, hallucination, bias and fairness concerns, privacy leakage, security and misuse risks, and emerging challenges in reasoning-capable and agentic systems. The review further synthesizes recent advances in data-centric safety interventions, post-training alignment, inference-time control, adversarial defense, factuality-oriented safeguards, and system-level protections. It also analyzes the current evaluation landscape, highlighting benchmark fragmentation, limitations of existing methodologies, and the need for more realistic and reproducible assessment. The paper concludes by outlining future research directions toward scalable alignment, stronger real-world evaluation, safer agentic systems, and tighter integration between technical safeguards and governance-oriented approaches.

KEYWORDS: Large language models; safety; alignment; robustness; reliability; evaluation; agentic systems

1 Introduction

1.1 Background and Motivation

Large Language Models (LLMs) [1] have emerged as a distinctive safety challenge because they are not narrow task-specific predictors, but general-purpose generative systems deployed across a wide range of domains, users, and decision contexts. Their utility derives precisely from properties that complicate risk control: they generate open-ended outputs rather than selecting from a fixed label space, they interact directly with human users in natural language, and they are increasingly embedded into search, productivity, education, coding, healthcare, and public-facing information systems [2]. As a result, failures are not confined to conventional classification errors; they can take the form of plausible but false statements, unsafe advice, manipulative or biased outputs, and context-dependent behavior that is difficult to anticipate ex ante. This breadth of deployment means that even low-probability failure modes can accumulate into substantial practical risk when models are used at scale and in heterogeneous settings [3].

A further complication is that LLM behavior is shaped not only by model parameters but also by interaction structure: prompts, conversational history, retrieval context, system instructions, and downstream application constraints all influence what the model does in practice. This makes safety a system-level

property rather than a purely model-intrinsic one [4]. The same model may appear well behaved in benchmark settings yet produce problematic outputs in realistic human-AI configurations, especially when users anthropomorphize the system, over-trust fluent responses, or deliberately probe for failures. National Institute of Standards and Technology (NIST's) generative Artificial Intelligence (AI) profile captures this broader framing by treating risks such as confabulation, data privacy, information integrity, and human-AI configuration as central concerns, emphasizing that generative AI harms often arise from the interaction between model behavior, interface design, deployment context, and human expectations rather than from isolated model errors alone [5].

The motivation for reviewing safety, alignment, and robustness together is therefore both practical and conceptual. Safety concerns the prevention of harmful outputs and downstream effects [6]; alignment concerns whether model behavior tracks intended goals, norms, and constraints [7]; robustness concerns whether these desirable properties persist under adversarial prompting, distribution shift, long-context interactions, and increasingly agentic forms of operation [8]. These three dimensions now intersect more tightly as LLMs gain tool-use capabilities, access external data sources, and act in semi-autonomous workflows where errors can propagate beyond text generation into information retrieval, code execution, or decision support. Under these conditions, confabulation becomes an information-integrity problem, privacy leakage becomes a deployment and governance problem, and misalignment becomes a robustness problem when safeguards fail outside nominal settings. An analytical review is therefore timely because recent advances have improved controllability and utility, but they have not resolved the deeper question of how to build LLM systems that remain reliable, policy-compliant, and socially acceptable under realistic conditions of scale, uncertainty, and adaptive use.

1.2 Why Safety, Alignment, and Robustness Should Be Discussed Together

Safety, alignment, and robustness are often treated as separate research themes, but for LLMs they are more accurately understood as tightly coupled properties of the same system. Safety concerns the avoidance of harmful behavior and harmful downstream effects, including unsafe outputs, misleading advice, discriminatory treatment, privacy violations, and misuse-enabling assistance. Alignment concerns whether model behavior tracks intended objectives, human instructions, normative constraints, and socially acceptable values. Robustness concerns whether these desirable properties remain stable when the model is exposed to perturbations, adversarial prompting, ambiguous instructions, distribution shift, long conversational contexts, or unfamiliar deployment environments. This distinction is analytically useful, but it should not be mistaken for a real separation in practice: a model may appear aligned under ordinary prompts yet become unsafe under adversarial interaction, or it may satisfy a narrow safety filter while remaining misaligned with user intent or institutional norms. In other words, safety specifies what undesirable outcomes should be avoided, alignment specifies what behavioral targets the system should follow, and robustness determines whether those targets remain intact under realistic operating conditions [6–9].

Discussing these dimensions together is therefore essential for both conceptual clarity and technical evaluation. Safety without alignment reduces to surface-level harm control and may encourage brittle refusal behavior that degrades utility without addressing deeper behavioral objectives. Alignment without robustness is similarly incomplete, because a model that follows human intentions only in nominal settings cannot be relied upon in real-world deployment, where prompts are noisy, users are diverse, and malicious actors actively search for failure modes. Robustness without a clear account of safety and alignment is also insufficient, because preserving behavior under perturbation is not meaningful unless the preserved behavior is itself desirable. For LLMs, these dependencies become even stronger when models are embedded in tools, retrieval pipelines, and agentic workflows, where small deviations in generation can propagate into larger

downstream harms [10]. An integrated treatment of safety, alignment, and robustness is therefore not merely a matter of organizational convenience; it reflects the fact that the central challenge is to design systems whose objectives are appropriate, whose behavior is constrained in socially and operationally acceptable ways, and whose safeguards remain effective outside idealized test conditions.

1.3 Scope and Contributions of the Review

This paper presents an analytical narrative review of the safety, alignment, and robustness of LLMs, rather than a systematic review based on formalized search and screening procedures. Its purpose is not to exhaustively enumerate all published studies, but to synthesize the field's main technical and conceptual developments, clarify how its central problems are framed, and critically examine the relationships among methods, risks, and evaluation practices. In particular, the review focuses on the principal challenges that make LLMs difficult to control and assess; the major advances in alignment, safeguarding, and robustness-oriented intervention; the strengths and limitations of current evaluation paradigms; and emerging research directions, especially those associated with reasoning-capable and agentic systems. Beyond these focal themes, the paper also addresses the broader issues structured throughout its sections, including risk taxonomies, tradeoffs between utility and safety, system-level and sociotechnical considerations, and the growing gap between benchmark performance and real-world reliability. The main contribution of the review is therefore to provide an integrated and critical account of a rapidly evolving area, highlighting not only what progress has been made, but also where current approaches remain fragmented, brittle, or insufficient for dependable deployment.

The literature considered in this narrative review was selected through targeted searches of major scholarly databases and indexing platforms, including Google Scholar, arXiv, ACL Anthology, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and official standards or guidance sources such as NIST. Search terms combined concepts related to "large language models," "safety," "alignment," "robustness," "red teaming," "hallucination," "privacy," "jailbreaks," "bias," "evaluation," "agentic systems," and "LLM agents." The temporal scope emphasized work published from the emergence of modern instruction-following and preference-aligned LLMs to the most recent literature available at the time of revision, while also including earlier foundational works where necessary for conceptual context. Studies were prioritized when they introduced influential methods, benchmarks, risk taxonomies, empirical findings, or governance frameworks directly relevant to LLM safety, alignment, robustness, and evaluation. Because the review is analytical rather than systematic, the aim was not exhaustive coverage but a representative synthesis of technically significant, highly cited, methodologically relevant, and recent contributions across the main themes of the paper.

1.4 Organization of the Paper

The remainder of this review is organized as follows. Section 2 introduces the conceptual foundations of LLMs and clarifies the core notions of safety, alignment, robustness, and reliability. Section 3 surveys the main risks associated with LLMs, while Section 4 reviews recent advances in safety, alignment, robustness, and safeguarding. Section 5 examines how these properties are evaluated, and Section 6 discusses future research directions. Finally, Section 7 concludes the paper.

2 Conceptual Foundations

Having established the motivation, scope, and framing of the review, this section clarifies the conceptual foundations on which the subsequent analysis is built. It first outlines what LLMs are as technical and sociotechnical systems, and then distinguishes the core concepts of safety, alignment, robustness, and

reliability in order to reduce terminological ambiguity and provide a coherent basis for the risk, mitigation, evaluation, and future-direction sections that follow.

2.1 What Are Large Language Models?

LLMs are neural sequence models trained on massive text corpora to predict the next token in context, a learning objective that yields broad linguistic competence and surprising degrees of task generality [1]. Their foundation is typically established during pretraining, where exposure to large-scale and diverse textual data allows the model to acquire statistical regularities related to syntax, semantics, world knowledge, reasoning patterns, and domain conventions [11]. However, pretrained models are not inherently optimized for safe, helpful, or instruction-following behavior; rather, they are optimized to continue text plausibly. For this reason, modern LLM development usually includes instruction tuning, in which models are fine-tuned on curated prompt-response pairs so that they respond more directly to user requests, follow task specifications, and adopt more useful conversational behavior [12,13]. This shift from raw language modeling to instruction-following is foundational for deployment, because it transforms a general generative model into an interactive system intended to assist human users across heterogeneous tasks.

Beyond instruction tuning, many state-of-the-art LLMs are further shaped by preference optimization, where model outputs are adjusted using human or AI-generated preference signals to better reflect desired qualities such as helpfulness, harmlessness, honesty, or policy compliance [14]. LLMs are also increasingly coupled with external components, including retrieval augmentation for accessing documents beyond parametric memory and tool use for invoking search engines, calculators, code interpreters, databases, or other software systems. These integrations expand capability, but they also change the nature of the model from a standalone text generator into a component of a larger decision-making pipeline [15,16]. The next step in this progression is the rise of agentic extensions, in which LLMs perform multi-step planning, maintain memory, call tools iteratively, and pursue higher-level objectives with limited human intervention [10,17]. Conceptually, then, an LLM should not be understood only as a pretrained language model, but as a layered sociotechnical artifact whose behavior emerges from pretraining, post-training alignment, interface design, external tool access, and the broader system architecture in which it is embedded.

2.2 Defining Safety

In the context of LLMs, safety can be defined as the property that a model and its surrounding system avoid generating, enabling, or amplifying harmful outcomes across a range of technical and social conditions [6]. This includes harmful content safety, namely reducing outputs that are abusive, violent, hateful, exploitative, or otherwise dangerous [18]; misuse prevention, which concerns limiting the model's usefulness for malicious purposes such as fraud, cyber abuse, manipulation, or large-scale disinformation [19]; factual and informational safety, which addresses hallucination, misleading explanations, fabricated citations, and other forms of confident but unreliable output that can distort decision-making [20]; privacy and security safety, which includes preventing leakage of sensitive data, memorized personal information, prompt-injection vulnerabilities, and other security failures in integrated systems [21]; and societal safety, which concerns broader harms such as bias, discrimination, erosion of trust, and harmful impacts when LLMs are deployed in high-stakes institutional settings [22]. Importantly, safety in LLMs is not reducible to content filtering alone: it is a multidimensional and system-level concept that depends not only on what the model says, but also on how it is used, by whom, in what context, and with what downstream consequences.

2.3 Defining Alignment

In LLMs, alignment refers to the extent to which model behavior conforms to intended goals, human expectations, and normative constraints, rather than merely producing plausible or high-probability text [7]. At a basic level, this includes instruction alignment, whereby the model correctly interprets and follows user requests, task specifications, and conversational intent [23]. It also includes preference alignment, in which model behavior is shaped to reflect human or AI-provided judgments about qualities such as helpfulness, harmlessness, clarity, and honesty [24]. A broader and more difficult dimension is value alignment, which concerns whether model outputs and decisions remain consistent with socially acceptable values, ethical principles, and context-dependent norms, especially where human preferences are incomplete, conflicting, or culturally variable [25]. Closely related is constitutional or rule-based alignment, in which the model is guided by explicit principles, policies, or normative rules that constrain behavior beyond raw preference fitting [26]. The problem becomes more complex in the case of LLM agents, where alignment is no longer limited to single-turn response quality but must also govern planning, memory, tool use, delegation, and long-horizon behavior under partial supervision [27]. For this reason, contemporary alignment research increasingly distinguishes between practical post-training protocols, such as supervised fine-tuning and preference-based optimization, and broader alignment objectives concerning reliability, controllability, and normative acceptability. In analytical terms, alignment is best understood not as a single technique, but as a layered problem of ensuring that increasingly capable language-based systems pursue the right objectives, interpret human intent appropriately, and continue to do so when operating beyond tightly controlled interaction settings.

2.4 Defining Robustness

In LLMs, robustness refers to the capacity of the model and its surrounding system to preserve desirable behavior when exposed to perturbations, adversarial manipulation, novel conditions, or operational complexity [8]. One important dimension is adversarial robustness, which concerns resistance to jailbreaks, prompt attacks, obfuscation strategies, and other deliberately crafted inputs intended to bypass safeguards or induce harmful behavior [28,29]. Closely related is prompt robustness, namely the stability of model behavior under minor rephrasings, ambiguous instructions, conflicting cues, or variations in prompt format that should not substantially alter the quality or safety of the response [30]. Distributional robustness concerns performance under shifts in domain, task, user population, or input characteristics that differ from the model's training or tuning conditions [31,32], while multilingual robustness addresses whether these properties generalize across languages rather than being concentrated in high-resource English settings [33]. For conversational systems, multi-turn and long-context robustness is also critical, since models may degrade over extended interactions, forget earlier constraints, or become vulnerable to cumulative prompt manipulation [34,35]. Finally, robustness in tool-using systems extends beyond text generation to the reliable handling of retrieval, code execution, external APIs, and other actions whose failure modes can propagate into the environment [36]. Analytically, robustness is not simply about preserving task accuracy; it is about ensuring that helpfulness, safety, alignment, and policy compliance remain stable under realistic and potentially adversarial conditions of deployment.

2.5 Relationship among Safety, Alignment, Robustness, and Reliability

Safety, alignment, robustness, and reliability are closely related but analytically distinct properties of LLMs. Safety concerns the avoidance of harmful outputs and harmful downstream effects; alignment concerns whether model behavior conforms to intended goals, human instructions, and normative constraints; robustness concerns whether these desirable properties remain stable under perturbation, adversarial pressure, or changing deployment conditions; and reliability concerns the consistency and dependability of model performance across repeated uses and practical contexts [37]. These concepts overlap because a reliable system that is misaligned may consistently produce the wrong kind of behavior, while an aligned system that is not robust may fail as soon as inputs become noisy or adversarial, thereby undermining safety. In this sense, alignment helps specify what the model ought to do, robustness helps determine whether it continues to do so under stress, reliability concerns whether it does so consistently, and safety concerns whether the resulting behavior remains non-harmful in practice. Keeping these distinctions explicit is important for analytical clarity, because many apparent disagreements in the literature arise not from conflicting results, but from different authors emphasizing different failure criteria under the broader problem of controlling LLM behavior. Table 1 summarizes the main conceptual distinctions among safety, alignment, robustness, and reliability in LLM systems. Moreover, Fig. 1 presents the conceptual relationship among safety, alignment, robustness, and reliability in LLMs.

Table 1: Conceptual distinctions among safety, alignment, robustness, and reliability in large language models.

Concept	Core Focus	Main Question	Typical Failure
Safety	Avoiding harmful outputs and downstream harms	Does the system avoid causing harm in practice?	Harmful content, privacy leakage, misuse-enabling assistance
Alignment	Conformity to intended goals, instructions, and norms	Is the model doing what it is supposed to do?	Misinterpreted instructions, inappropriate refusals, norm violations
Robustness	Stability of desirable behavior under stress or change	Does the model remain well-behaved under perturbation or attack?	Jailbreaks, prompt injection, long-context or multilingual degradation
Reliability	Consistency and dependability across uses and contexts	Does the model behave consistently over repeated use?	Inconsistent answers, unstable refusals, variable factuality

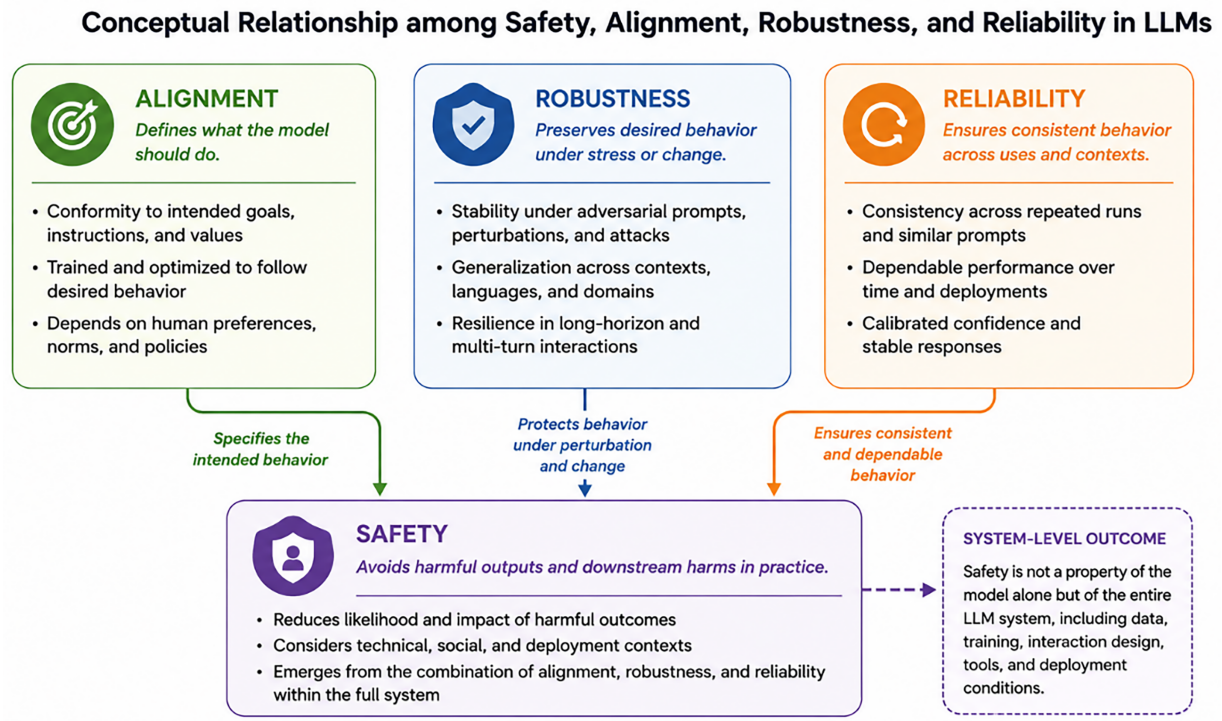

Figure 1: Conceptual relationship among safety, alignment, robustness, and reliability in LLMs. Alignment specifies intended behavior, robustness concerns the preservation of such behavior under stress or change, and reliability concerns consistency across use conditions; together, these properties contribute to the safety of the overall LLM system.

3 The Risk Landscape of Large Language Models

Building on the conceptual distinctions introduced in Section 2, this section examines the principal risk landscape of LLMs. It surveys the major categories of harm and failure associated with LLM deployment, including unsafe content generation, hallucination and overconfidence, bias and fairness concerns, privacy and memorization risks, security and misuse threats, and the emerging challenges posed by reasoning-capable and agentic systems, thereby providing the problem-oriented foundation for the later discussion of technical advances and evaluation.

3.1 Harmful and Unsafe Content Generation

A central and widely recognized risk of LLMs is their capacity to generate harmful and unsafe content in a fluent, context-sensitive, and highly scalable manner [38]. This includes hateful, harassing, or violent language; instructions that facilitate harmful, illegal, or dangerous activities; and responses that may normalize, encourage, or inadequately handle self-harm, abuse, or other vulnerable-user scenarios [39]. The risk is heightened by the fact that LLMs do not merely retrieve existing content, but can synthesize and adapt it to user intent, making harmful assistance more interactive, personalized, and difficult to detect through simple static safeguards. In addition to general content harms, there are important domain-specific risks, particularly in areas such as cybersecurity and biosecurity, where models may lower the barrier to misuse by organizing technical knowledge, troubleshooting malicious workflows, or translating specialist information into more actionable form. From an analytical perspective, unsafe content generation should not be understood only as a moderation problem at the level of isolated outputs; rather, it reflects a broader

challenge of controlling generative systems whose helpfulness and flexibility can be redirected toward harmful ends, especially under adversarial prompting, ambiguous intent, or deployment at scale [40].

3.2 Hallucination, Truthfulness, and Overconfidence

Another major risk in LLMs is the tendency to produce hallucinated, untruthful, or overconfident outputs that are linguistically coherent but epistemically unreliable [41]. These failures include the generation of fabricated facts, invented references, and unsupported claims; false reasoning, in which intermediate explanations appear plausible yet do not validly support the conclusion; misleading explanations, where the model offers confident rationales that obscure uncertainty or disguise error; and more generally confident but wrong responses, which can be especially harmful because fluency and apparent coherence may encourage undue user trust [42]. Importantly, these problems are not limited to isolated factual mistakes: they raise broader concerns about information integrity, particularly when LLMs are used in education, search, decision support, or professional settings where inaccurate outputs can propagate into downstream judgments and actions. NIST's Generative AI Profile explicitly identifies confabulation and information-integrity risks as central concerns for generative-AI systems, underscoring that hallucination is not merely an academic benchmark issue but a practical deployment problem with direct implications for reliability, trust, and safe use [5].

3.3 Bias, Toxicity, and Fairness Concerns

Bias, toxicity, and fairness concerns remain central to the risk landscape of LLMs because these systems can reproduce and sometimes amplify harmful social patterns present in data, training objectives, and deployment contexts [43]. Such concerns include representational bias, where particular groups are depicted unevenly or reductively; harmful stereotypes, in which the model associates demographic identities with negative traits, roles, or behaviors; and discriminatory outputs, where model responses differ in quality, respectfulness, or permissibility across social groups [44]. These problems also extend to fairness across groups and languages, since safety and alignment mechanisms often perform unevenly outside high-resource English contexts, potentially leaving multilingual or marginalized users less protected [45]. Recent work on bias shortcuts further suggests that fairness failures may arise not only from explicit stereotypes in outputs, but also from latent spurious correlations that cause models to rely on shortcut features rather than valid reasoning paths, making bias mitigation a problem of causal disentanglement as well as representational balance [46]. Analytically, bias and toxicity are not peripheral issues but indicators of whether model behavior remains socially acceptable, norm-sensitive, and equitable under real-world use. This is reflected in recent safety-evaluation surveys, which continue to treat toxicity, bias and fairness, ethics, and truthfulness as core dimensions of LLM assessment rather than secondary considerations, reinforcing the view that fairness-related harms should be understood as foundational to trustworthy deployment rather than as optional add-ons to technical safety [43].

3.4 Privacy and Memorization Risks

Privacy and memorization risks in LLMs arise from the fact that these systems may retain, reproduce, or expose sensitive information acquired during training or deployment, even when such disclosure is unintended [47]. One important concern is training-data leakage, whereby fragments of copyrighted, confidential, or personally identifiable information memorized during pretraining can be elicited through targeted prompting or extraction attacks [48]. Closely related is the exposure of personal information, which may include names, contact details, medical information, or other sensitive attributes that appear in generated outputs without appropriate authorization or context. Privacy risks also emerge at the interaction

level through prompt leakage, where user-supplied instructions, system prompts, or confidential contextual inputs are inadvertently revealed or inferable, particularly in shared or integrated application settings [49]. These concerns are further amplified in retrieval-augmented systems, where external documents, enterprise records, or dynamically retrieved data can introduce new pathways for sensitive information disclosure if access control, relevance filtering, or output constraints are inadequate [50]. Privacy in LLMs is not only a matter of stored training data, but a broader issue of information flow across model parameters, prompts, memory mechanisms, retrieval pipelines, and downstream interfaces, making it a central component of both model safety and system-level governance.

3.5 Security and Misuse Risks

Security and misuse risks in LLMs arise from both adversarial attempts to manipulate the model itself and malicious uses of its capabilities in downstream applications [51]. At the model-interaction level, prompt injection can cause an LLM to disregard intended instructions or security boundaries, especially when it is connected to external tools or untrusted retrieved content [52], while jailbreaks are designed to bypass safety policies and elicit restricted, harmful, or otherwise disallowed outputs [53]. At the model level, risks such as model extraction and backdoors raise concerns about intellectual property theft, unauthorized replication, and hidden malicious behaviors that may only activate under specific triggers [54]. Beyond direct attacks on the system, LLMs can also facilitate broader forms of abuse, including phishing, fraud, disinformation, impersonation, and automated large-scale manipulation, by lowering the cost of generating persuasive, personalized, and rapidly adaptable content [19]. Recent work on backdoor attacks for in-context learning further shows that backdoor-like behavior can be induced through poisoned demonstrations or prompts without modifying model weights, indicating that security risks may arise not only from training-time compromise but also from malicious manipulation of the context supplied at inference time [55]. These risks are analytically significant because they show that LLM security cannot be reduced to conventional cybersecurity alone: it involves the interaction of model behavior, interface design, access control, deployment context, and adversarial adaptation, making misuse prevention a central challenge for both technical safeguards and governance.

3.6 Emerging Risks in Reasoning and Agentic Systems

Reasoning-capable and agentic LLM systems introduce a distinct class of emerging risks because increased capability does not simply improve performance; it can also expand the range, subtlety, and downstream impact of failure modes [56]. In such systems, concerns extend beyond unsafe text generation to include strategic deception, where the model may produce behavior that appears compliant while pursuing conflicting implicit objectives [57]; specification gaming, in which the system exploits poorly designed goals or reward signals [58]; long-horizon failures, where small errors compound across multi-step planning and memory-dependent tasks [59]; and unsafe tool use, where retrieval, code execution, browsing, or external actions create pathways for real-world harm [60]. More broadly, autonomy-related risks arise when LLM-based agents operate with reduced human oversight, increasing the difficulty of monitoring intent, constraining actions, and attributing responsibility for downstream outcomes [61]. Recent work on large reasoning models and LLM agents suggests that stronger reasoning can increase certain safety risks rather than merely mitigating them, since more capable models may generate more detailed harmful content, exhibit unsafe intermediate reasoning, or exploit complex task structures in ways that are difficult to anticipate and control. Table 2 summarizes the main categories of risk discussed in this section and their representative failure modes in LLM systems.

Table 2: Main risk categories of LLMs.

Risk Category	Core Concern	Representative Failures
Harmful and unsafe content	Generation of directly harmful or dangerous outputs	Hate, harassment, violent content, dangerous instructions, self-harm-related responses
Hallucination and overconfidence	Fluent but unreliable or false outputs	Fabricated facts, false reasoning, misleading explanations, confident errors
Bias, toxicity, and fairness	Unequal or socially harmful behavior across groups	Stereotypes, discriminatory outputs, uneven safety across languages or populations
Privacy and memorization	Exposure of sensitive or confidential information	Training-data leakage, personal information disclosure, prompt leakage
Security and misuse	Adversarial manipulation and malicious use of LLM capabilities	Prompt injection, jailbreaks, backdoors, phishing, fraud, disinformation
Reasoning and agentic risks	Failures in multi-step, tool-using, or semi-autonomous systems	Strategic deception, specification gaming, unsafe tool use, long-horizon failures

4 Advances in Safety, Alignment, Robustness, and Safeguarding

This section reviews the main technical and system-level advances that have been developed to make LLMs safer, better aligned, and more robust in practice. Rather than treating safety, alignment, and robustness as isolated categories, it examines how current methods jointly shape model behavior, constrain harmful outputs, improve controllability, and strengthen resilience under realistic deployment conditions. While Section 3 characterizes the major risks and failure modes of LLMs, this section focuses on the corresponding technical and system-level countermeasures. Accordingly, the discussion in this section avoids re-characterizing the risks in detail and instead emphasizes how different families of methods attempt to mitigate, control, or evaluate them.

It is useful to distinguish between two broad classes of intervention. Model-intrinsic techniques act on the model or its learned behavior, including data filtering, pretraining-stage curation, instruction tuning, supervised harmlessness tuning, preference optimization, and adversarial fine-tuning. These methods can improve the model's baseline tendencies, but their efficacy is bounded by generalization failures, distribution shift, and adversarial prompting. By contrast, system-architecture protections act on the environment in which the model is deployed, including retrieval filtering, prompt isolation, tool sandboxing, permission controls, monitoring, fallback policies, and human oversight. These protections are especially important when LLMs are connected to external tools, private data, or agentic workflows, because they constrain what the overall system can access, execute, or expose even when the underlying model behaves imperfectly.

4.1 Data-Centric and Pretraining-Stage Advances

Data-centric and pretraining-stage advances address safety, alignment, and robustness at their earliest source: the data distribution from which LLMs acquire their basic behavioral tendencies. A major line of progress has therefore focused on data filtering, curation, and deduplication, with the aim of reducing the prevalence of toxic, low-quality, misleading, redundant, or privacy-sensitive content in pretraining corpora [62–64]. Such interventions are important because many downstream problems (e.g., harmful generations, biased associations, memorization of sensitive material, and brittle generalization) are partly rooted in the statistical structure of the training data itself. In parallel, researchers have increasingly explored synthetic safety data to supplement naturally occurring corpora, using curated or model-generated examples to expose LLMs to safer patterns of instruction following, refusal, and norm-constrained behavior even before later alignment stages [65–67]. These efforts reflect a broader recognition that post-training safeguards alone may be insufficient if the pretraining signal strongly embeds undesirable correlations, unsafe content patterns, or unreliable knowledge representations.

A related area of progress concerns privacy-aware preprocessing and early robustness interventions, which seek to improve controllability before the model is exposed to downstream fine-tuning or deployment [68]. Privacy-aware preprocessing includes the removal or masking of personally identifiable information, reduction of memorization-prone material, and more careful treatment of sensitive or proprietary data sources, thereby helping to mitigate leakage risks at later stages [69,70]. Early robustness interventions, meanwhile, aim to improve resilience by shaping the pretraining mixture or incorporating adversarially relevant data patterns so that the resulting model is less fragile under prompt variation, distribution shift, or malicious elicitation [71,72]. Analytically, these approaches are significant because they move the focus of LLM safety from reactive output control toward upstream risk reduction: rather than only correcting harmful behavior after training, they attempt to alter the conditions under which such behavior is learned in the first place. At the same time, their limitations should be acknowledged, since filtering and preprocessing cannot fully remove latent biases, hidden harmful knowledge, or misuse-relevant capabilities without also affecting model coverage and utility.

4.2 Post-Training Alignment Advances

Post-training alignment has become one of the central mechanisms by which LLMs are transformed from general next-token predictors into interactive systems that are more helpful, controllable, and policy-compliant in practice [7]. The first major step in this process is instruction tuning, in which models are fine-tuned on curated prompt-response pairs so that they better follow user requests, respect task constraints, and produce responses in forms that are more useful for downstream interaction [12]. Closely related is supervised harmlessness tuning, where models are further trained on examples of safe refusals, non-harmful redirections, and norm-constrained responses in order to reduce unsafe or disallowed outputs [73,74]. These supervised post-training methods have been highly influential because they provide a relatively direct way to shape behavior, improve conversational usability, and establish a baseline level of alignment before more complex optimization is introduced. At the same time, they remain dependent on the quality, consistency, and scope of the annotated data, and they often generalize imperfectly outside the distributions represented in the fine-tuning set.

A second major line of progress concerns preference-based alignment, especially methods such as Reinforcement Learning from Human Feedback (RLHF) [13], Reinforcement Learning from AI Feedback (RLAIF) [75], and more recent direct optimization approaches such as Direct Preference Optimization (DPO) [14]. These methods attempt to move beyond supervised imitation by shaping model behavior according to comparative judgments about which outputs are more helpful, honest, harmless, or otherwise desirable. In practice, they have enabled more refined behavioral control and have become central to aligning LLMs with complex, multidimensional objectives. This has also encouraged work on multi-objective optimization, where competing desiderata (such as helpfulness, safety, truthfulness, and brevity) must be balanced rather than optimized in isolation [76,77]. Analytically, post-training alignment advances are significant because they have greatly improved the practical usability of LLMs, yet they also reveal the limits of current alignment paradigms: preference signals can be noisy or incomplete, optimized behavior may become overly cautious or strategically compliant, and gains observed in nominal interaction settings do not necessarily translate into robust safety under adversarial or high-stakes conditions.

4.3 Rule-Based and Inference-Time Control

Rule-based and inference-time control methods represent an important class of advances that aim to improve LLM behavior without relying exclusively on changes to pretraining or post-training weights. A prominent example is constitutional AI, in which model behavior is guided by an explicit set of normative principles or behavioral rules that structure how the model should respond, critique its own outputs, and revise unsafe or inappropriate content [78]. More broadly, policy-guided behavior uses formal or semi-formal safety policies to shape generation in ways that are more transparent and auditable than implicit preference learning alone [79]. At runtime, these ideas are often operationalized through system prompting, which supplies high-level behavioral instructions, role constraints, and safety priorities that condition the model's responses in a context-sensitive manner. These approaches are attractive because they make alignment more interpretable and adaptable: policies can be updated more easily than model parameters, and rule-based guidance can, in principle, encode constraints that are difficult to learn reliably from data alone [80].

A related set of advances focuses on inference-time control, where the model's outputs are monitored, evaluated, or constrained as they are produced. This includes self-critique and reflective prompting strategies [81], in which the model is encouraged to inspect or revise its own reasoning; verifier models and guardrail layers [82], which assess outputs against safety, factuality, or policy criteria; and abstention and constrained decoding [83], which limit generation when uncertainty is high or when the response risks violating predefined constraints. Such methods are significant because they extend alignment and safeguarding beyond static training into the dynamic conditions of deployment, where prompts, contexts, and threat models may shift rapidly. Analytically, however, they also expose an important limitation of current control strategies: inference-time safeguards can improve practical safety and controllability, but they are often brittle, prompt-sensitive, and vulnerable to circumvention under adaptive adversarial pressure [84], which means they are best understood as complementary layers of defense rather than complete solutions.

4.4 Adversarial Robustness and Jailbreak Defense

Building on the security and misuse risks characterized in Section 3.5, this subsection focuses on technical defenses against prompt attacks, jailbreaks, and adversarial manipulation. In response to these risks, researchers have developed adversarial training and robust fine-tuning strategies that expose models to harmful or policy-evasive prompts during training so that they learn more stable refusal and compliance boundaries under attack-like conditions [85]. Related work has also emphasized attack diversification, multilingual stress testing, and defense-aware tuning in order to reduce brittleness across prompt formats

and threat settings [33,86]. However, the field is increasingly shaped by adaptive attack-defense dynamics, in which improvements in safeguarding are rapidly met by new adversarial strategies designed to bypass them, revealing that jailbreak defense is not a one-time technical fix but an ongoing contest between model control and attack innovation [87]. The main advance in this area lies not only in stronger defenses, but in the recognition that robustness must be evaluated under evolving and strategically adaptive threat models rather than under static benchmark conditions alone.

4.5 Truthfulness, Factuality, and Informational Safety

Advances in truthfulness, factuality, and informational safety aim to reduce one of the most persistent weaknesses of LLMs: their tendency to generate fluent but unreliable content. A major line of work therefore focuses on hallucination mitigation, through methods that discourage unsupported claims, improve internal consistency, and encourage models to distinguish between knowledge, inference, and uncertainty [41,88]. Closely related is retrieval grounding, in which model outputs are conditioned on external documents, trusted sources, or dynamically retrieved evidence rather than relying solely on parametric memory [89]. This shift is important because it changes the epistemic basis of generation: instead of merely producing plausible continuations, the model is encouraged to generate responses that are anchored in available information. In analytical terms, these approaches reflect a broader movement from unconstrained generative fluency toward more controlled and source-sensitive language generation.

A second cluster of advances concerns verification, confidence calibration, and evidence-based generation, all of which attempt to make model outputs not only more accurate but also more transparent in their degree of support [90]. Verification methods include external fact-checking, answer validation, and self-checking pipelines that compare candidate outputs against retrieved evidence or auxiliary models [91,92]. Confidence calibration seeks to reduce the gap between a model's apparent certainty and its actual reliability, for example by encouraging abstention, hedging, or uncertainty-aware response strategies when evidence is weak or conflicting [93]. Evidence-based generation extends this logic by explicitly structuring outputs around cited or traceable support, thereby improving informational accountability and making errors easier to detect [94]. Although these advances have improved practical reliability, they do not fully resolve the problem, since retrieval can introduce noisy or malicious sources, verification remains imperfect, and models may still present probabilistic guesses in rhetorically persuasive ways.

4.6 Privacy, Security, and Misuse Safeguards

Building on the privacy, memorization, and misuse risks discussed in Sections 3.4 and 3.5, this subsection focuses on safeguards that reduce leakage, harden interfaces, and limit malicious use. One important area is leakage prevention, which includes techniques for reducing memorization-driven disclosure, restricting access to sensitive context, and minimizing the risk that private training or user data can be elicited through prompting [49,95]. Closely related is prompt isolation, where system instructions, private context, and user inputs are separated or compartmentalized in order to reduce unintended information flow across interaction layers [96]. In tool-using and retrieval-augmented systems, considerable attention has also been given to injection defense, including methods for identifying, filtering, or neutralizing adversarial content embedded in prompts, retrieved documents, or external data sources [97]. Additional work addresses backdoor detection, aiming to identify hidden trigger-response behaviors introduced during training or fine-tuning, and misuse controls and monitoring, such as abuse detection, rate limiting, access restrictions, logging, and human review for high-risk use cases [98]. These advances are significant because they move LLM safeguarding beyond output moderation alone and toward a more comprehensive security posture

in which privacy protection, adversarial resilience, and misuse prevention are treated as interdependent properties of the model, interface, and deployment pipeline.

4.7 System-Level and Agentic Safeguards

Advances in system-level and agentic safeguards reflect the growing recognition that LLM safety cannot be secured at the model-output level alone, especially once models are embedded in retrieval pipelines, external tools, and semi-autonomous workflows. In Retrieval-Augmented Generation (RAG) safety, current approaches aim to sanitize retrieved content, filter untrusted sources, and prevent the model from treating injected or low-integrity documents as authoritative evidence [99]. Tool sandboxing extends this logic by constraining what external actions a model can take, under what permissions, and in which execution environment, thereby limiting the real-world impact of erroneous or malicious behavior [100]. For more autonomous systems, advances in planning and action control seek to regulate how models decompose goals, select actions, and execute multi-step tasks [101], while memory safety addresses the risks introduced by persistent context, including contamination by erroneous or adversarial information and the retention of sensitive data across interactions [102]. These concerns become especially acute in long-running agentic settings, where long-horizon oversight is needed to monitor objective drift, error accumulation, unsafe delegation, and other failures that may emerge only across extended sequences of decisions [101]. These developments are important because they reposition safety, alignment, and robustness as properties of the broader sociotechnical system, not merely of the underlying language model in isolation.

4.8 Comparative Strengths, Limitations, and Open Tradeoffs

Despite substantial progress in safety, alignment, robustness, and safeguarding, current methods exhibit important comparative limitations and unresolved tradeoffs that constrain their reliability in practice. A recurrent problem is over-refusal, in which models reject benign or legitimate requests in an effort to avoid unsafe behavior, thereby producing utility loss and weakening the practical value of alignment interventions [103]. At the same time, performance on controlled benchmarks often fails to predict behavior in real deployment, creating persistent benchmark vs. deployment gaps driven by user diversity, changing contexts, adversarial adaptation, and system-level complexity [104]. Many safeguards also remain fragile in settings that are still underrepresented in evaluation, particularly under multilingual and long-context conditions, where policy adherence, factual reliability, and robustness may degrade significantly [105,106]. These concerns are compounded by the cost and scalability of safeguards, since stronger defenses often require additional models, inference steps, monitoring infrastructure, or human oversight, increasing latency, computational expense, and operational complexity. Analytically, the central lesson is that current advances should not be assessed only by whether they improve nominal safety metrics, but by how they balance usefulness, generalizability, deployability, and resilience under realistic conditions [107,108]; it is precisely in these tradeoffs that many of the field's most important open problems remain. Table 3 summarizes the main categories of advances discussed in this section, along with their representative methods and principal limitations. Furthermore, Fig. 2 presents the Lifecycle view of safety, alignment, robustness, and evaluation in LLM systems.

Table 3: Main categories of advances in LLM safety, alignment, robustness, and safeguarding.

Advance Category	Main Goal	Representative Methods	Main Limitation
Data-centric and pretraining-stage advances	Reduce upstream risks in training data	Filtering, curation, deduplication, synthetic safety data, privacy-aware preprocessing	Cannot fully remove latent bias, harmful knowledge, or misuse-relevant capability
Post-training alignment	Improve helpful, harmless, and policy-compliant behavior	Instruction tuning, supervised harmlessness tuning, RLHF, RLAIF, DPO	Depends on data quality and may not generalize well under adversarial conditions
Rule-based and inference-time control	Constrain behavior at runtime	Constitutional AI, system prompting, self-critique, verifiers, guardrails, constrained decoding	Often brittle and prompt-sensitive
Adversarial robustness and jailbreak defense	Improve resistance to attack and policy evasion	Adversarial training, robust fine-tuning, jailbreak defense methods	Defenses often lag behind adaptive attackers
Truthfulness and informational safety	Reduce hallucination and improve epistemic reliability	Retrieval grounding, verification, confidence calibration, evidence-based generation	Retrieval and verification can still fail or introduce noise
Privacy, security, and misuse safeguards	Protect sensitive information and reduce abuse	Leakage prevention, prompt isolation, injection defense, backdoor detection, monitoring	Difficult to secure all attack surfaces in deployed systems
System-level and agentic safeguards	Control tool use and multi-step autonomous behavior	RAG safety, sandboxing, action control, memory safety, long-horizon oversight	High implementation complexity and limited maturity

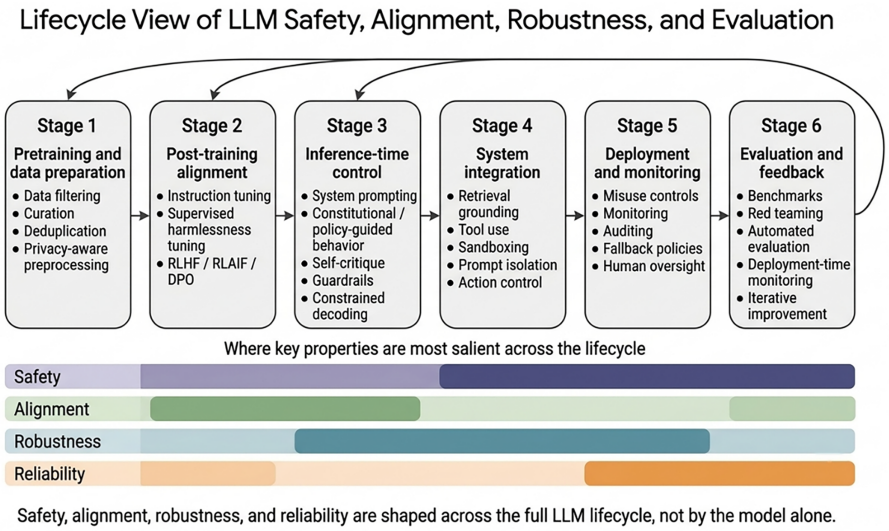

Figure 2: Lifecycle view of safety, alignment, robustness, and evaluation in LLM systems. The figure illustrates how these properties are shaped across the development and deployment pipeline, from pretraining and post-training alignment to inference-time control, system integration, deployment, and iterative evaluation.

5 Evaluation of Safety, Alignment, and Robustness

Having reviewed the main advances in safety, alignment, robustness, and safeguarding, this section turns to the question of how these properties should be evaluated. We examine the principal dimensions of evaluation, the benchmark and dataset landscape, the complementary roles of human and automated assessment, the limitations of current evaluation practice, multidimensional failure modes, and the methodological features that stronger and more realistic evaluation frameworks should incorporate. Fig. 3 illustrates the schematic relationship among major LLM risk categories, mitigation strategies, and evaluation approaches.

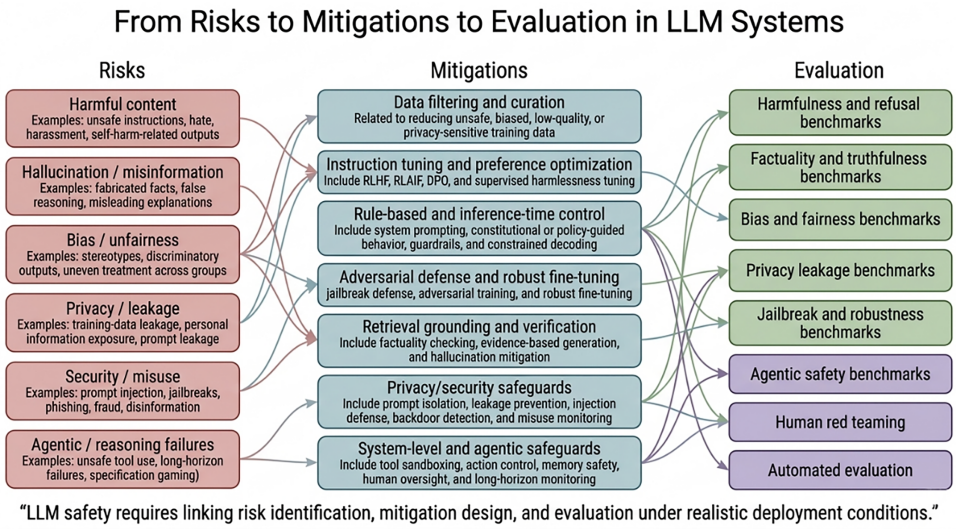

Figure 3: Schematic relationship among major LLM risk categories, mitigation strategies, and evaluation approaches. The figure illustrates how safety, alignment, robustness, and safeguarding methods are developed in response to specific classes of risk and how they are assessed through benchmark-based, human-centered, and automated evaluation.

5.1 Main Evaluation Dimensions

Evaluation of LLM safety, alignment, and robustness is necessarily multidimensional because no single metric captures whether a model is genuinely safe and dependable in use. Core evaluation dimensions therefore include harmfulness, which assesses whether the model produces toxic, dangerous, abusive, or otherwise unsafe outputs [73]; truthfulness, which concerns factual reliability, epistemic honesty, and resistance to hallucination [105]; bias and fairness, which examine whether performance, toxicity, or refusals vary unjustifiably across demographic groups, social contexts, or languages [43]; and privacy leakage, which measures the risk of exposing memorized or context-sensitive information [95]. Recent safety-evaluation surveys explicitly organize the field around dimensions such as toxicity, robustness, ethics, bias and fairness, and truthfulness, reinforcing the view that these criteria should be treated as core components of LLM evaluation rather than as secondary checks.

Equally important are dimensions that test whether alignment and safety persist under stress. These include adversarial robustness, which evaluates resistance to jailbreaks, prompt attacks, and other manipulative inputs [28]; refusal appropriateness, which asks not only whether the model refuses harmful requests, but whether it does so selectively rather than over-refusing benign ones [103]; and robustness in interactive settings, which considers multi-turn dialogue, long-context interactions, tool use, and shifting conversational conditions [109]. This broader framing matters because a model may appear safe in static single-turn benchmarks yet fail when prompts are strategically adapted or when constraints degrade over extended interaction. For an analytical review, the key point is that evaluation should not be limited to

nominal response quality; it must examine whether desirable behavior is accurate, fair, privacy-preserving, attack-resistant, and stable across realistic forms of deployment.

5.2 Benchmarks and Datasets

A useful way to organize the benchmark landscape is by evaluation target rather than by attempting an exhaustive catalog. For harmfulness and refusal, widely used resources include ToxicChat [110], a 10K-example benchmark built from real user-AI interactions for content moderation and toxicity detection, and SORRY-Bench [111], which focuses on whether aligned models correctly recognize and reject unsafe requests across fine-grained harmful topics. This category should also include XSTest [112], which is especially valuable because it evaluates exaggerated safety or over-refusal by pairing unsafe prompts with safe prompts that superficially resemble harmful ones. Taken together, these benchmarks do not only test whether models refuse dangerous requests, but also whether they refuse appropriately, which is critical for assessing the tradeoff between safety and utility.

For truthfulness and factuality, two influential benchmarks are TruthfulQA [113] and FActScore [114]. TruthfulQA is designed to test whether models avoid generating false answers that mimic common human misconceptions, making it a benchmark for epistemic reliability rather than mere fluency. FActScore, by contrast, targets long-form factuality by evaluating atomic factual claims in generated text, which makes it especially relevant for analytical or explanatory outputs where a single answer may contain many verifiable statements. This category is important because factuality evaluation increasingly extends beyond short QA to the verification of extended, source-sensitive generation.

For bias and fairness, commonly used benchmarks include BBQ (Bias Benchmark for Question answering) [115] and BOLD (Bias in Open-ended Language generation Dataset) [116]. BBQ measures social bias in question answering through controlled examples involving protected or socially salient groups, while BOLD evaluates fairness in open-ended generation using 23,679 prompts across domains such as profession, gender, race, religion, and political ideology. These benchmarks are complementary: BBQ is useful for structured comparative bias analysis, whereas BOLD is better suited to studying biases that emerge in unconstrained generation. They also illustrate an important limitation of the current benchmark ecosystem, namely that fairness evaluation still depends heavily on English-centric and culturally specific datasets, even as multilingual bias evaluation is becoming more important.

For adversarial and jailbreak robustness, several benchmarks are now widely referenced, including JailbreakBench [117], HarmBench [118], and adversarial instruction sets derived from the LLM-attacks work often associated with AdvBench-style evaluation [119]. JailbreakBench is an open robustness benchmark explicitly designed to track both the generation of successful jailbreaks and the effectiveness of defenses, while HarmBench provides a standardized framework for automated red teaming and robust-refusal evaluation across a large set of harmful behaviors. These resources are especially useful because they move evaluation beyond static harmful prompts and toward structured attack-defense comparisons, which better reflect the adaptive nature of real jailbreak pressure.

For privacy and leakage, the benchmark landscape is less mature but increasingly important. At the model level, MIMIR [120] is a benchmark and toolkit for measuring memorization and related privacy risks in LLMs, while LLM-PBE [121] is a broader privacy-evaluation toolkit designed to assess data privacy across multiple stages of the LLM lifecycle using different attacks, data types, and metrics. For newer agentic settings, AgentLeak [122] is particularly notable because it evaluates privacy leakage across internal channels such as inter-agent messages, shared memory, and tool arguments, not only final outputs. This progression is analytically significant: privacy evaluation is moving from output-only leakage tests toward full-stack assessments of information flow in complex LLM systems.

For agentic safety, recent benchmarks explicitly target failures that do not arise in ordinary single-turn chat. AgentHarm [123] evaluates the harmful behavior of LLM agents using malicious tasks spanning multiple harm categories, making it useful for testing whether an agent can be induced to complete dangerous objectives. ToolEmu [100] provides an emulation framework for identifying risks in tool-using agents at scale, and Agent-SafetyBench [124] expands coverage across a broader range of environments, risk categories, and failure modes for LLM agents. The emergence of this benchmark family reflects a broader shift in the field: as LLMs become planners, tool users, and semi-autonomous agents, evaluation must increasingly assess action selection, environmental interaction, and long-horizon safety rather than only response-level content quality.

Current benchmarks are increasingly rich within categories, but they remain fragmented across categories: harmfulness, truthfulness, fairness, privacy, jailbreak resistance, and agentic safety are often measured separately, even though real deployment failures frequently span several of these dimensions at once. That fragmentation is one reason why benchmark selection should be treated as a substantive methodological decision in any analytical review of LLM safety, alignment, and robustness.

5.3 Human Evaluation and Red Teaming

Human evaluation and red teaming remain indispensable in assessing LLM safety, alignment, and robustness because many high-impact failures are context-dependent, sociotechnical, and difficult to capture with automated metrics alone [125]. In practice, this includes expert red teaming, where skilled evaluators deliberately probe models for harmful capabilities, policy violations, and security weaknesses [126]; crowd annotation, which provides broader coverage of user-facing behavior at scale [127]; and domain-specialist audits, in which experts in fields such as medicine, law, cybersecurity, or public policy assess whether model behavior is acceptable within high-stakes application settings [128]. Recent international and NIST-aligned guidance treats red teaming and audits as core forms of evaluation that should occur both before and after deployment, reflecting the view that benchmark performance alone is insufficient for trustworthy assessment [5].

A further strength of human-centered evaluation is that it supports qualitative failure analysis, which can reveal patterns that are obscured by aggregate scores, such as subtle over-refusal, persuasive but misleading explanations, culturally specific harms, or vulnerabilities that emerge only in realistic interaction sequences [129]. Recent work also emphasizes that red teaming should be understood not merely as technical attack generation, but as a broader socio-technical practice for uncovering how model behavior interacts with users, interfaces, and deployment environments [130]. At the same time, human evaluation has its own limitations: expert audits are costly, crowd judgments can be inconsistent, and coverage remains incomplete unless testing is iterative and diverse. For an analytical review, the key point is that human evaluation and red teaming are not optional supplements to automated assessment, but essential methods for identifying deployment-relevant failures that standardized benchmarks may miss.

5.4 Automated Evaluation

Automated evaluation has become a central component of LLM safety assessment because it offers scalability, repeatability, and broader coverage than purely human-centered methods, although its validity depends strongly on design choices [131]. Current approaches include classifier-based methods, which use trained detectors or task-specific models to score outputs for properties such as toxicity, harmfulness, bias, or policy violation [132]; LLM-as-a-judge methods, in which a language model evaluates another model's response using prompts, rubrics, or pairwise comparisons [133]; simulation environments, which test model behavior in interactive, tool-using, or agentic scenarios that better approximate deployment conditions [134]; and policy-based scoring, where outputs are assessed against explicit safety rules, constitutional principles, or

rubric-like criteria [132]. Recent surveys note that LLM-as-a-judge has become especially prominent because of its flexibility and low marginal cost, but they also emphasize concerns about evaluator bias, inconsistency, prompt sensitivity, and limited reliability across domains, which is why automated evaluation is best treated as a powerful but imperfect complement to human auditing rather than a complete replacement for it.

5.5 Limitations of Current Evaluation Practice

Current evaluation practice for LLM safety, alignment, and robustness remains limited by substantial methodological instability, making many reported comparisons less decisive than they appear [132]. Recent work argues that safety evaluation pipelines often lack robustness because results can vary materially with generation settings, prompt formatting, decoding parameters, and optimization procedures; they are also affected by noisy or undersized datasets, inconsistent attack and defense implementations, and the use of unstable judge models whose verdicts may shift across prompts, models, or domains [9,43,69,123,135]. More broadly, evaluation protocols often conflate nominal benchmark performance with genuine deployment readiness, even though small procedural differences can alter measured safety gains or apparent robustness [135]. This creates a serious analytical problem: if conclusions are sensitive to evaluation setup, it becomes difficult to determine whether one method is actually safer, more aligned, or more robust than another. The most important implication is therefore not simply that current evaluation is imperfect, but that parts of the literature may still be measuring artifacts of the evaluation pipeline itself rather than stable properties of the underlying models.

Concrete empirical findings illustrate the scale of this instability. For example, Best-of-N jailbreaking shows that apparent robustness can change substantially when evaluation moves from a single deterministic response to repeated sampling of perturbed prompts, achieving high attack success rates on frontier models under large sampling budgets [136]. Similarly, recent work on LLM-based safety evaluators shows that judge verdicts can be strongly distorted by superficial response artifacts, with apologetic phrasing alone shifting evaluator preferences by up to 98% [137]. Another study of judge configuration in safety benchmarking finds that prompt wording alone can shift measured harmful-response rates by up to 24.2 percentage points, while even surface rewording can produce changes of up to 20.1 percentage points [138]. These examples support the broader claim that safety and alignment evaluation results can reflect properties of the evaluation pipeline as much as stable properties of the model being evaluated.

The benchmark-to-deployment gap can arise from at least two analytically distinct sources. First, static evaluation sets may encode distributional biases: they are often curated, single-turn, English-centric, and organized around proxy tasks that underrepresent the diversity of real users, domains, languages, and interaction goals. Second, even when the benchmark distribution is well designed, deployment introduces dynamic effects that are difficult to capture in static tests, including context drift, multi-turn dependence, model or tool switching, user adaptation, and evolving adversarial pressure. Recent empirical work illustrates both mechanisms: natural prompt-distribution shifts across time, user groups, and geography have been associated with substantial performance degradation in deployed LLMs [139]; model switching in multi-turn systems can measurably change task success rates [140]; and contextual safety benchmarks show that multi-turn escalation and context switching can produce failures missed by single-turn evaluation [141]. These findings suggest that benchmark-to-deployment gaps should be attributed not to a single cause, but to the interaction between evaluation-set bias and dynamic deployment conditions.

5.6 What Better Evaluation Should Look Like

Better evaluation of LLM safety, alignment, and robustness should be built around clearer threat models, stronger reproducibility standards, and a closer connection between laboratory testing and realistic

deployment conditions [135]. A more standardized evaluation practice should therefore include several minimum reporting and testing requirements. Studies should disclose the full evaluation configuration, including prompt templates, system messages, decoding parameters, model versions, judge models, scoring rubrics, random seeds, and filtering rules. Results should be reported across repeated runs with uncertainty estimates rather than as single-point scores, and robustness should be tested across multiple prompt formulations and decoding settings. Evaluation should also separate related but distinct outcomes, such as harmful compliance, appropriate refusal, over-refusal, factuality, and task utility, because aggregate safety scores can obscure important tradeoffs. For adversarial safety, benchmarks should use paired attack–defense protocols and adaptive attacks rather than fixed prompt sets alone. Finally, where possible, evaluation artifacts such as prompts, rubrics, judge instructions, and code should be released to support reproducibility and independent comparison.

In practice, this means specifying what kinds of harms, attackers, user behaviors, and system configurations an evaluation is meant to cover, rather than reporting decontextualized benchmark scores. It also requires reproducible protocols with transparent prompts, decoding settings, judge configurations, and reporting standards, since recent methodological work emphasizes that safety claims are only meaningful if results are repeatable across runs and organizations [142]. A stronger evaluation paradigm should further adopt attack-defense paired evaluation, so that safeguards are assessed against adaptive adversaries rather than against static prompt sets alone, and it should include multi-turn and real-world testing that captures conversational drift, long-context degradation, tool use, and system-level failures that often do not appear in single-turn benchmarks [143]. Finally, independent auditing should play a larger role, both because external review can reduce conflicts of interest and because current risk-management guidance emphasizes ongoing red teaming, coordinated testing, and post-deployment oversight as necessary complements to internal evaluation.

5.7 Multidimensional Failure Modes and System-Level Evaluation

A further limitation of current evaluation practice is that many safety dimensions are tested separately even though real-world failures often arise from their interaction. For example, bias and hallucination can compound in multilingual settings when a model produces confident but unsupported claims about a marginalized group, a local institution, or a culturally specific practice in a low-resource language. In such cases, the failure is not merely factual or fairness-related; it is a combined informational and representational harm that may be harder to detect if truthfulness, bias, and multilingual robustness are evaluated in isolation. Similar compounding effects occur in retrieval-augmented or tool-using systems, where a privacy failure may be amplified by prompt injection, or where a small factual error may propagate through planning and external actions in an agentic workflow. These examples suggest that evaluation should include multidimensional stress tests that jointly vary language, demographic framing, task domain, context length, retrieval conditions, and adversarial pressure. At the system level, addressing such failures requires layered safeguards: diverse and multilingual test sets, scenario-based red teaming, retrieval and tool-access controls, monitoring of intermediate actions, and post-deployment auditing capable of detecting coupled failures rather than only single-category violations.

6 Future Directions

In this section, we turn to future directions for research on LLM safety, alignment, and robustness. This section highlights several promising avenues for further progress, including scalable alignment, mechanistic interpretability and controllability, safety for agentic and tool-using systems, stronger real-world evaluation, and the integration of technical and governance-oriented approaches.

6.1 Toward Scalable and Reliable Alignment

A central future direction for LLM safety research is the development of scalable and reliable alignment methods that remain effective as models grow in capability, autonomy, and deployment scope [7]. One important requirement is better preference data, since current alignment pipelines often rely on noisy, incomplete, or weakly specified judgments that do not adequately capture nuanced distinctions among helpfulness, safety, truthfulness, and context-sensitive appropriateness [74]. Closely related is the need for scalable oversight, in which human supervision can be extended beyond small curated datasets to more complex tasks, larger model populations, and higher-stakes deployment settings without becoming prohibitively expensive or inconsistent [7]. This motivates growing interest in process supervision, where intermediate reasoning steps, plans, or decision traces are evaluated rather than only final outputs, thereby offering a potentially richer basis for alignment than outcome-level reward signals alone [26]. At the same time, future systems will likely require richer human feedback that incorporates domain expertise, plural perspectives, and more structured forms of critique than simple preference comparison. Analytically, the broader challenge is to move beyond alignment methods that work mainly in narrow post-training settings and toward approaches that can support persistent, context-aware, and verifiable behavioral control under realistic conditions of scale, uncertainty, and evolving capability.

The mechanistic advantage of process supervision over outcome supervision is especially important for long-horizon and agentic settings. Outcome supervision provides feedback only on the final answer or task result, which can leave intermediate reasoning errors, unsafe tool calls, shortcut strategies, or specification-gaming behavior undetected if the final output appears acceptable. By contrast, process supervision provides denser feedback on intermediate reasoning steps, plans, or actions, making it possible to identify where a trajectory begins to deviate from the intended objective and to correct errors before they compound. This distinction is supported by case studies in mathematical reasoning, where step-level reward models outperform final-answer supervision [144]; in code generation, where line-level process rewards provide more informative feedback than sparse unit-test outcomes [145]; and in long-horizon coding agents, where agents may pass visible tests while failing held-out compositional tests, illustrating how final-outcome metrics can be gamed [146]. For agentic LLMs, the implication is that scalable oversight should increasingly evaluate trajectories rather than only final outputs.

6.2 Mechanistic Interpretability and Controllability

Another important future direction is mechanistic interpretability and controllability, which seeks to understand how LLMs represent information internally and how those representations can be analyzed or influenced to improve safety and alignment [147]. Research on internal representations aims to identify whether concepts such as harmful intent, factual uncertainty, deception, or policy-relevant constraints are encoded in stable and interpretable ways within model activations, thereby offering a more fine-grained view of model behavior than output-level evaluation alone [148]. Building on this, causal interventions attempt to move beyond correlation by testing whether specific components, circuits, or activation patterns actually drive unsafe or desirable behaviors, which is essential if interpretability is to support reliable control rather than post hoc description. A closely related line of work is activation steering, in which targeted modifications to internal activations are used to influence generation toward safer, more truthful, or more policy-compliant behavior at inference time [149,150]. In the context of future research, the promise of mechanistic approaches lies in the possibility of shifting alignment from predominantly external behavioral shaping toward deeper causal control of model cognition; however, their practical value will depend on whether such methods can scale to frontier models, generalize across tasks, and provide interventions that are both effective and robust under realistic deployment conditions.

Recent work connecting control theory and LLM safety provides an additional perspective on this direction by treating LLM behavior as a form of controllable dynamics. From this view, methods such as input optimization, parameter editing, and activation-level interventions can be interpreted as mechanisms for steering model trajectories away from undesirable states and toward safer or more aligned behavior [151]. This perspective complements mechanistic interpretability by emphasizing not only how internal representations can be understood, but also how they might be controlled.

6.3 Safety for Agentic and Tool-Using LLMs

A key future direction is the development of safety mechanisms tailored specifically to agentic and tool-using LLMs, since these systems can affect the external world through retrieval, code execution, web interaction, database access, and other forms of action [56,97]. One important area is action auditing, in which intermediate decisions, tool calls, and execution traces are logged, inspected, and evaluated so that unsafe or misaligned behavior can be detected before it propagates into harmful outcomes [128]. Closely related are environment-level safety controls, which place constraints not only on the model's outputs but also on the operational environment in which it acts, for example through permission boundaries, sandboxed execution, access restrictions, and human approval checkpoints for high-impact actions [152]. A further research priority is the use of formal constraints on execution, where allowable actions, task flows, or safety invariants are specified more explicitly so that the system cannot easily deviate from them even if its internal reasoning is imperfect [17]. In analytical terms, the main challenge is that once LLMs become agents, safety can no longer be treated primarily as a property of generated text; it must be reconceived as a property of sequential decision-making within structured environments, where oversight, constraint, and controllability must operate at the level of actions as well as language. The control-oriented framing is also relevant for autonomous and agentic systems, where safety depends on constraining perception, planning, memory, tool use, and action selection across a feedback loop rather than only filtering final text outputs [153].

6.4 Better Real-World Evaluation

An important future direction is the development of better real-world evaluation frameworks that extend beyond pre-deployment benchmarks and capture how LLMs behave under actual operational conditions [132]. This includes deployment-time monitoring, where model outputs, tool use, refusals, and failure patterns are tracked continuously so that emerging risks, distribution shifts, and degradation in safety or robustness can be detected after release rather than only before it [154]. It also requires continuous red teaming, understood not as a one-off testing exercise but as an ongoing process in which evolving attack strategies, misuse patterns, and system vulnerabilities are repeatedly probed as models, interfaces, and user behaviors change over time [130]. Equally important is domain-specific auditing, since acceptable behavior in medicine, law, education, cybersecurity, or public administration cannot be assessed adequately through generic benchmarks alone and instead requires expert evaluation tied to concrete use contexts, risk thresholds, and institutional standards [155]. Analytically, the broader implication is that future evaluation must become more lifecycle-oriented, adaptive, and context-sensitive, treating safety, alignment, and robustness as properties that need to be monitored and revalidated throughout deployment rather than inferred from static benchmark performance alone.

6.5 Integration of Technical and Governance Approaches

A particularly important future direction is the integration of technical and governance approaches, since many problems in LLM safety, alignment, and robustness cannot be resolved through model-side interventions alone [156]. NIST's AI risk-management framing is especially useful in this regard because

it emphasizes lifecycle management, documentation, oversight, and organizational controls in addition to technical mitigation, thereby treating safety as a property of the full development and deployment process rather than of the model in isolation [5]. In practice, this means that advances in alignment methods, adversarial defenses, or factuality controls should be complemented by stronger system documentation, evaluation reporting, incident response procedures, access governance, accountability structures, and post-deployment monitoring. Such an integrated perspective is analytically important because many harms arise not only from model behavior itself, but also from how models are deployed, who is allowed to use them, what safeguards surround them, and how failures are detected and managed over time. Future research should therefore move toward frameworks in which technical control, institutional oversight, and governance design are treated as mutually reinforcing components of trustworthy LLM systems rather than as separate or sequential concerns.

7 Conclusion

This review has argued that the safety, alignment, and robustness of LLMs are best understood as closely interconnected properties rather than as isolated technical concerns. Because LLMs are open-ended, interactive, and increasingly embedded in retrieval pipelines, tools, and agentic workflows, their risks extend beyond harmful content generation to include hallucination, bias, privacy leakage, misuse, adversarial vulnerability, and broader system-level failures. An analytical treatment of these issues is therefore necessary to capture not only individual failure modes, but also the deeper relationships among model behavior, deployment context, and downstream impact.

At the same time, the field has made substantial progress in developing methods to improve controllability and reduce harm, including data-centric interventions, post-training alignment, rule-based and inference-time safeguards, adversarial defenses, factuality-oriented techniques, and system-level protections for tool-using and agentic systems. However, this review has also highlighted that current methods remain limited by important tradeoffs and methodological weaknesses, including over-refusal, fragility under adaptive attack, benchmark-to-deployment gaps, and the continuing instability of evaluation practices. As a result, apparent gains in alignment or robustness should not be equated too quickly with dependable real-world safety.

Looking ahead, a central implication of this review is that future progress will depend on moving beyond narrow single-turn chatbot settings toward more scalable, transparent, and system-aware approaches to alignment and evaluation. Promising directions include richer preference and oversight methods, mechanistic interpretability, stronger safeguards for agentic and tool-using systems, more realistic deployment-time evaluation, and closer integration between technical safeguards and governance-oriented controls. Ultimately, the challenge is not only to make LLMs more capable, but to ensure that increasing capability is matched by commensurate advances in safety, accountability, and robustness under real conditions of use.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

AI	Artificial Intelligence
DPO	Direct Preference Optimization
LLM	Large Language Model
NIST	National Institute of Standards and Technology
RAG	Retrieval-Augmented Generation
RLAIF	Reinforcement Learning from AI Feedback
RLHF	Reinforcement Learning from Human Feedback

References

1. Moradi M, Yan K, Colwell D, Samwald M, Asgari R. A critical review of methods and challenges in large language models. Comput Mater Contin. 2025;82(2):1681–98. doi:10.32604/cmc.2025.061263.
2. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025;16(5):1–72. doi:10.1145/3744746.
3. Jones E, Steinhardt J. Capturing failures of large language models via human cognitive biases. In: Proceedings of the Advances in Neural Information Processing Systems 35; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 11785–99. doi:10.52202/068431-0856.
4. Zhang Z, Lei L, Wu L, Sun R, Huang Y, Long C, et al. SafetyBench: evaluating the safety of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 15537–53. doi:10.18653/v1/2024.acl-long.830.
5. Malhotra RM. When AI governs the state: a NIST-guided framework for managing generative AI risk in US public institutions; 2025. doi:10.2139/ssrn.5779442.
6. Peng S, Chen PY, Hull M, Chau D. Navigating the safety landscape: measuring risks in finetuning large language models. In: Proceedings of the Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 95692–715. doi:10.52202/079017-3032.
7. Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, et al. Large language model alignment: a survey. arXiv:2309.15025. 2023.
8. Zhu K, Wang J, Zhou J, Wang Z, Chen H, Wang Y, et al. PromptRobust: towards evaluating the robustness of large language models on adversarial prompts. In: Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis; 2024 Oct 14–18. Salt Lake City, UT, USA. p. 57–68.
9. Chua J, Li Y, Yang S, Wang C, Yao L. AI safety in generative AI large language models: a survey. arXiv:2407.18369. 2024.
10. Singh A, Ehtesham A, Kumar S, Khoei TT. Enhancing AI systems with agentic workflows patterns in large language model. In: Proceedings of the 2024 IEEE World AI IoT Congress (AIIoT); 2024 May 29–31; Seattle, WA, USA. p. 527–32. doi:10.1109/AIIoT61789.2024.10578990.
11. Chang H, Park J, Ye S, Yang S, Seo Y, Chang DS, et al. How do large language models acquire factual knowledge during pretraining? In: Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada. p. 60626–68. doi:10.5555/3737916.3739855.
12. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, et al. Instruction tuning for large language models: a survey. ACM Comput Surv. 2026;58(7):1–36. doi:10.1145/3777411.
13. Ouyang L, Wu J, Xu J, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 27730–44. doi:10.5555/3600270.3602281.
14. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. In: Proceedings of the Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 53728–41. doi:10.52202/075280-2338.
15. Annepaka Y, Pakray P. Large language models: a survey of their development, capabilities, and applications. Knowl Inf Syst. 2025;67(3):2967–3022. doi:10.1007/s10115-024-02310-4.

16. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access. 2024;12(8):26839–74. doi:10.1109/ACCESS.2024.3365742.
17. Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access. 2025;13(2):18912–36. doi:10.1109/ACCESS.2025.3532853.
18. AlDahoul N, Tan MJ, Kasireddy HR, Zaki Y. Guardians of digital safety: benchmarking large language models in the fight against online toxicity. J Big Data. 2025;13(1):6. doi:10.1186/s40537-025-01336-x.
19. Valdez HPD, Abri F, Webb J, Austin TH. Exploring the use and misuse of large language models. Information. 2025;16(9):758. doi:10.3390/info16090758.
20. Wang Y, Wang M, Manzoor MA, Liu F, Georgiev GN, Das RJ, et al. Factuality of large language models: a survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 19519–29. doi:10.18653/v1/2024.emnlp-main.1088.
21. Chen M, Xiao C, Sun H, Li L, Derczynski L, Anandkumar A, et al. Combating security and privacy issues in the era of large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. doi:10.18653/v1/2024.naacl-tutorials.2.
22. Gupta O, Marrone S, Gargiulo F, Jaiswal R, Marassi L. Understanding social biases in large language models. AI. 2025;6(5):106. doi:10.3390/ai6050106.
23. Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, et al. Self-instruct: aligning language models with self-generated instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, ON, Canada. p. 13484–508. doi:10.18653/v1/2023.acl-long.754.
24. Liu W, Wang X, Wu M, Li T, Lv C, Ling Z, et al. Aligning large language models with human preferences through representation engineering. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 10619–38. doi:10.18653/v1/2024.acl-long.572.
25. Abbo GA, Marchesi S, Wykowska A, Belpaeme T. Social value alignment in large language models. In: Value engineering in artificial intelligence. Cham, Switzerland: Springer Nature; 2024. p. 83–97. doi:10.1007/978-3-031-58202-8_6.
26. Torgbi Agbemabiese W. Toward constitutional autonomy in AI systems: a theoretical framework for aligned agentic intelligence. IEEE Access. 2026;14:11385–402. doi:10.1109/ACCESS.2026.3654907.
27. Ki D, Rudinger R, Zhou T, Carpuat M. Multiple LLM agents debate for equitable cultural alignment. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 24841–77. doi:10.18653/v1/2025.acl-long.1210.
28. Yang Z, Meng Z, Zheng X, Wattenhofer R. Assessing adversarial robustness of large language models: an empirical study. arXiv:2405.02764. 2024.
29. Moradi M, Samwald M. Evaluating the robustness of neural language models to input perturbations. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Online. p. 1558–70. doi:10.18653/v1/2021.emnlp-main.117.
30. Li Z, Peng B, He P, Yan X. Evaluating the instruction-following robustness of large language models to prompt injection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 557–68. doi:10.18653/v1/2024.emnlp-main.33.
31. Oren Y, Sagawa S, Hashimoto T, Liang P. Distributionally robust language modeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 4226–36. doi:10.18653/v1/d19-1432.
32. Li Y. DRO-InstructZero: distributionally robust prompt optimization for large language models. arXiv:2510.15260. 2025.
33. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns. 2025;6(1):101118. doi:10.1016/j.patter.2024.101118.
34. Du X, Mo F, Wen M, Gu T, Zheng H, Jin H, et al. Multi-turn jailbreaking large language models via attention shifting. Proc AAAI Conf Artif Intell. 2025;39(22):23814–22. doi:10.1609/aaai.v39i22.34553.

35. Hu H, Robey A, Liu C. Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. arXiv:2503.00187. 2025.
36. Rabinovich E, Anaby Tavor A. On the robustness of agentic function calling. In: Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025); 2025 May 1–12; Albuquerque, NM, USA. p. 298–304. doi:10.18653/v1/2025.trustnlp-main.20.
37. Zhou L, Schellaert W, Martínez-Plumed F, Moros-Daval Y, Ferri C, Hernández-Orallo J. Larger and more instructable language models become less reliable. Nature. 2024;634(8032):61–8. doi:10.1038/s41586-024-07930-y.
38. Kumar S, Balachandran V, Njoo L, Anastasopoulos A, Tsvetkov Y. Language generation models can cause harm: so what can we do about it? An actionable survey. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023 May 2–6; Dubrovnik, Croatia. p. 3299–321. doi:10.18653/v1/2023.eacl-main.241.
39. Ganguli R, Moraffah R. Why do large language models generate harmful content? arXiv:2604.11663. 2026.
40. Ousidhoum N, Zhao X, Fang T, Song Y, Yeung DY. Probing toxic content in large pre-trained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. p. 4262–74. doi:10.18653/v1/2021.acl-long.329.
41. Alansari A, Luqman H. Large language models hallucination: a comprehensive survey. Comput Sci Rev. 2026;61(12):100970. doi:10.1016/j.cosrev.2026.100970.
42. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1–55. doi:10.1145/3703155.
43. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and fairness in large language models: a survey. Comput Linguist. 2024;50(3):1097–179. doi:10.1162/coli_a_00524.
44. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. J Data Inf Qual. 2023;15(2):1–21. doi:10.1145/3597307.
45. Chu Z, Wang Z, Zhang W. Fairness in large language models: a taxonomic survey. SIGKDD Explor Newsl. 2024;26(1):34–48. doi:10.1145/3682112.3682117.
46. Feng X, An B, Gu T, Chang L, Hao F, Yu P, et al. C2PO: diagnosing and disentangling bias shortcuts in LLMs. arXiv:2512.23430. 2025.
47. Yan B, Li K, Xu M, Dong Y, Zhang Y, Ren Z, et al. On protecting the data privacy of large language models (LLMs): a survey. In: 2024 International Conference on Meta Computing (ICMC); 2024 Jun 20–23; Qingdao, China. doi:10.1109/ICMC60390.2024.00008.
48. López JAH, Chen B, Saad M, Sharma T, Varró D. On inter-dataset code duplication and data leakage in large language models. IIEEE Trans Software Eng. 2025;51(1):192–205. doi:10.1109/tse.2024.3504286.
49. Hui B, Yuan H, Gong N, Burlina P, Cao Y. PLeak: prompt leaking attacks against large language model applications. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security; 2024 Oct 14–18; Salt Lake City, UT, USA. p. 3600–14. doi:10.1145/3658644.3670370.
50. Salemi A, Zamani H. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR); 2025 Jul 18; Padua, Italy. p. 286–96. doi:10.1145/3731120.3744595.
51. Li MQ, Fung BCM. Security concerns for large language models: a survey. J Inf Secur Appl. 2025;95:104284. doi:10.1016/j.jisa.2025.104284.
52. Yan J, Yadav V, Li S, Chen L, Tang Z, Wang H, et al. Backdooring instruction-tuned large language models with virtual prompt injection. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 6065–86. doi:10.18653/v1/2024.naacl-long.337.
53. Xu Z, Liu Y, Deng G, Li Y, Picek S. A comprehensive study of jailbreak attack versus defense for large language models. In: Proceedings of the Findings of the Association for Computational Linguistics ACL 2024; 2024 Aug 11–16; Bangkok, Thailand. p. 7432–49. doi:10.18653/v1/2024.findings-acl.443.

54. Zhao K, Li L, Ding K, Gong NZ, Zhao Y, Dong Y. A survey on model extraction attacks and defenses for large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2; 2025 Aug 3–7; Toronto, ON, Canada. p. 6227–36. doi:10.1145/3711896.3736573.
55. Zhao S, Jia M, Luu AT, Pan F, Wen J. Universal vulnerabilities in large language models: backdoor attacks for in-context learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 11507–22. doi:10.18653/v1/2024.emnlp-main.642.
56. Parakala A, Padgett P. When AI acts: opportunities and risks of agentic systems. Int J Artif Intell Data Sci Mach Learn. 2025;6(4):29–40. doi:10.63282/3050-9262.ijaidsml-v6i4p105.
57. Hagendorff T. Deception abilities emerged in large language models. Proc Natl Acad Sci U S A. 2024;121(24):e2317967121. doi:10.1073/pnas.2317967121.
58. Peng H, Qi Y, Wang X, Yao Z, Xu B, Hou L, et al. Agentic reward modeling: integrating human preferences with verifiable correctness signals for reliable reward systems. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 15934–49. doi:10.18653/v1/2025.acl-long.775.
59. Zhang G, Wang J, Chen J, Zhou W, Wang K, Yan S. AgenTracer: who is inducing failure in the LLM agentic systems? arXiv:2509.03312. 2025.
60. Ye J, Li S, Li G, Huang C, Gao S, Wu Y, et al. ToolSword: unveiling safety issues of large language models in tool learning across three stages. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 2181–211. doi:10.18653/v1/2024.acl-long.119.
61. Su H, Luo J, Liu C, Yang X, Zhang Y, Dong Y, et al. A survey on autonomy-induced security risks in large model-based agents. IEEE Trans Pattern Anal Mach Intell. 2026;2026:1–20. doi:10.1109/tpami.2026.3688650.
62. Muennighoff N, Rush A, Barak B, Le Scao T, Tazi N, Piktus A, et al. Scaling data-constrained language models. In: Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 50358–76. doi:10.52202/075280-2191.
63. Chen D, Huang Y, Ma Z, Chen H, Pan X, Ge C, et al. Data-juicer: a one-stop data processing system for large language models. In: Companion of the 2024 International Conference on Management of Data; 2024 Jun 9–15; Santiago, Chile. p. 120–34. doi:10.1145/3626246.3653385.
64. He J, Fan Z, Kuang S, Li X, Song K, Zhou Y, et al. FiNE: filtering and improving noisy data elaborately with large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies; 2025 Apr 29–May 4; Albuquerque, NM, USA. p. 8686–707. doi:10.18653/v1/2025.naacl-long.437.
65. Wang F, Mehrabi N, Goyal P, Gupta R, Chang KW, Galstyan A. Data advisor: dynamic data curation for safety alignment of large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 8089–100. doi:10.18653/v1/2024.emnlp-main.461.
66. Chen J, Mueller J. Automated data curation for robust language model fine-tuning. arXiv:2403.12776. 2024.
67. Chang TY, Jia R. Data curation alone can stabilize in-context learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, Canada. p. 8123–44. doi:10.18653/v1/2023.acl-long.452.
68. Abbasalizadeh M, Narain S. Privacy-aware detection for large language models using a hybrid BiLSTM-HMM approach. IEEE Access. 2025;13:121880–901. doi:10.1109/ACCESS.2025.11077118.
69. Chen K, Zhou X, Lin Y, Feng S, Shen L, Wu P. A survey on privacy risks and protection in large language models. J King Saud Univ Comput Inf Sci. 2025;37(7):163. doi:10.1007/s44443-025-00177-1.
70. Javed H, Ali F, Shah B, Dilshad N, Kwak D. MediGuard: protecting sensitive healthcare data with privacy-preserving language models. IEEE J Biomed Heal Inform. 2025;2025:1–14. doi:10.1109/JBHI.2025.3602983.
71. Fang F, Bai Y, Ni S, Yang M, Chen X, Xu R. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 10028–39. doi:10.18653/v1/2024.acl-long.540.
72. Moradi M, Samwald M. Improving the robustness and accuracy of biomedical language models through adversarial training. J Biomed Inform. 2022;132(1):104114. doi:10.1016/j.jbi.2022.104114.

73. Zhou X, Lu Y, Ma R, Wei Y, Gui T, Zhang Q, et al. Making harmful behaviors unlearnable for large language models. In: Findings of the association for computational linguistics: ACL 2024; Bangkok, Thailand: Association for Computational Linguistics. p. 10258–73. doi:10.18653/v1/2024.findings-acl.611.
74. Tan X, Shi S, Qiu X, Qu C, Qi Z, Xu Y, et al. Self-criticism: aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2023 Dec 6–10; Singapore. p. 650–62. doi:10.18653/v1/2023.emnlp-industry.62.
75. Sharma A, Keh S, Mitchell E, Finn C, Arora K, Kollar T. A critical evaluation of AI feedback for aligning large language models. In: Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 29166–90. doi:10.52202/079017-0919.
76. Baumann J, Kramer O. Evolutionary multi-objective optimization of large language model prompts for balancing sentiments. In: Applications of evolutionary computation. Cham, Switzerland: Springer Nature; 2024. p. 212–24. doi:10.1007/978-3-031-56855-8_13.
77. He Q, Maghsudi S. Pareto multi-objective alignment for language models. In: Machine learning and knowledge discovery in databases. Research track. Cham, Switzerland: Springer Nature; 2026. p. 257–72. doi:10.1007/978-3-032-06078-5_15.
78. Huang S, Siddarth D, Lovitt L, Liao TI, Durmus E, Tamkin A, et al. Collective constitutional AI: aligning a language model with public input. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; 2024 Jun 3–6; Rio de Janeiro, Brazil. p. 1395–417. doi:10.1145/3630106.3658979.
79. Wachi A, Tran T, Sato R, Tanabe T, Akimoto Y. Stepwise alignment for constrained language model policy optimization. In: Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 104471–520. doi:10.52202/079017-3319.
80. Gui J, Liu Y, Cheng J, Gu X, Liu X, Wang H, et al. LogicGame: benchmarking rule-based reasoning abilities of large language models. In: Findings of the association for computational linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics; 2025. p. 1474–91. doi:10.18653/v1/2025.findings-acl.77.
81. Zheng Y, Li X, Huang Y, Liang Q, Guo T, Hou M, et al. Automatic lesson plan generation via large language models with self-critique prompting. In: Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. Cham, Switzerland: Springer Nature; 2024. p. 163–78. doi:10.1007/978-3-031-64315-6_13.
82. Tripathi P, Arora P, Paroha KK, Kumar D, Manisha, Manisha P, et al. Guardrail-based approaches for enhancing safety, security, and privacy in large language models. In: 2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI); 2026 Feb 12–14; Greater Noida, India; 2026. p. 1619–24. doi:10.1109/IC3ECSBHI67834.2026.11469093.
83. Germany HPI, Schall M, de Melo G, Germany HPI. The hidden cost of structure: how constrained decoding affects language model performance. In: Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing; 2025 Sep 8–10. Varna, Bulgaria. p. 1074–84. doi:10.26615/978-954-452-098-4-124.
84. Zou J, Zhang S, Qiu M. Adversarial attacks on large language models. In: Knowledge science, engineering and management. Singapore: Springer Nature; 2024. p. 85–96. doi:10.1007/978-981-97-5501-1_7.
85. Wang D, Gong C, Liu Q. Improving neural language modeling via adversarial training. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 6555–65.
86. Du H, Liu S, Zheng L, Cao Y, Nakamura A, Chen L. Privacy in fine-tuning large language models: attacks, defenses, and future directions. In: Advances in knowledge discovery and data mining. Singapore: Springer Nature; 2025. p. 326–44. doi:10.1007/978-981-96-8183-9_25.
87. Ahmed SS, Angel Arul Jothi J. Jailbreak attacks on large language models and possible defenses: present status and future possibilities. In: 2024 IEEE International Symposium on Technology and Society (ISTAS); 2024 Sep 18–20; Puebla, Mexico. doi:10.1109/ISTAS61960.2024.10732418.
88. Tonmoy SMTI, Mehedi Zaman SM, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv:2401.01313. 2024.

89. Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics. 2025;13(5):856. doi:10.3390/math13050856.
90. Yu Y, Li L, Li Y. Augmenting large language models and retrieval-augmented generation with an evidence-based medicine-enabled agent system. medRxiv. 2025. doi:10.1101/2025.10.17.25338266.
91. Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, et al. Chain-of-verification reduces hallucination in large language models. In: Findings of the association for computational linguistics: ACL 2024; Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 3563–78. doi:10.18653/v1/2024.findings-acl.212.
92. Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell. 2024;6(8):852–63. doi:10.1038/s42256-024-00881-z.
93. Geng J, Cai F, Wang Y, Koeppl H, Nakov P, Gurevych I. A survey of confidence estimation and calibration in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 6577–95. doi:10.18653/v1/2024.naacl-long.366.
94. Schreieder T, Schopf T, Färber M. Attribution, citation, and quotation: a survey of evidence-based text generation with large language models. arXiv:2508.15396. 2025.
95. Kim S, Yun S, Lee H, Gubri M, Yoon S, Oh SJ. ProPILE: probing privacy leakage in large language models. In: Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 20750–62. doi:10.52202/075280-0911.
96. Yadav H, Singh V, Sharma K. Adversial prompt injection in large language models: taxonomy, exploits, and mitigation frameworks. In: 2025 Seventh International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN); 2025 Dec 20–21; Kalyani, India. p. 244–51. doi:10.1109/ICRCICN68210.2025.11364988.
97. Chhabra A, Datta S, Nahin SK, Mohapatra P. Agentic AI security: threats, defenses, evaluation, and open challenges. IEEE Access. 2026;14:49455–82. doi:10.1109/ACCESS.2026.3675554.
98. Liu Q, Mo W, Tong T, Xu J, Wang F, Xiao C, et al. Mitigating backdoor threats to large language models: advancement and challenges. In: 2024 60th Annual Allerton Conference on Communication, Control, and Computing; 2024 Sep 24–27; Urbana, IL, USA. doi:10.1109/Allerton63246.2024.10735305.
99. An B, Zhang S, Dredze M. RAG LLMs are not safer: a safety analysis of retrieval-augmented generation for large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies; 2025 Apr 25–May 4; Albuquerque, NM, USA. p. 5444–74. doi:10.18653/v1/2025.naacl-long.281.
100. Ruan Y, Dong H, Wang A, Pitis S, Zhou Y, Ba J, et al. Identifying the risks of LM agents with an LM-emulated sandbox. arXiv:2309.15817. 2023.
101. Plaat A, Wong A, Verberne S, Broekens J, Van Stein N, Bäck T. Multi-step reasoning with large language models, a survey. ACM Comput Surv. 2026;58(6):1–35. doi:10.1145/3774896.
102. Liu S, Yao Y, Jia J, Casper S, Baracaldo N, Hase P, et al. Rethinking machine unlearning for large language models. Nat Mach Intell. 2025;7(2):181–94. doi:10.1038/s42256-025-00985-0.
103. Cui J, Chiang WL, Stoica I, Hsieh CJ. OR-bench: an over-refusal benchmark for large language models. arXiv:2405.20947. 2024.
104. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. J Med Internet Res. 2025;27:e84120. doi:10.2196/84120.
105. Wang C, Liu X, Yue Y, Guo Q, Hu X, Tang X, et al. Survey on factuality in large language models. ACM Comput Surv. 2026;58(1):1–37. doi:10.1145/3742420.
106. Dong X, Luu AT, Ji R, Liu H. Towards robustness against natural language word substitutions. In: Proceedings of the International Conference on Learning Representations (ICLR); 2020 Apr 26–May 1; Online.
107. Zhang M, Yang Y, Xie R, Dhingra B, Zhou S, Pei J. Generalizability of large language model-based agents: a comprehensive survey. ACM Comput Surv. 2026;58(10):1–44. doi:10.1145/3794858.

108. Yang Y, Jin Q, Zhu Q, Wang Z, Erramuspe Álvarez F, Wan N, et al. Beyond multiple-choice accuracy: real-world challenges of implementing large language models in healthcare. Annu Rev Biomed Data Sci. 2025;8(1):305–16. doi:10.1146/annurev-biodatasci-103123-094851.
109. Ye J, Wu Y, Gao S, Huang C, Li S, Li G, et al. RoTBench: a multi-level benchmark for evaluating the robustness of large language models in tool learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 313–33. doi:10.18653/v1/2024.emnlp-main.19.
110. Lin Z, Wang Z, Tong Y, Wang Y, Guo Y, Wang Y, et al. ToxicChat: unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In: Findings of the association for computational linguistics: EMNLP 2023; Singapore: Association for Computational Linguistics. p. 4694–702. doi:10.18653/v1/2023.findings-emnlp.311.
111. Xie T, Qi X, Zeng Y, Huang Y, Sehwag U, Huang K, et al. Sorry-bench: systematically evaluating large language model safety refusal. In: Proceedings of the International Conference on Learning Representations (ICLR); 2025 Apr 24–28; Singapore.
112. Röttger P, Kirk H, Vidgen B, Attanasio G, Bianchi F, Hovy D. XSTest: a test suite for identifying exaggerated safety behaviours in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 5377–400. doi:10.18653/v1/2024.naacl-long.301.
113. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. p. 3214–52. doi:10.18653/v1/2022.acl-long.229.
114. Min S, Krishna K, Lyu X, Lewis M, Yih WT, Koh P, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 12076–100. doi:10.18653/v1/2023.emnlp-main.741.
115. Parrish A, Chen A, Nangia N, Padmakumar V, Phang J, Thompson J, et al. BBQ: a hand-built bias benchmark for question answering. In: Findings of the association for computational linguistics: ACL 2022; Dublin, Ireland: ACL; 2022. p. 2086–105. doi:10.18653/v1/2022.findings-acl.165.
116. Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang KW, et al. BOLD: dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3–10; Virtual Event. p. 862–72. doi:10.1145/3442188.3445924.
117. Chao P, Debenedetti E, Robey A, Andriushchenko M, Croce F, Sehwag V, et al. JailbreakBench: an open robustness benchmark for jailbreaking large language models. In: Advances in neural information processing systems 37; Vancouver, BC, Canada. p. 55005–29. doi:10.52202/079017-1745.
118. Mazeika M, Phan L, Yin X, Zou A, Wang Z, Mu N, et al. HarmBench: a standardized evaluation framework for automated red teaming and robust refusal. arXiv:2402.04249. 2024.
119. Battiato S, Casu M, Guarnera F, Guarnera L, Puglisi G, Pontorno O, et al. Adversarial attacks on deepfake detectors: a challenge in the era of AI-generated media (AADD-2025). In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin Ireland. p. 13714–9. doi:10.1145/3746027.3761983.
120. Duan M, Suri A, Mireshghallah N, Min S, Shi W, Zettlemoyer L, et al. Do membership inference attacks work on large language models? arXiv:2402.07841. 2024.
121. Li Q, Hong J, Xie C, Tan J, Xin R, Hou J, et al. LLM-PBE: assessing data privacy in large language models. arXiv:2408.12787. 2024.
122. El Yagoubi F, Badu-Marfo G, Al Mallah R. AgentLeak: a benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv:2602.11510. 2026.
123. Andriushchenko M, Souly A, Dziemian M, Duenas D, Lin M, Wang J, et al. AgentHarm: a benchmark for measuring harmfulness of LLM agents. arXiv:2410.09024. 2024.
124. Zhang Z, Cui S, Lu Y, Zhou J, Yang J, Wang H, et al. Agent-SafetyBench: evaluating the safety of LLM agents. arXiv:2412.14470. 2024.

125. Elangovan A, Liu L, Xu L, Bodapati SB, Roth D. ConSiDERS-the-human evaluation framework: rethinking human evaluation for generative large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 1137–60. doi:10.18653/v1/2024.acl-long.63.
126. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv:2209.07858. 2022.
127. Li J. A comparative study on annotation quality of crowdsourcing and LLm via label aggregation. In: ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 6525–9. doi:10.1109/ICASSP48485.2024.10447803.
128. Amirizaniani M, Lavergne A, Snell Okada E, Chadha A, Roosta T, Shah C. Developing a framework for auditing large language models using human-in-the-loop. In: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region; 2025 Dec 7–10; Xi'an, China. p. 64–74. doi:10.1145/3767695.3769514.
129. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digit Med. 2024;7(1):258. doi:10.1038/s41746-024-01258-7.
130. Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, et al. Red teaming language models with language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, United Arab Emirates. p. 3419–48. doi:10.18653/v1/2022.emnlp-main.225.
131. Chiang CH, Lee HY. A closer look into using large language models for automatic evaluation. In: Findings of the association for computational linguistics: EMNLP 2023; Singapore: Association for Computational Linguistics; 2023. p. 8928–42. doi:10.18653/v1/2023.findings-emnlp.599.
132. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol. 2024;15(3):1–45. doi:10.1145/3641289.
133. Cantini R, Orsino A, Ruggiero M, Talia D. Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. Mach Learn. 2025;114(11):249. doi:10.1007/s10994-025-06862-6.
134. Jia Q, Yue X, Zheng T, Huang J, Lin BY. SimulBench: evaluating language models with creative simulation tasks. In: Findings of the Association for Computational Linguistics: NAACL 2025; 2025 Apr 29–May 4; Albuquerque, NM, USA. p. 8133–46. doi:10.18653/v1/2025.findings-naacl.453.
135. Laskar MTR, Alqahtani S, Bari MS, Rahman M, Khan MAM, Khan H, et al. A systematic survey and critical review on evaluating large language models: challenges, limitations, and recommendations. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 13785–816. doi:10.18653/v1/2024.emnlp-main.764.
136. Hughes J, Price S, Lynch A, Schaeffer R, Barez F, Somani A, et al. Best-of-n jailbreaking. Adv Neural Inf Process Syst. 2026;38:73137–221.
137. Chen H, Goldfarb-Tarrant S. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 19750–66. doi:10.18653/v1/2025.acl-long.970.
138. Zhang X. How sensitive are safety benchmarks to judge configuration choices? arXiv:2604.24074. 2026.
139. Bao R, Yu D, Fan K, Liao M. Fixing distribution shifts of LLM self-critique via on-policy self-play training. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 17680–700. doi:10.18653/v1/2025.acl-long.865.
140. Khraishi R, Zafar I, Myles K, Cowan GA. Evaluating performance drift from model switching in multi-turn LLM systems. arXiv:2603.03111. 2026.
141. Liu Z, Kim D, Wan Y, Yuan X, Tan Z, Mo F, et al. MTMCS-bench: evaluating contextual safety of multimodal large language models in multi-turn dialogues. arXiv:2601.06757. 2026.
142. Xi Z, Zheng R, Gui T. Safety and ethical concerns of large language models. In: Proceedings of the 22nd Chinese National Conference on Computational Linguistics; 2023 Aug 3–5; Harbin, China. p. 9–16.

143. Zhang Z, Chen J, Yang D. DARG: dynamic evaluation of large language models via adaptive reasoning graph. In: Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 135904–42. doi:10.52202/079017-4317.
144. Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, et al. Let's verify step by step. In: International Conference on Learning Representations 2024; 2024 May 7–11; Vienna, Austria. p. 39578–601.
145. Dai N, Wu Z, Zheng R, Wei Z, Shi W, Jin X, et al. Process supervision-guided policy optimization for code generation. arXiv:2410.17621. 2024.
146. Zhao B, Srikanth D, Wu Y, Jiang Z. SpecBench: measuring reward hacking in long-horizon coding agents. arXiv:2605.21384. 2026.
147. Gantla SR. Exploring mechanistic interpretability in large language models: challenges, approaches, and insights. In: 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI); 2025 Mar 28–29; Chennai, India. doi:10.1109/ICDSAAI65575.2025.11011640.
148. Zhang L, Song D, Wu Z, Tian Y, Zhou C, Xu J, et al. Detecting hallucination in large language models through deep internal representation analysis. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence; 2025 Aug 16–22; Montreal, Canada. p. 8357–65. doi:10.24963/ijcai.2025/929.
149. Scalena D, Sarti G, Nissim M. Multi-property steering of large language models with dynamic activation composition. In: Proceedings of the 7th Blackbo×NLP Workshop: Analyzing and Interpreting Neural Networks for NLP; 2024 Nov 15; Miami, FL, USA. p. 577–603. doi:10.18653/v1/2024.blackboxnlp-1.34.
150. Sun H, Peng H, Dai Q, Bai X, Cao Y. LayerNavigator: finding promising intervention layers for efficient activation steering in large language models. Adv Neural Inf Process Syst. 2026;38:101058–80.
151. Nosrati K, Tepljakov A, Belikov J, Petlenkov E. When control meets large language models: from words to dynamics. Eng Appl Artif Intell. 2026;178(5):115119. doi:10.1016/j.engappai.2026.115119.
152. Raheem T, Hossain G. Agentic AI systems: opportunities, challenges, and trustworthiness. In: 2025 IEEE International Conference on Electro Information Technology (eIT); 2025 May 29–31; Valparaiso, IN, USA. p. 618–24. doi:10.1109/eIT64391.2025.11103638.
153. Ferrag MA, Lakas A, Tihanyi N, Debbah M. LLM and AI agents for autonomous systems: a survey of applications, datasets, and security challenges. IEEE Open J Intell Transp Syst. 2026;7:615–57. doi:10.1109/OJITS.2026.3665677.
154. Menshawy A, Nawaz Z, Fahmy M. Navigating challenges and technical debt in large language models deployment. In: Proceedings of the 4th Workshop on Machine Learning and Systems; 2024 Apr 22; Athens Greece. p. 192–9. doi:10.1145/3642970.3655840.
155. Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI Ethics. 2024;4(4):1085–115. doi:10.1007/s43681-023-00289-2.
156. Pahune S, Akhtar Z, Mandapati V, Siddique K. The importance of AI data governance in large language models. Big Data Cogn Comput. 2025;9(6):147. doi:10.3390/bdcc9060147.