A Survey on Large Language Model Benchmarks

Shiwen Ni[1], Guhong Chen[1, 2], Shuaimin Li[1], Xuanang Chen[9], Siyi Li[1, 4], Bingli Wang[6]
Qiyao Wang[1, 3], Xingjian Wang[5], Yifan Zhang[7], Liyang Fan[8]
Chengming Li[10], Ruifeng Xu[11], Le Sun[9], Min Yang[1, 12, *]
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Southern University of Science and Technology [3]University of Chinese Academy of Sciences
[4]University of Science and Technology of China [5]Shanghai University of Electric Power
[6]Shanghai AI Lab [7]South China University of Technology [8]Shenzhen University
[9]Institute of Software, Chinese Academy of Sciences [10]Shenzhen MSU-BIT University
[11]Harbin Institute of Technology, Shenzhen [12]Shenzhen University of Advanced Technology

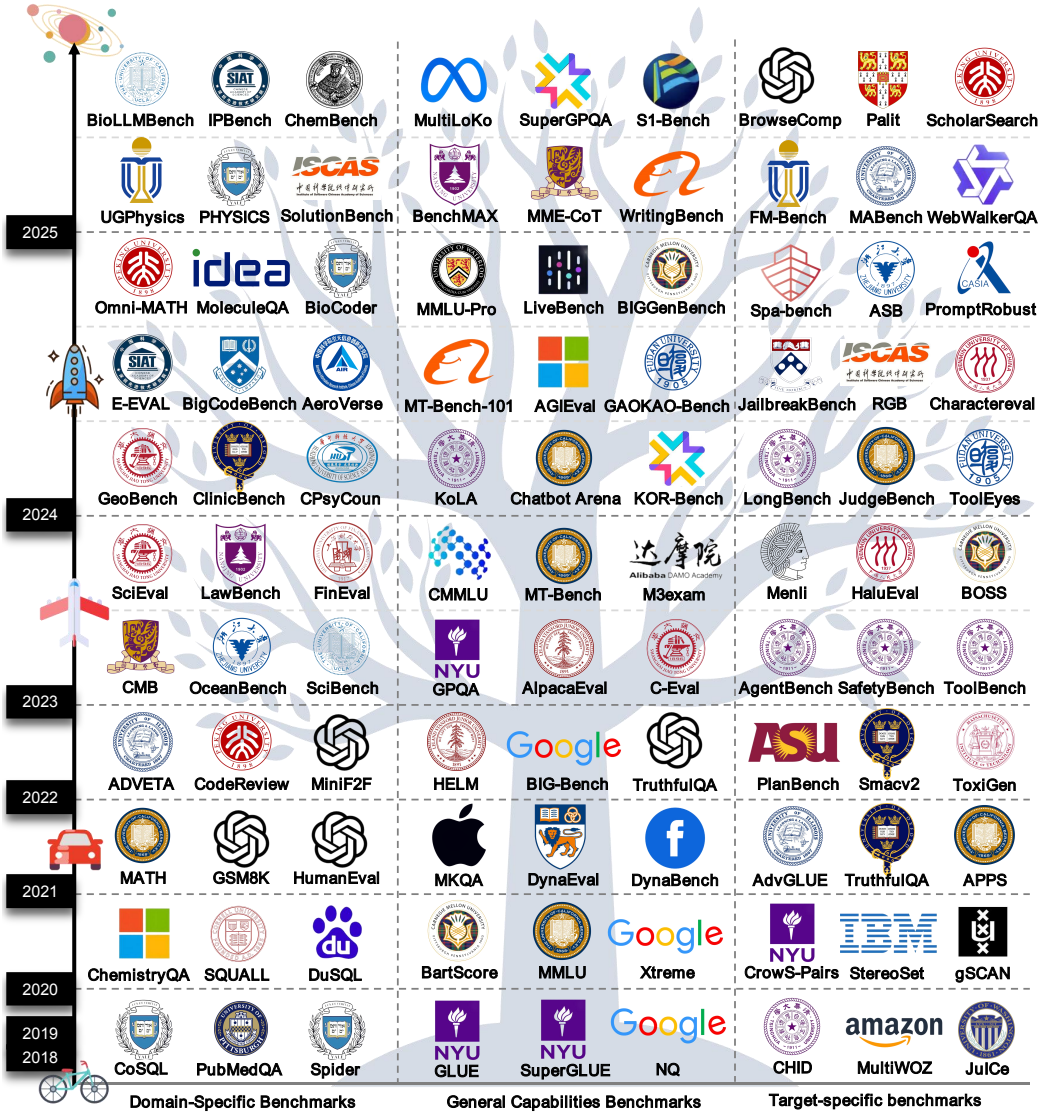

Figure 1: A timeline of representative LLM benchmarks.

Contents

Abstract

In recent years, with the rapid development of the depth and breadth of large language models' capabilities, various corresponding evaluation benchmarks have been emerging in increasing numbers. As a quantitative assessment tool for model performance, benchmarks are not only a core means to measure model capabilities but also a key element in guiding the direction of model development and promoting technological innovation. We systematically review the current status and development of large language model benchmarks for the first time, categorizing 283 representative benchmarks into three categories: general capabilities, domain-specific, and target-specific. General capability benchmarks cover aspects such as core linguistics, knowledge, and reasoning; domain-specific benchmarks focus on fields like natural sciences, humanities and social sciences, and engineering technology; target-specific benchmarks pay attention to risks, reliability, agents, etc. We point out that current benchmarks have problems such as inflated scores caused by data contamination, unfair evaluation due to cultural and linguistic biases, and lack of evaluation on process credibility and dynamic environments, and provide a referable design paradigm for future benchmark innovation.

1 Introduction

Since the Transformer [1] architecture was introduced in 2017, large language models (LLMs) have launched a revolutionary wave in the field of Artificial Intelligence (AI) with their powerful natural language processing capabilities. From basic natural language understanding and text generation tasks to complex logical reasoning and intelligent body interactions, LLMs continue to expand the boundaries of AI and reshape the human-computer interaction paradigm and information processing model. With the successive introduction of GPT series [2, 3, 4], LLaMA series [5, 6, 7], Qwen series [8, 9, 10], and other models, LLMs have widely penetrated into intelligent customer service, content creation, education, medical care, law and other fields, and have become the core driving force to promote the development of the digital economy and the intelligent transformation of society.

With the acceleration of the iteration of the LLM technology, it is urgent to establish a scientific and comprehensive evaluation system; Benchmarks, as a quantitative assessment of model performance, are not only a core tool to measure the ability of the model, but also a key element to guide the direction of model and promote technological innovation. Through benchmarks, researchers can objectively compare the strengths and weaknesses of different models, accurately locate technical bottlenecks, and provide data support for algorithm optimization and architectural design; At the same time, standardized evaluation results can help build user trust and ensure that the models comply with the social and ethical norms in terms of security and fairness. However, compared with the earlier language model evaluation benchmarks represented by GLUE [11] and SuperGLUE [12], the number of model parameters in the LLM era has increased exponentially, the capability dimension has expanded from single task to multitask and multidomain MMLU [13], GIG-bench [14], GPQA [15], SuperGPQA [16], and the evaluation paradigm has shifted from fixed task to multitask and multidomain. These changes put forward higher requirements on the scientific and adaptive nature of evaluation systems.

Currently, the field of LLM evaluation still faces many challenges that need to be solved. First, data leakage [17, 18] is becoming increasingly prominent, and some models are exposed to evaluation data during the training phase, leading to inflated evaluation results and fails to truly reflect the model generalization ability; second, static evaluation [13, 19] is difficult to simulate dynamic real-world scenarios and it is difficult to predict model performance when faced with new tasks and new domains. The singularity of evaluation indexes (e.g., over-reliance on accuracy rate and BLEU score) fails to comprehensively portray the complex capabilities of LLMs, and key requirements such as detection of bias and security loopholes and systematic evaluation of instruction compliance have not yet been effectively met. In addition, the high cost of arithmetic and manpower required for large-scale evaluation, and the difficulty of task design to cover the complexity of the real world are serious constraints to the healthy development of LLMs. Figure 1 shows a timeline of representative LLM benchmarks, illustrating this rapid evolution.

This paper is the first to conduct a systematic review and prospective analysis focusing on LLM benchmarks, with its contributions summarized as follows:

1. For the first time, 283 LLM benchmarks are analyzed and summarized under three categories: General Capabilities Benchmarks, Domain-Specific Benchmarks, and Target-specific Benchmarks.
2. This paper examines the design motivations and limitations of each benchmark from multiple perspectives, including data sources, data formats, data volume, evaluation methods, and evaluation metrics, providing a directly referable design paradigm for subsequent benchmark innovation.
3. We point out three major issues faced by current LLM benchmarks: inflated scores caused by data contamination; unfair evaluation due to cultural and linguistic biases; and the lack of evaluation on "process credibility" and "dynamic environments".

2 Background

2.1 Large Language Models

Research on language models dates back to Shannon in the 1950s, who pioneered modeling human language with information theory using n-gram models [20]. Its evolution has gone through several stages: statistical language models (e.g., n-gram models relying on co-occurrence statistics and independence assumptions [21, 22]), neural language models (utilizing distributed representations and architectures like recurrent neural networks [23, 24, 25], with works such as word2vec advancing representation learning [26, 27]), and subsequently pretrained language models (PLMs).

Pretrained language models learn context-aware representations from large unlabeled corpora for downstream fine-tuning. ELMo introduced bidirectional LSTM pretraining for dynamic embeddings [28]. The Transformer architecture, with its self-attention mechanism, became foundational for large-scale PLMs [1]. Based on this, models like BERT [29], GPT/GPT-2 [2, 3], BART [30], and T5 [31] emerged, following the "pretraining and fine-tuning" paradigm, with subsequent refinements (e.g., RoBERTa [32]).

Large language models (LLMs) developed from PLMs, driven by scaling laws that link increased parameters and data to improved performance [33]. With parameter counts growing to billions/trillions, LLMs exhibit emergent capabilities like few-shot learning and in-context learning [34]. The ecosystem includes proprietary models (e.g., OpenAI's ChatGPT, GPT-4 [4]; Anthropic's Claude; Google's Gemini [35]) and open-source ones (e.g., Meta's LLaMA series [5]; Alibaba's Qwen series [8, 9, 10]), excelling in diverse tasks from dialogue to multimodal reasoning.

2.2 LLM Benchmarks

The rapid advancement of large language models (LLMs) has fundamentally reshaped the landscape of natural language processing. As LLMs grow in scale—from millions to billions and now trillions of parameters—their emergent capabilities, including complex reasoning, instruction following, multi-turn dialogue, and tool usage, have presented unprecedented opportunities and challenges. In parallel with these developments, the design and evolution of benchmarks have become essential to accurately measure, compare, and guide the progress of LLMs. Benchmarks and LLMs are not independent trajectories; instead, their evolution is deeply intertwined, forming a mutually reinforcing cycle that continuously pushes the boundaries of the field.

In the early stages of language model development, benchmarks such as GLUE [11], BERTScore [42] and SuperGLUE [12] played a crucial role in driving research progress. These benchmarks primarily focused on natural language understanding (NLU) through relatively small-scale, single-task evaluations. However, as LLMs rapidly scaled up in size and began to exhibit emergent generalization abilities. In response, a new wave of LLM-specific benchmarks has emerged, such as MMLU [13], BIG-bench [14], HELM [55], AGIEval [56], GPQA [15]. These benchmarks aim to assess a wider range of capabilities, including reasoning, factual knowledge, robustness, multilingual understanding, and generalization to unseen tasks. Moreover, many of them are designed to evaluate LLMs in zero-shot or few-shot settings, aligning more closely with how these models are used in practice.

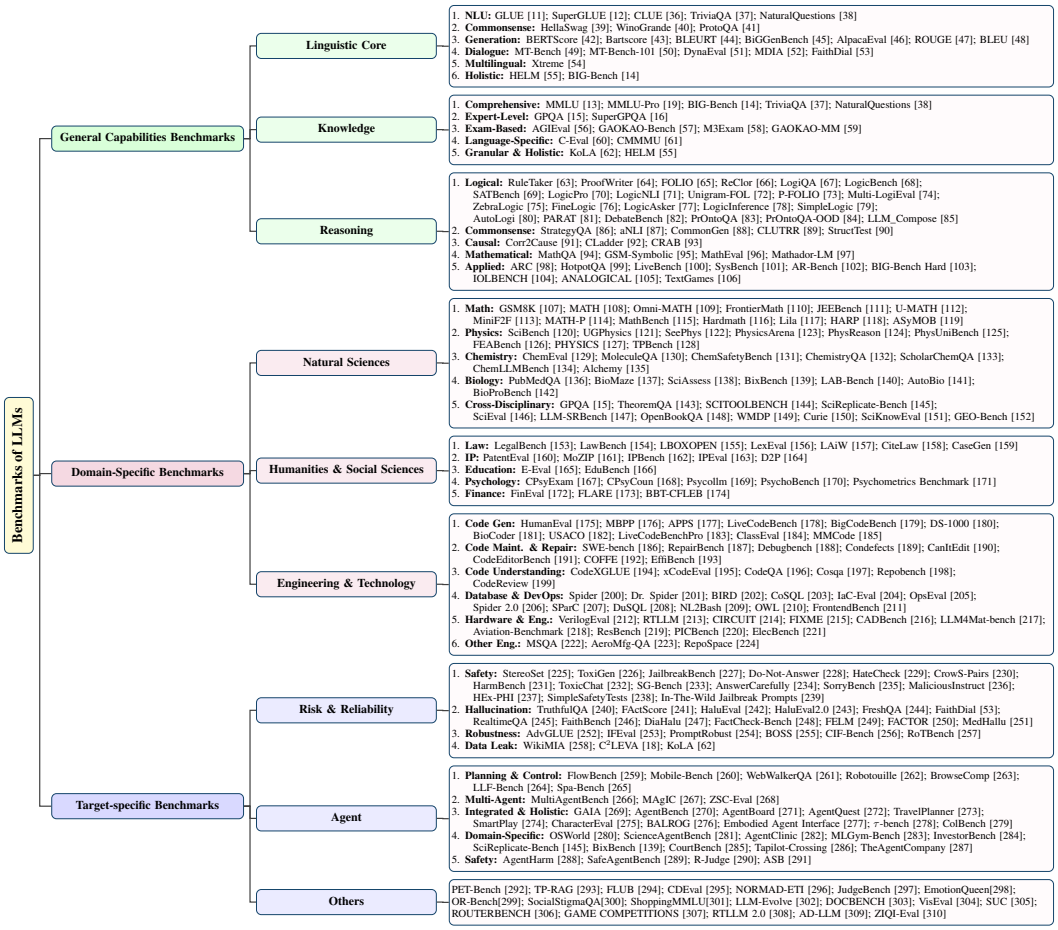

Figure 2: A taxonomy of representative benchmarks for Large Language Models, categorized by their primary evaluation focus.

In this survey, we present a comprehensive review of LLM benchmarks, delving into their design principles, coverage scope, inherent limitations, and emerging trends. Our objective is to crystallize the current landscape of LLM evaluation and offer actionable insights to inform the development of benchmarking strategies tailored for next-generation language models. Figure 2 provides a detailed taxonomy of these benchmarks, which serves as the organizational structure for the remainder of this paper.

3 General Capabilities Benchmarks

3.1 Linguistic Core

The evolution of linguistic capability benchmarks embodies a continuous arms race between model advancement and evaluation methodology. This progression is driven by the core challenge of measuring generalized linguistic competence, moving beyond surface-level pattern matching to assess deeper aspects of syntax, semantics, and pragmatics. This section chronicles how benchmarks evolved from fragmented task evaluations to dynamic, multilingual ecosystems, revealing fundamental shifts in how we define and quantify linguistic intelligence. A summary of representative benchmarks is provided in Table 1.

3.1.1 The Evolution of Linguistic Benchmarks

Phase 1: The Fragmentation Crisis and GLUE's Unification (2018)

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
GLUE [11]	Natural Language Understanding	Monolingual	Hybrid	Hybrid	AE	Accuracy	415,354	No	8607
SuperGLUE [12]	Natural Language Understanding	Monolingual	Hybrid	Hybrid	AE	Accuracy	20,483	No	2662
HellaSwag [39]	Commonsense Inference	Monolingual	Web	MCQA	AE	Accuracy	10,004	Yes	2475
WinoGrande [40]	Commonsense Inference	Monolingual	Hybrid	MCQA	AE	Accuracy	44,000	Yes	2025
BERTScore [42]	Text Generation	Multilingual	Open Datasets	Generation	LLM	Precision, Recall, F1	3 datasets	Yes	6977
CLUE [36]	Natural Language Understanding	Monolingual	Hybrid	Hybrid	ME	Accuracy, EM	900k+	No	397
Xtreme [54]	Multilingual Performance	Multilingual	Hybrid	Hybrid	AE	Accuracy, EM, F1	3 datasets	No	1017
Bartscore [43]	Text Generation	Multilingual	Open Datasets	Generation	LLM	BartScore	16 datasets	Yes	906
DynaEval [51]	Dialogue	Monolingual	Open Datasets	Classification	AE	Accuracy, F1	50k+	Yes	80
HELM [55]	Holistic Capability	Multilingual	Hybrid	Hybrid	AE	Accuracy, 6 Designed Metrics	17,431,479	No	1507
MT-Bench [49]	Multi-Turn Dialogues	Monolingual	Manual Design	Generation	LLM	8 Designed Metrics	80	Yes	4033
MDIA [52]	Multilingual Performance	Multilingual	Web	Generation	ME	4 Automated Metrics & SSA	380,914	No	17
BIG-Bench [14]	Holistic Capability	Multilingual	Manual Design	Hybrid	ME	EM, MC_Acc, Breakthroughness, etc	200k+	No	1628
BiGGenBench [45]	Text Generation	Multilingual	Hybrid	Generation	ME	Instr. Follow, Grounding, etc	765	No	7

Table 1: Summary of representative linguistic-core benchmarks. Evaluation methods are abbreviated as MCQA (Multiple Choice Question Answering), AE (Automated Evaluation), ME (Mixed Evaluation), SSA (Sensibleness and Specificity Average), MC_ACC (Multiple-Choice Accuracy). The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

Early natural language understanding (NLU) systems excelled at narrow tasks but failed to transfer knowledge across domains—a critical flaw for real-world applicability. GLUE [11], introduced in 2018, was a pivotal development, confronting this by integrating 9 diverse English NLU tasks (e.g., sentiment analysis, textual entailment) under a unified framework. Its diagnostic suite was crucial, exposing models' reliance on spurious statistical cues and lexical overlap rather than an understanding of syntactic structure or semantic roles. By including limited-data subtasks, GLUE incentivized the development of models that could build more robust and transferable linguistic representations, establishing multi-task evaluation as the new paradigm.

Phase 2: The Adversarial Turn and Deeper Linguistic Probes (2019)

BERT's rapid domination of GLUE [11] (surpassing human performance) revealed a deeper crisis: benchmarks were vulnerable to dataset-specific biases. SuperGLUE [12] responded with harder tasks requiring complex reasoning, but the field soon uncovered models' tendency to exploit annotation artifacts. This spurred a wave of adversarially constructed benchmarks.

Benchmarks like HellaSwag [39] were designed to be difficult for models yet trivial for humans by generating distractors that were semantically plausible but pragmatically absurd, directly probing commonsense and script knowledge. Concurrently, WinoGrande [40] used the AFLITE algorithm to de-bias 44,000 pronoun disambiguation problems. This forced models to properly handle anaphora and perform true coreference resolution—a fundamental syntactic-semantic challenge—rather than relying on word-association shortcuts. These innovations redefined benchmarks as active adversaries, dynamically evolving to test for deeper linguistic phenomena beyond surface patterns.

Phase 3: Beyond Linguistic Hegemony (2020)

The anglophone focus of GLUE [11] and SuperGLUE [12] limited the assessment of models' ability to generalize across languages with different structural properties. CLUE [36] was a significant first step for Chinese NLU. Xtreme [54] dramatically expanded this effort to 40 languages across 12 language families, systematically testing generalization across diverse typological properties (e.g., morphology, word order). The benchmark revealed significant performance degradation when models were transferred from high-resource, analytically-structured languages (like English) to morphologically rich or agglutinative ones. This "multilingual awakening" culminated in benchmarks like MDIA [52], which extended dialogue evaluation to 46 languages, emphasizing the need to evaluate models on a wide range of morphosyntactic and cultural contexts.

Phase 4: The Generation Paradigm Shift (2019–2021)

With the rise of generative models, metrics based on n-gram overlap like BLEU [48] and ROUGE [47] proved inadequate, as they fail to capture semantic equivalence. The field responded with a new class of semantic-aware metrics. BERTScore [42] leveraged contextual embeddings to measure semantic similarity, while BLEURT [44] trained a regression model on 6.5M synthetically perturbed sentence pairs to better align with human judgments of quality. Bartscore [43] reframed evaluation as a conditional language modeling task, directly assessing the probability of a reference given a generated output, thus aligning the metric with the model's pre-training objective. Concurrently, dialogue evaluation advanced from turn-level metrics to other more creative evaluation methods. DynaEval [51], for instance, used graph-based modeling to capture the coherence and logical flow of a conversation, assessing dependencies between utterances.

Phase 5: The Holistic Era and Fine-Grained Assessment (2022–Present)

Static benchmarks struggled to keep pace with the rapidly expanding linguistic capabilities of LLMs. In response, HELM [55] introduced a "living benchmark" concept, dynamically integrating emergent linguistic dimensions—from cross-lingual robustness to toxicity detection—through continuous scenario expansion. This fluidity found complement in BIG-Bench [14], where crowdsourced frontier tasks (204 challenges co-created by 442 researchers) deliberately targeted capabilities beyond contemporary models' reach, probing complex abilities like multi-step reasoning, metaphor interpretation, and theory of mind.

Concurrently, the LLM-as-Judge revolution redefined open-ended evaluation. MT-Bench [49] and MT-Bench-101 [50] leveraged GPT-4 to score open-ended dialogues along dimensions like perceptivity and adaptability, achieving high correlation with human judgments. BiGGenBench [45] pushed further, assigning instance-specific criteria (e.g., "Evaluate safety in this medical advice context") to overcome the limitations of coarse-grained metrics by enabling context-sensitive evaluation, a cornerstone of pragmatics.

This development forged an adaptive evaluation ecosystem that treats linguistic capability not as fixed competencies but as evolving phenotypes.

3.1.2 Cross-Cutting Design Innovations

Benchmark evolution reveals several tectonic shifts.

From Static to Living Frameworks: Early benchmarks were often presented as static collections of datasets with a single canonical metric, usually accuracy. Modern frameworks like HELM [55] and BIG-Bench [14] are dynamic, continuously incorporating new tasks to probe an expanding range of linguistic phenomena and mitigate benchmark saturation.

From Monolingual to Multilingual Stress Testing: Monolingual benchmarks (GLUE [11]/SuperGLUE [12]) implicitly assumed linguistic universality. Xtreme [54] and MDIA [52] shattered this illusion by establishing typological diversity as a core robustness probe.

From Task-Accuracy to Multi-Dimensional Intelligence: The LLM-as-Judge paradigm shifted evaluation from a single score like accuracy or F1 to a multi-faceted profile, assessing qualities like coherence, safety, and creativity which are previously unquantifiable at scale. Concurrently, adversarial filtering (HellaSWAG [39], WinoGrande [40]) and synthetic data infusion (BLEURT [44]) emerged as essential tools to combat dataset bias, ensuring models perform compositional reasoning rather than exploiting spurious statistical cues.

3.1.3 Summary and Future Directions

The relentless evolution of benchmarks has exposed fractures in their ability to authentically measure core linguistic capabilities, revealing three critical gaps demanding urgent resolution. **Persistent cross-linguistic inequities** remain entrenched despite expanded multilingual coverage—typological biases continue to distort performance measurements, as morphosyntactic divergences between analytic and agglutinative languages manifest in systemic evaluation gaps. Benchmarks like Xtreme [54] and MDIA [52] expose how shallow language tagging fails to capture structural phenomena like ergativity or vowel harmony, reducing linguistic diversity to mere metadata rather than a embedded variable in assessment frameworks.

Compounding these limitations, **the self-referential trap of LLM-as-Judge methodologies** threatens to calcify evaluation into stylistic monocultures. When frontier models like GPT-4 assess conversational depth or instruction fidelity in MT-Bench [49] or BiGGenBench [45], they risk circularly validating their own generative patterns, privileging familiarity over authentic capability. This epistemological crisis demands adversarial auditing frameworks and ensembles of specialized, domain-tuned judges enforcing pluralism while preserving human alignment.

Meanwhile, **the specter of resource asymmetry** corrupts benchmark integrity at its foundation. HELM [55]'s computational burden and MDIA [52]'s data scarcity for low-resource languages perpetuate exclusion. This violates linguistic justice—the principle that evaluation accessibility must scale with linguistic diversity. Emerging solutions like assessment with data decentralization, collective intelligence and dynamic task sampling offer pathways, but require rigorous fairness guarantees.

3.2 Knowledge

The capacity to store and accurately retrieve vast quantities of real-world information is a foundational pillar of modern Large Language Models (LLMs). These models function as veritable repositories of knowledge assimilated from extensive training corpora, making the quantification of this knowledge's extent and reliability a critical axis of evaluation. Consequently, benchmarks designed to probe this dimension have become a de-facto standard for gauging model progress. These evaluations typically simulate rigorous, "closed-book examinations," compelling models to rely solely on their internal, parameterized knowledge. This section presents a critical survey of the landscape of knowledge-oriented benchmarks, analyzing their methodological underpinnings, evolutionary trajectories, and the persistent challenges that shape future research. A summary of representative benchmarks is provided in Table 2.

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
MMLU [13]	Comprehensive knowledge	Monolingual	Web	MCQA	AE	Accuracy	15,908	No	4322
MMLU-Pro [19]	Robust knowledge	Monolingual	Hybrid	MCQA	AE	Accuracy	12,032	No	387
GPQA [15]	Google-Proof Q&A	Monolingual	Manual Design	MCQA	AE	Accuracy	448	No	724
SuperGPQA [16]	Graduate-level knowledge	Monolingual	Hybrid	MCQA	AE	Accuracy	26,529	No	17
C-Eval [60]	Chinese eval. suite	Monolingual	Web	MCQA	AE	Accuracy	13,948	No	204
AGIEval [56]	Human-centric exams	Bilingual	Std. Exams	Hybrid	AE	Accuracy, EM	8,062	No	480
GAOKAO-Bench [57]	Chinese college exam	Bilingual	Std. Exams	Hybrid	ME	Accuracy, Scoring Rate	2,811	No	97
KoLA [62]	Hierarchical knowledge	Monolingual	Open Datasets	Generation	AE	Custom Scores	19 datasets	No	138
BIG-Bench [14]	Extrapolating capabilities	Multilingual	Manual Design	Hybrid	ME	Multiple Metrics	200k+	No	1628
HELM [55]	Holistic evaluation	Multilingual	Hybrid	Hybrid	AE	7 Core Metrics	17M+	No	1507
M3Exam [58]	Multilingual/ modal exams	Multilingual	Std. Exams	MCQA	AE	Accuracy	12,317	No	141
CMMMU [61]	Chinese multi-modal understanding	Monolingual	Hybrid	Hybrid	AE	Accuracy	12,012	No	14

Table 2: Summary of representative knowledge-oriented benchmarks. Evaluation methods are abbreviated as AE (Automated Evaluation), ME (Mixed Evaluation). Data sources are abbreviated as Std. Exams (Standardized Exams). The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

3.2.1 Evolution of Knowledge Evaluation Paradigms

The trajectory of knowledge evaluation in LLMs mirrors the escalating capabilities of the models themselves, marked by a conceptual pivot from assessing information retrieval to probing internalized knowledge. Early paradigms often centered on open-domain question answering, such as in TriviaQA [37] and NaturalQuestions [38], where models were primarily evaluated on their ability to locate answers within provided documents.

A seminal shift occurred with the introduction of MMLU [13], which established a new and influential paradigm. By presenting a massive, multi-task benchmark of multiple-choice questions across 57 diverse disciplines without external context, MMLU forced the evaluation to focus squarely on the models' parameterized knowledge. This established a rigorous standard and catalyzed an arms race

in both model development and benchmark design. In response to emergent model saturation on MMLU, subsequent benchmarks have pushed the frontiers of difficulty and scope. For instance, MMLU-Pro [19] raised the adversarial bar by increasing the number of choices and the proportion of reasoning-intensive questions. Concurrently, benchmarks like GPQA [15] were designed by domain experts to be "Google-Proof," directly addressing the challenge of models retrieving answers from web search rather than relying on internalized knowledge, while SuperGPQA [16] further escalated the challenge into hundreds of highly specialized, graduate-level domains. This evolutionary arc reflects a continuous effort to create evaluations that remain challenging for even the most capable models.

3.2.2 Methodological Landscape and Divergent Philosophies

While sharing the common goal of knowledge assessment, these benchmarks are built upon a set of shared methodological foundations yet exhibit divergent evaluation philosophies. The predominant evaluation format is Multiple-Choice Question Answering (MCQA), a choice motivated by its scalability and amenability to objective, automated evaluation using accuracy as the primary metric. This approach, while logistically advantageous, has inherent limitations in assessing the nuances of knowledge generation and reasoning.

Beyond this common architecture, a number of distinct philosophical trajectories have emerged. **One prominent trajectory is the pursuit of human-centric alignment**, where evaluation is grounded in established human standards. Benchmarks like AGIEval [56] and GAOKAO-Bench [57] epitomize this approach by curating questions directly from high-stakes human examinations (e.g., college entrance and professional qualification tests). This methodology offers a more interpretable measure of a model's capabilities relative to human intellect. **Another direction focuses on achieving finer-grained analysis**. KoLA [62], for example, moves beyond a single accuracy score to propose a hierarchical framework that dissects knowledge into levels of recall, understanding, and application. **A third philosophy advocates for holistic and multi-faceted evaluation**. Rather than isolating knowledge, benchmarks like HELM [55] and BIG-Bench [14] integrate knowledge assessment (as accuracy) into a broader suite of metrics, including robustness, fairness, and calibration, providing a more comprehensive profile of model behavior. Finally, the expansion towards **multilingual and multimodal knowledge**, exemplified by benchmarks such as the multilingual M3Exam [58] and the Chinese-centric multimodal benchmarks GAOKAO-MM [59] and CMMMU [61], marks a critical effort to generalize evaluation beyond English-only, text-based paradigms.

3.2.3 Summary and Future Directions

In summary, while knowledge-oriented benchmarks have evolved to become more rigorous and diverse, they continue to face critical challenges that define the key directions for future research. **The first and most pervasive challenge is the specter of data contamination**. As models are trained on ever-expanding web-scale datasets, the probability of benchmark questions being present in the training data increases, potentially inflating scores and compromising the validity of the evaluation. This necessitates the development of dynamic or "Google-Proof" benchmarks, as seen in GPQA, as well as robust statistical methods for detecting contamination.

A second challenge lies in the methodological limitations of closed-form evaluation. The dominance of the MCQA format, while scalable, fails to capture a model's ability to generate coherent explanations, synthesize information, or admit uncertainty. This limitation may reward models adept at pattern matching rather than genuine comprehension. Consequently, a move towards hybrid evaluation frameworks that incorporate open-ended generation, assessed by either human experts or increasingly sophisticated LLM-as-a-judge systems [49, 46], is a crucial future direction.

Finally, the issues of static evaluation and cultural bias are intertwined. Most benchmarks represent a static snapshot of knowledge at a particular time and, often, from a predominantly Western, English-centric perspective. This not only makes them incapable of assessing a model's grasp of evolving, real-time information but also risks penalizing models with different cultural or linguistic knowledge bases. Addressing this requires a concerted effort to build more dynamic, culturally diverse, and multilingual benchmarks, following the path forged by comprehensive Chinese-language suites like CLUE [36] and C-Eval [60].

3.3 Reasoning

The ability to reason—spanning formal logic, commonsense inference, and applied problem-solving—is a cornerstone of higher intelligence. Evaluating this capability in Large Language Models (LLMs) is crucial for understanding their cognitive limits and practical potential. This section surveys a wide array of benchmarks designed to test these facets of reasoning, from structured logical puzzles to complex, real-world scenarios. A comprehensive overview of these benchmarks, categorized by reasoning type, is presented in Table 3.

3.3.1 Logical Reasoning

The domain of logical reasoning represents the most mature and densely populated area of LLM evaluation. This focus is understandable, as formal logic provides the bedrock of structured thought. The overall landscape reveals a clear developmental arc, beginning with foundational benchmarks testing discrete deductive steps (e.g., SimpleLogic [79]) and evolving towards assessments of highly complex, multi-step, and even programmatic reasoning (e.g., LogicPro [70]). This progression reflects the community's growing ambition, moving from asking "Can LLMs perform logical operations?" to "Can LLMs think like a reasoner?".

A primary commonality across these benchmarks is their reliance on controlled environments where logical correctness is unambiguous. As shown in Table 3, most datasets are either human-authored (e.g., FOLIO [65]) or synthetically generated (e.g., LogicBench [68], ProofWriter [64]), facilitating automated evaluation where Accuracy is the dominant metric. However, the uniqueness of these benchmarks lies in the specific facets of logic they target. We see a rich tapestry of challenges, from verifying natural language statements against first-order logic rules (LogicNLI [71]) and solving constraint-satisfaction puzzles (ZebraLogic [75], SATBench [69]) to generating verifiable proofs (ProofWriter [64]). This diversity allows for a fine-grained diagnosis of model capabilities, exposing specific failure points in their reasoning processes, such as compositional generalization (LLM_Compose [85], PrOntoQA-OOD [84]).

Several key trends and challenges are shaping the future of this domain. First, there is a clear push towards scalability and complexity, exemplified by datasets like LogicPro [70] with its 540,000 program-guided examples and the intricate, long-context challenges in DebateBench [82]. A second trend is the move towards programmatic and verifiable reasoning, where models generate structured outputs like code that can be executed and checked, providing more robust evaluation than simple string matching. The primary challenge remains bridging the gap between formal logic and the nuances of natural language. A further challenge is the brittleness of accuracy as a metric; future work must continue developing benchmarks that not only measure correctness but also evaluate the faithfulness and efficiency of the reasoning chain itself.

3.3.2 Specialized and Commonsense Reasoning

This category of benchmarks signifies a crucial expansion of the field, acknowledging that intelligence requires more than formal logic. It delves into the nuanced, often implicit, reasoning that underpins daily human cognition, such as understanding causality, leveraging commonsense knowledge, and performing mathematical calculations. The landscape here is newer and more diverse than that of pure logic, reflecting a frontier of active research united by the goal of quantifying abilities that are critical for real-world interaction.

While many of these benchmarks retain scalable automated evaluation, we see the introduction of more novel evaluation methods tailored to specific reasoning types. As detailed in Table 3, these include LLM-based judges to assess open-ended causal explanations (CRAB [93]) and specialized metrics like the Mahalanobis distance for evaluating analogical reasoning (ANALOGICAL [105]). Their uniqueness is their strength. Benchmarks like Corr2Cause [91] and CLadder [92] pioneer the evaluation of causal inference, a critical step towards moving models from correlation to understanding. Others, like the highly-cited StrategyQA [86] and aNLI [87], probe the implicit, multi-step, and abductive reasoning that is central to human problem-solving. Furthermore, the emergence of benchmarks for active reasoning (AR-Bench [102]) and linguistic rule induction (IOLBENCH [104]) represents a paradigm shift, moving evaluation from passive pattern recognition to active, agentic problem-solving.

3.3.3 Applied and Contextual Reasoning

This final category represents the crucible where all forms of reasoning are tested: the complex, noisy, and practical world of applied knowledge. These benchmarks assess an LLM's ability to deploy its skills to solve multi-faceted problems that mirror real-world tasks, serving as capstone evaluations of the entire pipeline of information retrieval, integration, reasoning, and synthesis. These are typically large-scale efforts, often with high citation counts (e.g., HotpotQA [99], SuperGLUE [12], ARC [98]), marking them as flagship measures of progress in artificial intelligence.

The common thread uniting these benchmarks is their grounding in realistic, web-scale data and their focus on tasks requiring integrative reasoning. For instance, HotpotQA [99] demands that a model locate and connect disparate pieces of evidence for Multi-hop Inferential Reasoning, while ARC [98] requires the application of scientific knowledge. Their uniqueness comes from the specific, complex reasoning processes they target. BIG-Bench Hard [103] is distinguished by its focus on Challenging Compositional Reasoning across 23 diverse tasks, while LiveBench [100] is particularly innovative for its use of live, Private user queries, creating a dynamic challenge that inherently resists data contamination.

A dominant trend in this area is the push for greater robustness and explainability. It is no longer sufficient for a model to produce the correct answer; benchmarks increasingly demand that the model "show its work" by providing supporting evidence (HotpotQA [99]) or by succeeding on a battery of diverse and challenging tasks (SuperGLUE [12]). The most significant challenge facing these benchmarks is data contamination. As their Web-sourced data is public, preventing test sets from leaking into training corpora is nearly impossible. The creation of dynamic, non-public benchmarks like LiveBench [100] is a direct and necessary response. Future evaluation will likely move towards more dynamic, real-time, and interactive scenarios (TextGames [106]) that test not just what a model knows, but its ability to adapt and reason in a constantly changing world.

3.3.4 Summary and Future Directions

The evaluation of reasoning in LLMs has evolved significantly, progressing from siloed tests of formal logic to complex, integrated assessments that mirror real-world demands. Our survey, summarized in Table 3, reveals a clear trajectory across three major categories. The journey begins with Logical Reasoning, where controlled, often synthetic datasets are used to probe deductive and formal abilities. It then expands into Specialized and Commonsense Reasoning, tackling more nuanced domains like causality, mathematics, and abductive inference, often requiring novel evaluation metrics beyond simple accuracy. Finally, Applied and Contextual Reasoning serves as a capstone, evaluating the synthesis of all reasoning skills on complex, multi-hop tasks drawn from web-scale data. Methodologically, this evolution is marked by a shift from human- or model-generated data towards web-crawled and now live, private data sources; a diversification of evaluation from accuracy-based automated scoring to include LLM judges and specialized metrics; and an increase in task complexity from single-step classification to interactive, multi-step generation.

Building on these trends and identified gaps, several key future directions emerge for the field:

Embracing Dynamic and Interactive Evaluation: The challenge of data contamination in static, web-sourced benchmarks (e.g., HotpotQA, SuperGLUE) is a critical threat to valid assessment. The future lies in dynamic benchmarks like LiveBench [100], which use a continuous stream of new, private data. This paradigm should be expanded. Furthermore, a move towards more interactive environments, as initiated by AR-Bench [102] and TextGames [106], is essential for evaluating agentic reasoning, where models must plan, act, and adapt based on feedback.

Deepening the Evaluation of Reasoning Processes: Current evaluations predominantly focus on the final output. Future benchmarks must increasingly scrutinize the reasoning process itself. This involves not just demanding a chain of thought, but developing metrics to assess its faithfulness, logical consistency, and efficiency. The verifiable, program-guided approach of LogicPro [70] is a promising step. Future work could involve causal tracing to understand which parts of a model's knowledge and context influenced its final decision, moving beyond correctness to true explainability.

Expanding to Underexplored Reasoning Domains and Languages: While deductive reasoning is well-covered, other critical forms of reasoning remain underexplored. The development of robust benchmarks for abductive (e.g., aNLI [87]), analogical (e.g., ANALOGICAL [105]), and especially causal reasoning (e.g., CLadder [92]) is a pressing need. Moreover, the vast majority of reasoning

benchmarks are monolingual (English). Creating multilingual and cross-lingual reasoning challenges, building on initial efforts like Multi-LogiEval [74], is vital for ensuring that progress in AI reasoning is equitable and globally applicable.

Integrating Reasoning with Action and Tools: The ultimate test of reasoning is its application to achieve goals in the world. The next frontier of evaluation will require LLMs to function as agents that use tools, search for information, and interact with complex systems. This moves beyond text-based problems to scenarios where reasoning directly informs actions with tangible outcomes, representing the convergence of reasoning, planning, and agency.

Benchmark	Focus	Language	Source	Data type	Eval.	Indicators	Amount	Method	Citations
Logical Reasoning									
RuleTaker [63]	Deductive Reasoning	Monolingual	Human	Classification	AE	Accuracy	1.2M	Yes	375
ProofWriter [64]	Proof Generation	Monolingual	Model	Hybrid	AE	Accuracy	1.2M+	Yes	275
LogicNLI [71]	First-Order Logic	Monolingual	Human	Classification	AE	Accuracy, P-EM, P-AC	96,000	No	91
Unigram-FOL [72]	First-Order Logic	Monolingual	Human	Classification	AE	Accuracy	100K	Yes	2
ReClor [66]	Reading Comprehension	Monolingual	Web	MCQA	AE	Accuracy	6,138	Yes	338
LogiQA [67]	Reading Comprehension	Monolingual	Web	MCQA	AE	Accuracy	8,678	No	316
FOLIO [65]	First-Order Logic	Monolingual	Human	Generation	AE	Accuracy	1,438	Yes	114
P-FOLIO [73]	First-Order Logic	Monolingual	Human	Classification	AE	Accuracy	19,000	Yes	5
LogicBench [68]	Logical Patterns	Monolingual	Model	Hybrid	AE	Accuracy	2,020	No	64
Multi-LogiEval [74]	Multi-task Logic	Bilingual	Hybrid	Hybrid	AE	Accuracy, F1	25,000	No	1
ZebraLogic [75]	Matrix-based Puzzles	Monolingual	Human	Generation	AE	Cell/Puzzle Acc.	1,000	No	15
FineLogic [76]	Fine-grained Logic	Monolingual	Human	Classification	AE	Accuracy	1,175	Yes	8
LogicAsker [77]	Atomic Logical Rules	Monolingual	Human	Classification	AE	Accuracy	5,200	No	20
LogicInference [78]	Long-Tailed Inference	Monolingual	Hybrid	Classification	AE	Accuracy	9,990	Yes	21
SimpleLogic [79]	Systematic Generalization	Monolingual	Human	Classification	AE	Accuracy	7,000	Yes	145
AutoLogi [80]	Logic Puzzles	Bilingual	Model	LLM	Accuracy	2,300	No	1	
SATBench [69]	SAT Problems	Monolingual	Hybrid	Classification	AE	Accuracy	2,100	No	1
PARAT [81]	SAT Solving	Monolingual	Human	Classification	AE	Accuracy	100K+	Yes	1
LogicPro [70]	Program-guided Logic	Monolingual	Model	Generation	LLM	Accuracy	540K	No	3
DebateBench [82]	Long-context Debate	Monolingual	Web	Classification	AE	Position Diff., Score	256 speeches	Yes	0
PrOntoQA [83]	"Greedy" Chain-of-Thought	Monolingual	Human	Generation	AE	Accuracy	3 tasks	Yes	323
PrOntoQA-OOD [84]	Compositional Generalization	Monolingual	Web	Generation	AE	Accuracy	1,760	Yes	77
LLM_Compose [85]	Compositional Generalization	Monolingual	Human	Generation	AE	Accuracy	Adaptive	Yes	60
Specialized and Commonsense Reasoning									
StrategyQA [86]	Multi-step Strategy	Monolingual	Human	Classification	AE	Accuracy	2,780	Yes	700
aNLI [87]	Abductive/Commonsense	Monolingual	Human	MCQA	AE	Accuracy	169K	Yes	800
CommonGen [88]	Generative Commonsense	Monolingual	Web	Generation	AE	SPICE, BLEU-4	77K	No	600
CLUTRR [89]	Inductive Reasoning	Monolingual	Human	Classification	AE	Accuracy	6,016	Yes	230
Corr2Cause [91]	Causal Reasoning	Monolingual	Hybrid	Classification	AE	Accuracy, F1	1,000	Yes	63
CRAB [93]	Causal Reasoning	Monolingual	Human	Generation	LLM	Win Rate	3,923	Yes	17
CLadder [92]	Causal Reasoning	Monolingual	Model	AE	Hybrid	Accuracy	10K	Yes	24
MathQA [94]	Mathematical Reasoning	Monolingual	Web	MCQA	AE	Accuracy	37,298	Yes	706
GSM-Symbolic [95]	Symbolic Math	Monolingual	Web	Generation	AE	Accuracy	8,500	Yes	21
Mathador-LM [97]	Mathematical Reasoning	Monolingual	Human	Generation	AE	Accuracy	N/A	Yes	8
MathEval [96]	Mathematical Reasoning	Bilingual	Hybrid	Generation	AE	Accuracy	64,171	No	23
AR-Bench [102]	Active Reasoning	Monolingual	Human	Generation	AE	Success Rate, Efficiency	5,500	Yes	0
IOLBENCH [104]	Linguistic Reasoning	Multilingual	Web	Hybrid	AE	Accuracy	1,500	Yes	1
ANALOGICAL [105]	Long-text Analogy	Monolingual	Hybrid	Classification	AE	Mahalanobis dist.	13 datasets	Yes	34
StructTest [90]	Structured Output	Monolingual	Human	Generation	AE	Rule Compliance	N/A	Yes	2
ProtoQA [41]	Prototypical Common-sense	Monolingual	Human	Generation	AE	Exact Match, Similarities, Max @K	10K	Yes	62
Applied and Contextual Reasoning									
ARC [98]	Scientific Reasoning	Monolingual	Web	MCQA	AE	Accuracy	7,787	Yes	1693
SuperGLUE [12]	Broad-Spectrum Reasoning	Monolingual	Web	Hybrid	AE	Accuracy, F1, EM	~110K	No	4000
HotpotQA [99]	Multi-hop Inferential Reasoning	Monolingual	Web	Hybrid	AE	F1, EM, Ans. Acc.	112,779	No	4100
BIG-Bench Hard [103]	Challenging Compositional Reasoning	Monolingual	Web	Hybrid	AE	Accuracy, etc.	23 tasks	No	1677
SysBench [101]	Algorithmic/Planning	Monolingual	Web	Hybrid	AE	Accuracy	10 tasks	No	11
TextGames [106]	Interactive Reasoning	Monolingual	Human	Hybrid	AE	Accuracy	8 games	No	0
LiveBench [100]	Real-world Applied Reasoning	Multilingual	Private	Generation	LLM	Win Rate	32,156	No	114

Table 3: This table provides a comprehensive overview of various benchmarks used to evaluate reasoning in LLMs, categorized into three sections. **Logical Reasoning:** This section includes benchmarks that specifically target deductive, first-order logic, and other formal reasoning abilities. **Specialized and Commonsense Reasoning:** This category covers benchmarks that evaluate reasoning in specialized domains like mathematics and broader commonsense understanding. **Applied and Contextual Reasoning:** These benchmarks assess how well LLMs can apply their reasoning skills to complex, multi-step tasks that often mirror real-world scenarios.
Abbreviations: AE: Automated Evaluation; LLM: LLM-based Judge; MCQA: Multiple Choice Question Answering; EM: Exact Match; P-EM: Probability-based Exact Match; P-AC: Probability-based Accuracy. The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

4 Domain-Specific Benchmarks

4.1 Natural Sciences

Shifting the evaluation perspective from general capabilities to specialized domains is a critical step in testing the boundaries of Large Language Models (LLMs). As one of the most logically rigorous and structurally organized areas of human knowledge, the natural sciences present a significant challenge to an LLM's knowledge base and reasoning abilities. This field, which spans core disciplines like Mathematics, Physics, Chemistry, and Biology, shares a common set of features. Success in this area not only requires a model to have good general-purpose abilities, but also demands strong capacities for abstract reasoning, symbolic manipulation, and following complex causal chains. For example, a physics problem might require the application of a specific mathematical theorem, while a model must be able to refuse to answer a chemistry question about how to make explosives.

As summarized in Table 4, this section will review and analyze these representative domain-specific benchmarks, discussing their design philosophies, evaluation dimensions, and the common challenges they face. To systematically examine the performance of LLMs in different branches of the natural sciences, existing evaluation benchmarks are typically categorized by discipline, each focusing on the unique challenges of that specific field.

4.1.1 Mathematics

Mathematics is the language of the natural science, evaluating a model's performance in this area is fundamental to measuring its abstract and logical reasoning capabilities. The evaluation of a model's mathematical capabilities is similar to human examinations, primarily utilizing multiple-choice questions and open-ended problems.

With the rapid advancement of Large Language Model (LLM) capabilities, the difficulty of mathematical evaluation benchmarks has increased.GSM8K [107], which **focus on grade school-level** word problems requiring models to perform multi-step arithmetic operations. To increase the difficulty, MATH [108] and JEEBench [111] collect problems from **high school and university entrance competitions**, covering more complex topics in algebra, geometry, and other fields. As model capabilities have grown, the difficulty of benchmarks has continued to rise, leading to several benchmarks aimed at higher levels. U-MATH [112] cover **undergraduate-level** mathematics problems; Omni-MATH [109] and MiniF2F [113] focus on **Olympiad-level** problems and formal theorem proving; while FrontierMath [110], designed by top mathematicians, aims to evaluate a model's ability to solve **cutting-edge advanced mathematics problems**, representing the current peak of difficulty in mathematical reasoning evaluation.

Mathematical evaluation benchmarks have a significant portion consists open-ended questions, with accuracy or pass rate serving as the core metric.This paradigm introduces the outcome-based problem that **open-ended problems without partial credit**, a right reasoning process can receive zero points due to a minor calculation error. To address this issue, new evaluation paradigms have been proposed. MATH-P [114] tests model robustness and generalization by applying difficult perturbations to problems; ASyMOB [119] focuses on university-level symbolic mathematical operations to assess a model's symbolic manipulation skills; and U-MATH [112] introduces **LLM-as-a-Judge** evaluation method for more nuanced assessment.

4.1.2 Physics

As a bridge connecting the abstract world of mathematics with the physical world, physics presents unique demands on the reasoning capabilities of Large Language Models (LLMs). Physics problems require not only mathematical computation but also a profound conceptual understanding, the ability to ground abstract problems in physical laws.

Early **comprehensive scientific benchmarks** laid the foundation for physical evaluation, and were subsequently followed by the emergence of more specialized and in-depth physics benchmarks. SciBench [120] is an early comprehensive science benchmark at the university level, it covers chemistry, physics, and mathematics three disciplines. It is designed to **test a model's capabilities in multi-step reasoning, understanding scientific concepts, knowledge retrieval, and complex numerical calculations.** The majority of other benchmarks follow this same evaluation paradigm, like JEEEBench [111]. PHYSICS [127] and UGPhysics [121] have built English and Chinese-English

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
Mathematics									
GSM8K [107]	Grade School Math	Monolingual	Manual Design	Generation	AE	Accuracy	85k	Yes	3068
MATH [108]	Competition Math	Monolingual	Hybrid	MCQA	AE	Accuracy	12.5k	No	1719
U-MATH [112]	Undergraduate Math	Monolingual	Std. Exams	Hybrid	LLM	Accuracy, PPV, TPR, etc.	1.1k	Yes	12
Omni-MATH [109]	Olympiad Math	Monolingual	Web	Generation	AE	Accuracy	4.4k	No	93
MiniF2F [113]	Formal Theorem	Monolingual	Open Datasets	Generation	AE	Pass rate	488	No	208
FrontierMath [110]	Advanced Math	Monolingual	Manual Design	Generation	AE	Accuracy	300	Yes	62
MATH-P [114]	Perturbed Math	Monolingual	Open Datasets	Generation	AE	Accuracy	279	Yes	20
ASyMOB [119]	Symbolic Math	Monolingual	Hybrid	Generation	AE	Accuracy	17k	Yes	0
MathBench [115]	Multi-difficulty Math	Bilingual	Hybrid	Hybrid	ME	Accuracy,CE	3.7k	Yes	80
Hardmath [116]	Graduate Math	Monolingual	Auto Design	Generation	ME	Accuracy	1.4k	Yes	14
Lila [117]	Multi-domain Math	Monolingual	Open Datasets	Hybrid	AE	F1	134k	Yes	143
HARP [118]	Olympiad Math	Monolingual	Web	Hybrid	AE	Accuracy	5.4k	No	9
Mathador-LM [97]	Math Game	Monolingual	Auto Design	Generation	AE	Accuracy	1k	Yes	8
Physics									
PHYSICS [127]	Undergrad. Physics	Monolingual	Manual Design	Generation	AE	Accuracy	1.3k	No	0
UGPhysics [121]	Undergrad. Physics	Bilingual	Web	Hybrid	ME	Accuracy	11k	Yes	5
PhysReason [124]	Multimodal Physics	Monolingual	Web	Generation	AE	Accuracy	1.2k	Yes	9
PhysicsArena [123]	Multimodal Physics	Monolingual	Web	Generation	ME	Accuracy	5.1k	Yes	0
PhysUniBench [125]	Vision-Essential	Bilingual	Std. Exams	Hybrid	ME	Accuracy	3.3k	No	0
SeePhys [122]	Vision-Essential	Bilingual	Std. Exams	Generation	ME	Accuracy	2k	No	1
FEABench [126]	Eng. Simulation	Monolingual	Web	Generation	AE	Resolved Ratio	1.4k	Yes	2
TPBench [128]	Theoretical Physics	Monolingual	Hybrid	Generation	ME	Accuracy	57	Yes	9
Chemistry									
ChemEval [129]	Foundational Chem.	Monolingual	Hybrid	Hybrid	ME	Accuracy, F1	1.5k	No	7
ChemistryQA [132]	Literature-based	Monolingual	Web	Generation	AE	Acc, Precision	4.4k	No	3
ScholarChemQA [133]	Literature-based	Monolingual	Web	MCQA	AE	Accuracy, F1	40k	No	7
MoleculeQA [130]	Molecular Prop.	Monolingual	Open Datasets	MCQA	AE	Accuracy	61.5k	No	9
ChemSafetyBench [131]	Chemical Safety	Monolingual	Web	Generation	LLM	Acc, Safety	30k+	No	2
ChemLLMBench [134]	Molecular Prop.	Monolingual	Open Datasets	Hybrid	AE	Accuracy,F1 BLEU,ROUGE Validity,Exact Match	100k	No	216
Alchemy [135]	Molecular Prop.	Monolingual	Open Datasets	Generation	AE	MAE	110k	No	100
Biology									
PubMedQA [136]	Biomedical QA	Monolingual	Web	MCQA	AE	Accuracy, F1	274k	Yes	984
BioMaze [137]	Pathway Reasoning	Monolingual	Web	Hybrid	ME	Accuracy	5.1k	No	0
SciAssess [138]	Paper Analysis	Monolingual	Open Datasets	Hybrid	AE	Acc, Recall, F1, Mol. Similarity	6.9k	No	29
BioPreDyn-bench [311]	Biological modeling	Monolingual	Manual Design	Generation	AE	NRMSE	6	No	99
BixBench [139]	Computational Biology	Monolingual	Manual Design	Generation	LLM	Accuracy	296	Yes	8
LAB-Bench [140]	Biology research	Monolingual	Hybrid	MCQA	AE	Accuracy Precision,Coverage	2.4k	Yes	54
AutoBio [141]	Biological experiment	Monolingual	Manual Design	Generation	AE	Accuracy	100	Yes	2
BioProBench [142]	Biological experiment	Monolingual	Open Datasets	Hybrid	AE	Acc,F1,Recall BLEU,METEOR ROUGE-L EM	556k	Yes	1
Cross-Disciplinary									
JEEBench [111]	Math Physics Chemistry	Monolingual	Web	Hybrid	ME	Accuracy	515	No	63
SciBench [120]	Math Physics Chemistry CS	Monolingual	Web	Generation	ME	Accuracy	972	Yes	162
TheoremQA [143]	Multi-domain	Monolingual	Manual Design	Hybrid	AE	Accuracy	800	No	150
OpenBookQA [148]	Sci. Commonsense	Monolingual	Manual Design	MCQA	AE	Accuracy	5.9k	No	1171
GPQA [15]	Physics, Chem., Biology	Monolingual	Manual Design	MCQA	AE	Accuracy	1.1k	No	724
WMDP [149]	Biology,Chem.,CS	Monolingual	Web	MCQA	AE	Accuracy	3.6k	Yes	226
Curie [150]	Physics,Chem., Biology	Monolingual	Web	Generation	ME	ROUGE-L, BERTScore F1, IoU, IDr, LMScore, LLMSim	580	Yes	2

Table 4: Summary of representative benchmarks in the Natural Sciences. Data sources are abbreviated as Std. Exams (Standardized Exams). The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

undergraduate-level physics problem sets. Respectively, UGPhysics [121] specifically designed to prevent data leakage and it was found that LLMs specialized for mathematics is not always outperform other models, while PhysReason [124] and PhysicsArena [123] have introduced a large number of multimodal problems that require analysis of diagrams and charts.

Diagrams are often function as an indispensable part of the problem itself in physics. Therefore, multimodal questions are an essential component of physics evaluation benchmarks.PhysUniBench [125] equips each problem with a corresponding diagram, while SeePhys [122] designs the majority of its problems to be vision-essential. Furthermore, PhysicsArena [123] introduces a fine-grained multimodal evaluation paradigm based on the physics problem-solving process, includes variable identification, physical process modeling, and reasoning-based solving 3 stages, moving beyond a single answer judgment.

Trends and challenges Physics reasoning is fundamentally more than solving "mathematical word problems," as it requires a unique, comprehensive set of capabilities that transcend simple mathematics, including **Conceptual Grounding, Multimodal Interpretation, and Process Formulation**.

Like the evaluation of mathematical problems, a simple outcome-based approach cannot evaluate a model's capabilities in physics comprehensively. Such as the MARJ framework within UGPhysics and methods like LLM-as-a-Judge are trying to overcome this limitation. However, a model's physics capabilities must be grounded in practical applications, requiring it to move beyond rote problem-solving to the construction of accurate physical models for real-world scenarios. Like the way forged by FEABench [126], the evaluation criteria shift from correctness to the ability to construct a valid physical model and derive verifiable results from simulation software.

4.1.3 Chemistry

Chemical evaluation benchmark not only focus on traditional problem-solving abilities but also extends to the critical areas of factual accuracy, the comprehension of literature, and the model's understanding of safety.ChemEval [129] have established multi-level evaluation systems to evaluate model's foundational knowledge. ChemistryQA [132] and ScholarChemQA [133] extract questions from chemical literature and papers to assess a model's understanding of scientific texts. Regarding the evaluation of LLMs in subfields of chemistry, MoleculeQA [130] builds a large-scale dataset specifically to evaluate the capability of model's regarding molecular structure, properties, and more. ChemSafetyBench [131] as a pioneering work in chemical safety area, construct a massive test set of over 30,000 samples to systematically evaluate a model's safety and responsibility when handling potentially hazardous chemical knowledge.

Chemistry benchmarks uniquely place non-technical and social dimensions—accuracy and safety—at the core of their evaluation, to an extent that surpasses benchmarks in mathematics and physics. While the focus in mathematics and physics remains on the correctness of solutions and the logical consistency of the reasoning process, chemistry domain has build benchmarks like MoleculeQA [130], which is entirely focus on verifying factual accuracy, and ChemSafetyBench [131], designed to assess safety and ethical risks.This divergence in focus origin from the different real-world implications of these disciplines: an incorrect mathematical answer is merely an error, but an incorrect chemical statement can be actively dangerous. When an LLM generates false information about molecular properties or provides synthesis methods for hazardous substances, it can directly lead to real-world harm. This indicates that as LLM evaluation engages with disciplines more closely tied to the real world, the standard of a "good" model is expanding beyond mere problem-solving ability to encompass a broader range of capabilities, including reliability, trustworthiness, and ethical alignment.

4.1.4 Biology

Biology evaluation benchmark primarily focus on the comprehension of relevant scientific literature.Building on classic biomedical question-answering benchmarks like PubMedQA [136], new benchmarks are expanding into more specialized and in-depth reasoning tasks. BioMaze [137] focuses on reasoning about biological pathways, requiring models to understand and predict the downstream effects that arise when a biological system is subjected to interventions, such as gene mutations, viral infections, or drug treatments. SciAssess [138] is dedicated to evaluating a model's ability to analyze biology literature in real scientific research scenarios, with tasks ranging from basic knowledge to advanced analytical reasoning. AutoBio [141] and BioProBench [142] introduce a new paradigm for assessing biological competence by conducting biological experiments or evaluating experimental protocols to test the LLM's understanding of experimental standards.

The uniqueness of biology lies in its vast, fragmented, and often incomplete knowledge graph. While complexity exists in other scientific fields, in biology, it is concentrated in the complex, multi-step biological pathways composed of genes, proteins, and metabolites. Early benchmarks primarily tested text comprehension. However, BioMaze [137] highlights that true biological reasoning involves understanding complex networks, where a minor perturbation can trigger a cascade of non-linear biological chain reactions. It introduced the PATHSEEKER agent, which integrates an LLM with structured navigation of a biological knowledge graph, propelling "Graph-Augmented LLMs" as a highly promising direction in the field of biology.

4.1.5 Cross-Disciplinary and General Scientific Abilities

True scientific research is often interdisciplinary, so a series of benchmarks have been designed to evaluate a model's comprehensive scientific capability.**Comprehensive Problem Solving** benchmarks like JEEBench [111] (Physics, Chemistry, Mathematics), SciBench [120](Physics, Chemistry,

Mathematics, Computer Science), and GPQA [15] (Biology, Physics, Chemistry) assess a model's overall problem-solving skills through difficult questions spanning multiple disciplines. GPQA [15] in particular, is authored by domain experts and is designed to be "Google-Proof" to effectively test models.**Higher-Order Reasoning and Tool Use** benchmarks like TheoremQA [143] requires models to apply theorems from disciplines like mathematics and physics to solve problems across fields. LLM-SRBench [147]focuses on discovering equations from data. In natural science domain is also gradually beginning to test their **ability to use tools.** SCITOOLBENCH [144] provides a series of API tools that models must call to solve complex scientific calculations and reasoning tasks, has taken a key step in this direction. In addition to specialized domains, benchmarks like OpenBookQA [148], SciEval [146], and SciKnowEval [151] assess a model's **common sense science knowledge**, multi-level scientific knowledge, and research capabilities from an overall perspective, while CURIE [150] focuses on evaluating a model's understanding in long scientific literature text. Natural sciences evaluation benchmarks are gradually expanding to encompass other disciplines. GEO-Bench [152] uses Earth monitoring data to evaluate pre-trained models in processing geospatial data.

Comprehensive scientific evaluation benchmarks are evolving from testing a model's static knowledge reserve to measuring its dynamic, process-oriented application capabilities. As the development of general models increasingly orients towards creating usable "research assistants," evaluation benchmarks may also transcend the boundaries of "problem-solving." It is vital to build interactive environments capable of assessing scientific research abilities. That these benchmarks will not only provide a more holistic measure of a model's utility but also play a crucial role in breakthroughs for artificial general intelligence in the natural science domain.

4.1.6 Summary and Future Directions

Evaluation benchmarks across the natural sciences and other domains face similar critical challenges. **A primary concern is data contamination**. if evaluation data is included in model's training set, the assessment becomes an "open-book exam" that cannot truly measure the model's reasoning capabilities.

Furthermore, the reliability of evaluation methods is under scrutiny. Traditional methods based solely on final answers are often deemed insufficient, while some paradigms like "LLM-as-a-Judge" are also being questioned for their robustness. Indeed, some reviewers have noted the self-contradiction within LLM-generated evaluation reports, casting further doubt on the reliability of these automated assessment methods.

Evaluating the generalization capabilities of model is a core challenge for benchmarks, which must determine whether a model has truly understood the knowledge or has merely memorized solution templates for similar problems. Perturbation-based benchmarks, such as MATH-Perturb [114] and ASyMOB [119], provide a potential approach. By making some modifications to problems, they reveal that even SOTA models often rely on "shortcut learning" rather than generalizable reasoning.

To address the challenges, several new evaluation paradigms are being explored. In the domain of the natural sciences, the role of an LLM is shifting from a simple knowledge retrieval tool to a research assistant. Therefore, evaluation benchmarks are **transitioning from assessing LLMs as "knowledge bases" to evaluating their capabilities as agents.** Following the path forged by benchmarks like SciAgent [144], FEABench [126], and BioMaze [137], the focus now is on assessing an LLM's ability to use provided tools to approach a goal.

A move towards Holistic, Multi-faceted Frameworks. Evaluation is shifting from a single score to a comprehensive assessment of a model's overall capabilities. Benchmarks such as ChemEval [129], SciKnowEval [151], and PhysicsArena [123] build multi-dimensional or multi-stage frameworks to fine-grained evaluate model capabilities, thereby providing more actionable guidance for model improvement.

4.2 Humanities & Social Sciences

Beyond the rational evaluation of large language models (LLMs) in the natural sciences, their anthropomorphic conversational traits enable more natural and effective communication with humans, enhancing interactive applications. Social sciences, as one of the most human-centered fields, play a crucial role in this context. A key question is whether LLMs can effectively address real-world challenges in areas such as **Law**, **Intellectual Property (IP)**, **Education**, **Psychology**, and

Finance. For example, can LLMs comprehend human emotions well enough to provide meaningful emotional support? Can they reliably retain and apply legal knowledge to offer sound legal advice? This section focuses on the human-centered capabilities of LLMs within social science domains. It reviews and analyzes relevant benchmarks that investigate these aspects, examining their task design principles, data construction methods, and evaluation strategies — including those involving subjective judgments.

All these humanities and social sciences domains are highly applicable in real-world scenarios. One of the biggest challenges is determining how to evaluate an LLM's knowledge within these domains, which involves defining appropriate tasks, constructing relevant datasets, and selecting suitable evaluation methods. These three key aspects are precisely what existing domain-specific benchmarks focus on and claim to address. In this section, we will follow the structure of these three aspects—task definition, dataset construction, and evaluation methods—to present the content for each domain, including Law, Intellectual Property (IP), Education, Psychology, and Finance. We provide detailed information on the representative benchmarks in the humanities and social sciences discussed in each subsection below, as summarized in Table 5.

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
Law									
LegalBench [153]	Law	Monolingual	Hybrid	Hybrid	ME	Exact Match, F1-Score Correct, Analysis	91,206	No	162
LBOX OPEN [155]	Law	Monolingual	Open Datasets	Hybrid	AE	Exact Match, F1-Score	147K	Yes	51
LawBench [154]	Law	Monolingual	Hybrid	Hybrid	AE	Accuracy, ROUGE F-Score, nLog-distance	–	No	133
LAiW [157]	Law	Monolingual	Open Datasets	Hybrid	ME	Accuracy, Miss Rate, F1-Score Entity-Acc, ROUGE, Win Rate	11,605	No	39
LexEval [156]	Law	Monolingual	Hybrid	Hybrid	AE	Accuracy, ROUGE	14,150	No	24
CiteLaw [158]	Law	Monolingual	Hybrid	Generation	AE	MAUVE, ROUGE, Citation Quality	1,000	No	3
CaseGen [159]	Law	Monolingual	Open Datasets	Generation	ME	ROUGE, BertScore, LLM Judge	2,000	No	3
Intellectual Property (IP)									
PatentEval [160]	IP	Monolingual	Open Datasets	Generation	ME	SemSim, N-grams Coverage FactGraph, QAFactEval, EntityGrid	400	No	4
MoZIP [161]	IP	Multilingual	Web	Hybrid	AE	Accuracy	3,121	Yes	8
IPEval [163]	IP	Bilingual	Std. Exams	MCQA	AE	Accuracy	2,657	No	3
D2P [164]	IP	English	Open Datasets	Generation	ME	ROUGE, BLEU, BERTScore	1,933	Yes	12
IPBench [162]	IP	Bilingual	Hybrid	Hybrid	ME	Accuracy,Exact-Match,LLM Score BLEU, ROUGE, BertScore	10,374	No	0
Education									
E-Eval [165]	Education	Monolingual	Web	MCQA	AE	Accuracy	4,352	No	9
EduBench [166]	Education	Bilingual	Model	Generation	ME	Scenario Adaptation Criteria, Factual Reasoning Accuracy Criteria Pedagogical Application Criteria	4,019	No	0
Psychology									
CPsyExam [167]	Psychology	Monolingual	Web	MCQA	AE	Accuracy	22,400	No	0
CPsyCoun [168]	Psychology	Monolingual	Web	Generation	AE	Comprehensiveness, Professionalism Authenticity, Safety	4,700	Yes	32
Psycollm [169]	Psychology	Monolingual	Std. Exams	Hybrid	AE	ROUGE, BLEU, BERTScore Standard Accuracy, Elastic Accuracy	3,863	Yes	27
Psychometrics Benchmark [171]	Psychology	Monolingual	Hybrid	Hybrid	AE	Standard Deviation, Match Rate Cohen's kappa coefficient, Agreement Rate	3,545	No	43
PsychoBench [312]	Psychology	Monolingual	Std. Exams	Hybrid	AE	Mean and Standard Deviation of the Scale	512	No	38
Finance									
BBT-CFLEB [174]	Finance	Monolingual	Web	Hybrid	AE	Rouge, F1, Accuracy	220M	Yes	86
FLARE [173]	Finance	Monolingual	Open Datasets	Hybrid	AE	Accuracy, F1, Exact-Match	136K	Yes	202
FinEval [172]	Finance	Monolingual	Web	Hybrid	AE	Accuracy, ROUGE	8,351	No	51

Table 5: Summary of representative benchmarks in the humanities and social sciences. Data sources are abbreviated as Std. Exams (Standardized Exams). Evaluation methods (listed under the Eval. column) are abbreviated as AE (Automated Evaluation) and ME (Mixed Evaluation). Abbreviations used in the Data Type column include MCQA (Multiple-Choice Question Answering). The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

4.2.1 Law

Legal Task Taxonomy. The most practical application of the legal domain in the real world is to provide legal support for clients. This requires LLMs to memorize concise statutes, apply them appropriately, and respond with clear and logical reasoning. Existing benchmarks adopt different philosophical approaches to task taxonomy in the legal domain, dividing real-world application tasks into distinct levels of cognition. Bloom's Taxonomy [313] is a commonly used framework for categorizing the cognitive capabilities of LLMs, with its latest version comprising the following levels: **Remember**, **Understand**, **Apply**, **Analyze**, **Evaluate**, and **Create**. LawBench [154] proposed a legal task taxonomy based on Bloom's Taxonomy for the domain of Chinese judicial jurisdiction, classifying legal-related abilities into three cognitive levels: **(a) legal knowledge memorization**, (**b) legal knowledge understanding**, and **(c) legal knowledge application**. These three cognitive levels are a direct adaptation and summarization of Bloom's Taxonomy, and are further divided into 20 fine-grained tasks, including article and knowledge recall, document-level element recognition and information processing, as well as legal reasoning involving penalties based on real-world cases.

LexEval [156], also a benchmark for the Chinese judicial domain and built upon Bloom's Taxonomy, proposes a Legal Cognitive Ability Taxonomy (LexAbility Taxonomy), which is similar to that of LawBench but more fine-grained, comprising six dimensions: **Memorization**, **Understanding**, **Logical Inference**, **Discrimination**, **Generation**, and **Ethics**. LexEval is more fine-grained than LawBench, comprising 23 tasks spanning six cognitive levels. Beyond what LawBench covers, it also incorporates key aspects of the legal domain, such as the evolution of law and ethical considerations, particularly in relation to bias, discrimination, and privacy. This helps evaluate whether LLMs effectively capture the nature of this specific legal domain framework for their human-like chat support, without introducing intrinsic bias. Beyond Bloom's Taxonomy, LAiW [157], based on legal domain practice and also focused on the Chinese judicial jurisdiction, proposes a syllogism-based taxonomy reflecting the thinking process of legal experts and legal practice, classifying tasks into three levels from easy to difficult: **basic information retrieval**, **legal foundation inference**, and **complex legal application**. This benchmark focuses on concrete, real-world legal practice. Beyond legal reasoning and case understanding, it introduces element recognition and Named Entity Recognition (NER) tasks, specifically tailored for legal domain retrieval.

In addition to legal benchmarks specifically designed for the Chinese judicial domain, there are also benchmarks developed for other languages and legal systems. LBOX OPEN [155] is the first large-scale legal benchmark specifically designed for the Korean judicial domain. Unlike cognition-oriented evaluations, it focuses on legal case analysis, including two tasks for case name and statute classification, two tasks for legal judgment prediction, and one for case summarization. To better understand the legal reasoning capabilities of LLMs, LegalBench [153] focuses primarily on legal reasoning and comprises 162 tasks, including **issue spotting**, **rule recall**, **rule application and conclusion**, **statutory interpretation**, and **rhetorical understanding**. Each of these five task types includes more scene-specific subtasks with substantial real-world application value—such as applying diversity jurisdiction tests based on information about the plaintiff, defendant, and the amount in controversy for different claims, which requires both arithmetic and logical reasoning.

All the legal benchmarks above evaluate the comprehensive legal capabilities of LLMs; however, there are also benchmarks that focus on specific legal abilities within particular scenarios. As a result, their task taxonomies are more fine-grained and concrete. CiteLaw [158] aims to evaluate LLMs whether could produce legally sound responses with appropriate citations. Legal case documents play a critical role in the legal domain. CaseGen [159] proposes an automated legal document generation task taxonomy within the Chinese judicial domain, including tasks such as **drafting defense statements**, **writing trial facts**, **composing legal reasoning**, and **generating judgment results**. These tasks require LLMs not only to accurately recall and apply legal knowledge, but also to process the internal logic within the concrete context of each case.

Different judicial domains exhibit significant differences in legal practice. As the benchmarks mentioned above are monolingual, they typically correspond to the legal systems of a single country. There is a lack of multilingual legal benchmarks for LLMs, highlighting the need to design a cross-jurisdictional evaluation task taxonomy—such as tasks focused on legal comparison across countries and recognition of cross-system differences—which could serve as a promising direction for future research in advancing legal AI.

Legal Dataset Source and Construction. Real-world legal documents and cases are the most important data sources for these benchmarks. LBOX OPEN [155] compiles its corpus from Korea's first- and second-instance trials as well as Supreme Court precedents.However, most benchmarks such as LawBench [154], LegalBench [153], and LexEval [156] rely on existing datasets derived from legal competitions like CAIL and LAIC, or on publicly available legal corpora—most of which were originally created for non-LLM evaluation settings. Considerable effort has been made to recompile and adapt these existing datasets. For example, LegalBench [153] restructures the CUAD dataset [314] to formulate a binary classification task for each type of contractual clause. The last approach to dataset construction in these legal benchmarks involves substantial human effort, often including the participation of legal professionals to manually create task-specific data. This not only enhances the domain expertise of the dataset but also reduces the risk of data leakage.

Future Direction for Better Legal AI. There are already many legal benchmarks available for individual countries—particularly for China, the United States, and Korea—but there is a lack of benchmarks that cover multiple languages or judicial domains, which are essential for evaluating

multilingual LLMs. However, this is not merely about introducing additional languages into the legal domain; rather, it encourages researchers to consider real-world legal demands by developing more relevant and comprehensive task taxonomies—such as capturing the differences between Chinese and U.S. legal systems, or between civil law and common law traditions. Moreover, there is a lack of benchmarks addressing multimodal scenarios in the legal domain. Researchers should explore the multimodal demands of legal practice and leverage multimodal LLMs to meet these needs.

4.2.2 Intellectual Property

Intellectual property (IP) is an emerging field that is attracting increasing attention from natural language processing researchers. This is due to the diverse nature of intellectual property—including patents, copyrights, trademarks, and more—and its inherent legal significance in protecting the value and originality of creators' work. In essence, intellectual property possesses a dual attribution: legal and technical. Among the various IP mechanisms, patents have been the most extensively researched. Before the emergence of LLMs, researchers primarily focused on three key areas: patent retrieval, patent classification, and patent generation. With the advancement of LLMs, processing the complex legal language of patents has become easier, enabling researchers to extend these tasks by leveraging LLM capabilities.

PatentEval [160] focuses on patent generation tasks—particularly abstract generation and next-claim prediction—by evaluating models' patent drafting capabilities and introducing a comprehensive error taxonomy. D2P [164] was introduced to generate complete, long-form patent documents from user-provided drafts, simulating real-world patent application scenarios. It also proposes a multi-agent framework, AutoPatent, to handle this complex generation task. Both of these benchmarks focus on evaluating the generation capabilities of LLMs in the intellectual property domain, where models must not only handle complex technical terminology and sentence structures, but also align with the legal language style and content specificity characteristic of IP documents. Both benchmarks aim to facilitate patent drafting and reduce the reliance on manual effort in these scenarios.

Beyond patent drafting and generation, there are also benchmarks designed to evaluate models' capabilities in processing IP-related legal knowledge and technical content. MoZIP [161] is a multilingual benchmark covering ten different languages, comprising three tasks: IPQuiz, IPQA, and PatentMatch. The IPQuiz and IPQA tasks primarily focus on the legal aspects of intellectual property, aiming to evaluate models' knowledge and their ability to apply it in practical IP scenarios. The PatentMatch task places greater emphasis on the technical aspects of patents, requiring the model to select the most similar patent from four carefully constructed candidates. A later literature, IPEval [163], focuses solely on the legal aspects of patents. It uses patent bar exam data in both Chinese and English, presented in a multiple-choice question answering format.

The most recent study, IPBench [162], introduces the most comprehensive taxonomy in this domain with four hierarchical levels: **Information Processing**, **Logical Reasoning**, **Discriminant Evaluation**, and **Creative Generation**. This taxonomy is based on the American educational evaluator Norman L. Webb's Depth of Knowledge Theory (DOK) [315], which categorizes students' cognitive levels into four tiers: Recall and Reproduction, Skill and Concept Application, Strategic Thinking, and Extended Thinking. From this cognitive perspective, it comprises 20 fine-grained tasks across the four levels, covering eight different IP mechanisms. These tasks—including multiple-choice question answering, patent classification, and generation—span a wide range of cognitive and practical demands, from IP-related legal memorization and interpretation to process guidance reflecting real-world practice. They also cover tasks such as infringement behavior determination and compensation calculation, requiring models not only to possess concrete legal knowledge, but also to demonstrate reasoning and mathematical computation abilities.

Patent Data Source. Similar to legal benchmarks, the China National Intellectual Property Administration (CNIPA), the United States Patent and Trademark Office (USPTO), and the European Patent Office (EPO)—particularly the USPTO and EPO—provide patent access APIs. Additionally, the Google Patents Dataset offers relevant corpora. Benchmarks such as PatentEval [160], D2P [164], and MoZIP [161] utilize patent data from these sources. IPEval [163] uses questions from national standard patent bar exams, while MoZIP's IPQuiz and IPQA tasks collect FAQs from the official websites of IP organizations and agencies worldwide. In contrast, IPBench [162] leverages extensive expert annotation to construct datasets that align with real-world application needs. All of these benchmarks are limited to the text modality, using only the textual portions of patents. However,

patents also contain images, and domains such as trademarks and logos involve even richer visual information. Future research should focus on integrating and categorizing multimodal IP tasks to enable more comprehensive and intelligent IP services.

4.2.3 Education

Existing benchmarks such as GPQA [15] and MMLU [13] have been proposed to evaluate the knowledge levels of LLMs (e.g., at the graduate level). Although possessing relevant knowledge is a prerequisite for LLMs to support educational outcomes, these benchmarks differ from education-oriented ones, particularly in their alignment with real-world educational scenarios. Especially for kindergarten through twelfth grade (K–12) education in the Chinese context, E-Eval [165] focuses on real-world classroom scenarios spanning primary, middle, and high school levels. It categorizes tasks into two main types: **arts** (including Chinese, English, and History) and **sciences** (such as Mathematics, Physics, and Chemistry). However, it is not sufficient for LLMs to merely know, understand, and apply knowledge. The education domain is complex and involves multiple stages, such as student preview, in-class teaching, and student ability assessment, among other concrete educational scenarios. These scenarios require dedicated design and exploration of effective methods to integrate LLMs into each stage of the educational process.

EduBench [166] takes into account the practical nature of the education domain, covering nine core educational scenarios and over 4,000 diverse synthetic education tasks. These tasks can be categorized into two types based on different teaching targets: **Student-Oriented Scenarios**, which include Problem Solving, Error Correction, Idea Provision, Personalized Learning Support, and Emotional Support; and **Teacher-Oriented Scenarios**, which include Question Generation, Automatic Grading, Teaching Material Generation, and Personalized Content Creation. These tasks are highly practical, but unlike E-Eval [165], which adopts a multiple-choice question answering format, they pose greater challenges in evaluating LLM performance within each specific scenario. For each scenario, EduBench [166] evaluates whether LLMs can fulfill the teaching objectives and scenario-specific expectations across three dimensions: Scenario Adaptation, Factual & Reasoning Accuracy, and Pedagogical Application, using DeepSeek-V3 as the evaluator.

Although EduBench [166] provides a richer simulation of educational scenarios compared to E-Eval [165], making it more suitable for this domain, it offers only a coarse-grained framework. The education domain remains underexplored, and researchers need to focus on more fine-grained educational scenarios and investigate real-world teaching practices. This includes incorporating multimodal information and exploring how to better leverage LLMs to assist teachers in class preparation, as well as helping students learn effectively and correct their knowledge, ultimately benefiting their exam performance and personal knowledge framework construction.

4.2.4 Psychology

Recent literature has increasingly focused on human health in the context of using large language models (LLMs); however, most studies emphasize physical or biological aspects of health, with comparatively less attention given to mental health. Given the human-like conversational style of interactions with LLMs, it is possible for them to provide support for mental health, such as offering psychological insights and advice. CPsyExam [167] is designed to benchmark the capabilities of LLMs in understanding psychological concepts through the Chinese Standard Examination.Its taxonomy includes two task types—**psychological knowledge** and **case analysis skills**—and incorporates three question formats: single-choice, multiple-choice, and question-answering. However, similar to the education domain, in this domain closely related to human interaction, it is important to explore concrete, context-specific practices. Relying solely on knowledge-based evaluation formats is insufficient to deeply assess LLMs' capabilities in the psychological domain. Motivated by this, Zhang et al. [168] propose CPSYCOUN, a benchmark aimed at evaluating multi-turn dialogues between humans and LLMs to explore whether LLMs can be effectively applied in Chinese psychological counseling scenarios. The benchmark covers nine topics and seven classic schools of psychological counseling, including Psychoanalytic Therapy and Cognitive Behavioral Therapy. Although CPsyCoun [168] recognizes the practical need to use LLMs for serving people and evaluating them in psychological settings, it still lacks a more fine-grained taxonomy of psychological scenes and tasks. Moreover, both CPsyExam and CPsyCoun focus primarily on the Chinese psychological context, leaving a gap in multilingual benchmarks for this human-centric domain. Psycollm [169] also focuses

on Chinese, constructing a benchmark based on authoritative psychological counseling examinations in China, using both single-turn and multi-turn question answering formats, including assessments of **professional ethics**, **theoretical proficiency**, and **case analysis**. These types of questions are highly related to the scenarios.

Beyond the psychological services, many literatures explore whether LLMs possess attributes similar to those of humans and examines the stability of these attributes. Psychometrics Benchmark [171] introduces a framework leveraged to measure LLMs' psychological attributions, covering six dimensions: personality, values, emotion, theory of mind, motivation, and intelligence. PsychoBench [170] also focuses on LLMs' personality, temperament, and emotion. Using thirteen scales commonly applied in clinical psychology, PsychoBench further classifies these scales into four categories: **personality traits**, **interpersonal relationships**, **motivational tests**, and **emotional abilities**. Both benchmarks use scenario-based scales to quantify models' psychological characteristics.

4.2.5 Finance

Financial technology has advanced alongside the development of language models. Before the surge of LLMs, many finance-specific models were built on BERT as their backbone, such as FinBERT and FLANG. BBT-CFLEB [174] was introduced to benchmark model performance—particularly that of BERT-based models—using data primarily composed of financial news and social media content. Although LLM performance on BBT-CFLEB has not been reported, it represents an important component of financial news analysis, which involves many domain-specific terms. PIXIU [173] is a benchmark specifically designed for financial LLMs. It includes not only NLP-related tasks but also financial prediction tasks such as stock movement prediction, which require deeper domain-specific knowledge. As a real-world and practice-oriented domain, finance also contains many scenarios that remain to be explored. FinEval [172] categorizes financial knowledge and practical abilities into four key types: **Financial Academic Knowledge**, **Financial Industry Knowledge**, **Financial Security Knowledge**, and **Financial Agent**. These categories cover areas such as financial and economic knowledge, the use of financial tools, and financial reasoning. All of these financial benchmarks are monolingual, highlighting the need for researchers to explore financial environments across different countries and languages.

4.2.6 Summary and Future Directions

Fine-grained Task Taxonomy. All these human and social science domains are highly grounded in real-world practice and hold significant practical value, especially for individuals. This characteristic makes the definition of concrete, scenario-based tasks particularly important. Therefore, before evaluating the capabilities of LLMs in these domains—especially in finance, psychology, and education—it is essential to establish a fine-grained task taxonomy.

Developing Robust Methods for Evaluation. In domains with rich real-world applications, relying solely on multiple-choice questions is insufficient for evaluation. There is an urgent need to develop scenario-oriented evaluation methods to better assess the concrete performance of language models. This necessitates that researchers, grounded in a fine-grained taxonomy, design evaluation formats that are more detailed, concrete, and robust for these domains.

4.3 Engineering & Technology

The engineering and technology domain represents a crucible for Large Language Models, testing their capabilities on tasks that demand not only linguistic fluency but also logical rigor, functional correctness, and deep, specialized knowledge. Unlike general-purpose tasks, engineering applications often have a single correct answer or a narrow range of acceptable solutions, governed by physical laws, mathematical principles, or strict syntax. Success in this area requires models to act as functional tools rather than just conversational partners. Consequently, this field has fostered some of the most sophisticated and mature evaluation frameworks. This section surveys the landscape of engineering benchmarks, tracing their evolution from foundational code generation to complex, multi-domain problem-solving across software, electrical, mechanical, and other engineering disciplines. A summary of representative benchmarks is provided in Table 6.

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
Software Engineering & Information Technology									
HumanEval [175]	Code Generation	Monolingual	Manual Design	Generation	AE	Pass@k	164	No	4970
MBPP [176]	Python Code Gen	Monolingual	Manual Design	Generation	AE	Pass@k	974	No	1935
SWE-bench [186]	GitHub Issue Repair	Monolingual	Open Datasets	Generation	AE	Pass Rate	2294	No	680
LiveCodeBench [178]	Live Contest Programming	Monolingual	Web	Generation	AE	Pass@1	511	Yes	401
ClassEval [184]	Class-level Gen.	Monolingual	Manual Design	Generation	AE	Pass@k, DET	100	Yes	154
xCodeEval [195]	Multilingual Tasks	Multilingual	Web	Hybrid	AE	Pass@k, macro-F1	7514	Yes	53
CodeReview [199]	Code Review	Multilingual	Web	Generation	ME	Acc, BLEU, EM	534k	No	202
Spider [200]	Cross-Domain SQL	Monolingual	Hybrid	Generation	AE	EM, EX	10k	No	1471
BIRD [202]	Large-scale SQL	Monolingual	Hybrid	Generation	AE	Execution Acc	12.7k	No	479
CoSQL [203]	Conversational SQL	Monolingual	Hybrid	Generation	AE	Acc, BLEU	3k	No	110
IaC-Eval [204]	IaC Generation	Monolingual	Manual Design	Generation	AE	Pass@k	458	Yes	10
OpsEval [205]	AIOps	Bilingual	Hybrid	Hybrid	ME	FAE-Score	9070	Yes	7
FrontendBench [211]	Frontend Dev.	Monolingual	Hybrid	Generation	AE	PassRate, Consistency	148 pairs	Yes	0
Electrical & Electronic Engineering									
VerilogEval [212]	Verilog Generation	Monolingual	Hybrid	Generation	AE	Pass@k	156	Yes	218
RTLLM [213]	RTL Generation	Monolingual	Manual Design	Generation	AE	Syntax/Func, Quality	30	Yes	182
CIRCUIT [214]	Analog Circuit	Monolingual	Manual Design	Generation	ME	Global Acc, Template Acc	510 pairs	Yes	0
ElecBench [221]	Power Dispatch	Monolingual	Hybrid	Hybrid	ME	Authenticity, Logicality, Stability, Security, etc.	N/A	Yes	0
FIXME [215]	HW Verification	Monolingual	Hybrid	Hybrid	AE	Pass rate, Correctness	180 tasks	Yes	0
Mechanical, Manufacturing, Aerospace & Transportation Engineering									
CADBench [216]	CAD Script Gen.	Monolingual	Hybrid	Generation	ME	Avg scores, Syntax error	700	Yes	0
LLM4Mat-bench [217]	Material Property	Monolingual	Web	Hybrid	AE	MAD:MAE ratio, AUC	1.98M	Yes	14
AeroMfg-QA [223]	Aerospace Mfg.	Monolingual	Manual Design	MCQA	AE	Custom score, Acc	2480	Yes	0
RepoSpace [224]	Aerospace Repo.	Monolingual	Private	Generation	AE	Rouge-L, CodeBLEU	825	Yes	0
Aviation-Benchmark [218]	Aviation Domain	Monolingual	Open Datasets	Hybrid	AE	Accuracy	150k	Yes	0

Table 6: Summary of representative benchmarks in Engineering & Technology. AE: Automated Evaluation; ME: Mixed Evaluation; Gen.: Generation. The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

4.3.1 Software Engineering and Information Technology

As the discipline most intertwined with the development of AI, software engineering has the most extensive and mature collection of benchmarks. These evaluations span the entire software development lifecycle, from initial ideation to long-term maintenance.

Software Development and Maintenance The journey of evaluation began with foundational **code generation** tasks. Benchmarks like HumanEval [175] and MBPP [176] established the now-standard paradigm of assessing function-level code synthesis from natural language prompts, using functional correctness (pass@k) via unit tests as the primary metric. This initial focus quickly expanded to address greater complexity and realism. For instance, APPS [177] and USACO [182] introduced problems from programming competitions, demanding more advanced algorithmic reasoning. To combat the pervasive issue of benchmark contamination, LiveCodeBench [178] and its expert-level successor LiveCodeBenchPro [183] pioneered the use of problems from live, ongoing contests, ensuring that models are evaluated on truly unseen data.

The scope of generation tasks has also broadened from simple, self-contained functions to more complex software artifacts. ClassEval [184] was the first to specifically target class-level generation, a crucial step for evaluating object-oriented programming skills. The evaluation of domain-specific code generation has also become a major trend, with benchmarks like DS-1000 [180] for data science libraries, BioCoder [181] for bioinformatics, and the recent MMCode [185], which challenges models with multimodal problems containing visual information like diagrams.

Beyond initial creation, a model's ability to work with existing code is critical. Comprehensive frameworks for **code understanding and completion**, such as CodeXGLUE [194] and the multilingual xCodeEval [195], offer a suite of tasks including code summarization, translation, and retrieval. Other benchmarks target more specific understanding tasks, such as code question-answering (CodeQA [196]), code search within large corpora (Cosqa [197]), and repository-level code completion that mimics a developer's IDE experience (Repobench [198]).

Code maintenance, a significant portion of real-world software engineering, is another vital evaluation area. Automated program repair is a key focus, with benchmarks like RepairBench [187] and the highly influential SWE-bench [186]. The latter is particularly notable for sourcing its tasks directly from real GitHub issues and pull requests in popular open-source projects, providing an unparalleled level of realism. Complementary benchmarks like Debugbench [188] and Condefects [189] focus on the related skills of debugging and defect localization. Furthermore, precise code editing is evaluated by CanItEdit [190] and CodeEditorBench [191], while code efficiency—a crucial non-functional requirement—is measured by benchmarks such as COFFE [192] and EffiBench [193].

Database Systems and DevOps In the realm of database systems, Text-to-SQL translation remains the predominant evaluation task, as it is key to democratizing data access. The Spider [200] benchmark is the established standard for complex, cross-domain queries. Its successors, such as Spider 2.0 [206] and BIRD [202], have increased the difficulty by incorporating more realistic enterprise workflows and value-based queries. The evaluation has also evolved to include conversational context, where models must understand multi-turn dialogues (CoSQL [203], SParC [207]), handle robustness against perturbations (Dr. Spider [201]), and support multilingual queries (DuSQL [208]).

For System Administration and DevOps, benchmarks are emerging to assess the automation of operational tasks. This includes translating natural language to shell commands (NL2Bash [209]), generating Infrastructure-as-Code (IaC) for cloud services (IaC-Eval [204]), and solving broader AIOps problems (OpsEval [205], OWL [210]). In Human-Computer Interaction, FrontendBench [211] specifically evaluates the generation of code for interactive web user interfaces.

4.3.2 Specialized Engineering Disciplines

Beyond software, evaluation frameworks are being developed for hardware and physical engineering domains, which introduce challenges related to physical laws, safety constraints, and highly specialized languages.

In **Electrical and Electronic Engineering**, benchmarks for chip design automation are a primary focus. VerilogEval [212] and RTLLM [213] assess the ability to generate Hardware Description Languages (HDL) like Verilog, a critical skill for designing digital integrated circuits. The evaluation goes beyond mere syntax to include functional correctness through simulation. The focus also extends to the efficiency of the generated hardware, with ResBench [219] measuring FPGA resource utilization, and to the crucial task of design verification, with FIXME [215] providing an end-to-end framework. Other specialized areas include analog circuit design, assessed by CIRCUIT [214], and the niche field of photonic circuits, covered by PICBench [220]. In the domain of power systems, ElecBench [221] evaluates LLM performance on complex power dispatch and fault diagnosis tasks.

In **Mechanical and Manufacturing Engineering**, benchmarks like CADBench [216] assess the generation of scripts for Computer-Aided Design (CAD) software, a key task in automating mechanical design. Materials science is another active area, where benchmarks like LLM4Mat-bench [217] and MSQA [222] test the ability of LLMs to accelerate materials discovery by predicting chemical properties and demonstrating graduate-level reasoning.

For **Aerospace and Transportation Engineering**, a safety-critical domain, specialized benchmarks have been developed to evaluate knowledge and code generation. These include AeroManufacturing-QA [223] for assessing expertise in aerospace manufacturing processes, RepoSpace [224] for evaluating repository-level code generation for satellite systems, and the broad Aviation-Benchmark [218] which covers over ten specific aviation tasks.

4.3.3 Summary and Future Directions

The engineering domain has driven the development of some of the most rigorous, functionally-grounded, and execution-based evaluations for LLMs. The clear trend is a progression from assessing isolated, function-level skills towards evaluating performance on complex, system-level problems that mirror complete engineering workflows. Despite this progress, a significant gap persists between high performance on benchmarks and reliable deployment in real-world, mission-critical engineering applications.

Future research in this area must prioritize several key directions. First, there is a pressing need for benchmarks that can evaluate **holistic engineering workflows**, integrating tasks from requirements

analysis and high-level design through to implementation, verification, and long-term maintenance. Second, establishing robust and standardized evaluation protocols for **safety, reliability, and security** is paramount, especially for domains where failure can have catastrophic consequences. Third, the community must continue to develop **dynamic and contamination-resistant benchmarks** to ensure that evaluations remain a true test of generalization for ever-more-powerful models. Finally, as LLMs become integrated into engineering teams, it is crucial to develop frameworks that assess **human-AI collaboration**, measuring metrics like efficiency gains, error reduction, and trust calibration, rather than evaluating the AI in isolation. Addressing these challenges will be essential for unlocking the full potential of LLMs as transformative tools in engineering practice.

5 Target-specific benchmarks

5.1 Risk & Reliability

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
Safety									
StereoSet[225]	Safety	Monolingual	Manual Design	MCQA	Automated	LM Score SS Score ICAT Score	2.12k	No	1238
CrowS-Pairs[230]	Safety	Bilingual	Manual Design	MCQA	Automated	Bias Score	1,508	No	821
HateCheck[229]	Safety	Multilingual	Manual Design	Classification	Automated	Accuracy	3,728	Yes	313
ToxiGen[226]	Safety	Monolingual	Model Generation	Classification	Automated	AUC	274,186	No	547
Do-Not-Answer[228]	Safety	Bilingual	Manual Design	Classification	Hybrid	Accuracy, Precision, Recall, F1	939	No	133
SG-Bench[233]	Safety	Monolingual	Hybrid	Hybrid	Hybrid	Failure Rate	1442	Yes	10
JailbreakBench[227]	Safety	Monolingual	Hybrid	Generation	Hybrid	Jailbreak Success Rate Defense Effectiveness	200	No	244
AnswerCarefully[234]	Safety	Bilingual	Manual Design	Generation	Automated	Violation Rate Acceptable Response Rate	1800	No	2
SorryBench[235]	Safety	Multilingual	Hybrid	Generation	Automated	fulfillment Rate	9240	No	120
MaliciousInstruct[236]	Safety	Monolingual	Hybrid	Generation	Automated	Attack Success Rate Harmfulness Percentage	100	No	362
HarmBench[231]	Safety	Monolingual	Manual Design	Generation	Automated	Attack Success Rate	500	Yes	434
HEx-PHI[237]	Safety	Monolingual	Manual Design	Generation	Automated	Harmfulness Score Harmfulness Rate	330	No	722
SimpleSafetyTests[238]	Safety	Monolingual	Hybrid	Classification	Automated	Accuracy	3000	No	48
ToxicChat[232]	Safety	Monolingual	Hybrid	Generation	Human	Precision, Recall, F1 jalibreaking recall	10166	No	139
In-The-Wild Jailbreak Prompts[239]	Safety	Monolingual	Hybrid	Generation	Automated	Attack Success Rate	107250	No	721
Hallucination									
TruthfulQA[240]	Hallucination	Multilingual	Manual Design	Generation	Human	Truthfulness Score	817	No	2098
FActScore[241]	Hallucination	Multilingual	Open Datasets	Generation	Hybrid	FActScore	683	Yes	707
RealtimeQA[245]	Hallucination	Monolingual	Web	Generation	Hybrid	Real-time Accuracy	30	Yes	168
FaithBench[246]	Hallucination	Monolingual	Hybrid	Generation	Human	Faithfulness Score	800	Yes	8
DiaHalu[247]	Hallucination	Monolingual	Model Generation	Generation	Human	Hallucination Rate	1103	Yes	15
FactCheck-Bench[248]	Hallucination	Monolingual	Model Generation	Hybrid	Hybrid	Fact-check Accuracy	94	Yes	39
FELM[249]	Hallucination	Monolingual	Hybrid	Generation	Hybrid	Factual Consistency Score	847	Yes	106
FACTOR[250]	Hallucination	Monolingual	Open Datasets	MCQA	Hybrid	Factual Accuracy	4266	Yes	106
FreshQA[244]	Hallucination	Monolingual	Manual Design	Generation	Hybrid	Freshness Score	600	Yes	246
MedHallu[251]	Hallucination	Monolingual	Hybrid	Classification	Human	Medical Accuracy	10000	Yes	5
HaluEval[242]	Hallucination	Monolingual	Hybrid	Generation	Hybrid	Hallucination Rate	35000	No	533
HaluEval2.0[243]	Hallucination	Monolingual	Hybrid	Generation	Hybrid	Hallucination Rate	8770	No	138
FaithDial[53]	Hallucination	Monolingual	Hybrid	Generation	Human	Faithfulness Score	5649	No	100
Robustness									
AdvGLUE[252]	Robustness	Monolingual	Hybrid	Classification	Automated	Robustness Score	5716	No	284
BOSS[255]	Robustness	Monolingual	Open Datasets	Classification	Automated	Robustness Score	900	No	8
IFEval[253]	Robustness	Monolingual	Manual Design	Generation	Automated	Inference Robustness	541	Yes	438
CIF-Bench[256]	Robustness	Monolingual	Hybrid	Generation	Automated	Consistency Score	45000	No	16
PromptRobust[254]	Robustness	Monolingual	Hybrid	Hybrid	Automated	Robustness Score	4788	Yes	52
RoTBench[257]	Robustness	Monolingual	Manual Design	Generation	Automated	Robustness Score	4077	Yes	18
Data Leak									
WikiMIA [258]	Data Leak	Monolingual	Open Datasets	Classification	Automated	Min-k% Prob	250	Yes	368
KoLA[62]	Data Leak	Bilingual	Hybrid	Generation	Automated	Accuracy	500	No	144
C^2LEVA[18]	Data Leak	Bilingual	Hybrid	Hybrid	Automated	Mean Win Rate	16115	Yes	8

Table 7: Comprehensive Summary of Risk & Reliability Benchmarks. The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

The rapid advancement of Large Language Models (LLMs)[4, 10, 316] has unlocked unprecedented potential across a wide spectrum of applications. However, as these models transition from research prototypes to real-world deployment, particularly in high-stakes scenarios such as medical consultation, legal reasoning, financial advising, or customer support, their immense capabilities are shadowed by equally significant risks. Issues like hallucinations, biased outputs, adversarial susceptibility, and privacy violations are no longer theoretical, they have tangible consequences for users, organizations, and society at large.

Consequently, Risk & Reliability assessment has evolved into a central pillar of modern LLM benchmarking frameworks, rather than a peripheral addition. Its core motivations are:

1. Identification and Quantification: To systematically probe LLMs for various negative impact patterns (e.g., generating harmful content[317], hallucinating facts[318], leaking private data[319]) and quantify the frequency and severity of these risks. This necessitates testing under diverse, challenging inputs, including extremes, adversarial prompts, and edge cases (e.g., jailbreak attempts[320], biased prompts[226], fact-intensive queries[240]).

2. Risk Mitigation: To utilize the weaknesses revealed by benchmarks to drive technical improvements (e.g., more robust RLHF, factuality enhancement, privacy-preserving training) by developers, and inform more effective safeguards (e.g., content filters, usage policies) for deployers. The ultimate goal is to minimize the likelihood of models malfunctioning or causing harm.

3. Alignment with Expectations: To verify that model behavior adheres to predefined ethical principles, legal boundaries, and safety requirements (i.e., the alignment problem) during complex, real-world interactions, demonstrating robustness especially on sensitive topics.

4. Building and Sustaining Trust: To provide rigorous, reproducible evidence of risks to demonstrate to users, regulators, and society that a specific LLM is sufficiently reliable, safe, and trustworthy, thereby fostering healthy ecosystem growth and responsible widespread adoption.

In essence, the core question this research direction addresses is: **Beyond impressive capabilities, is this model safe, reliable, and trustworthy enough?** It aims to supply the empirical foundation for the liability guarantee of the model, serving as the essential security checkpoint for LLMs transitioning from research labs to the real world.

5.1.1 Safety

Following pre-training, large language models (LLMs) rely on safety alignment to balance helpfulness and harmlessness. However, ensuring harmlessness often necessitates imposing strict constraints on the model's output space, creating a fundamental conflict with their core strength of deep instruction-following capability, whose objective is to broadly understand and respond to user requests. The inherent flaws in pre-training data, namely the inevitable inclusion of harmful content—embed a latent vulnerability within this conflict: users can employ "jailbreak" techniques to subtly activate these remnants of harmful knowledge retained internally by the model, thereby compromising its safety barriers. Particularly noteworthy is that the explicit reasoning chain approach adopted starting with models like GPT o1, while enhancing interpretability, also unintentionally introduces an observable and steerable pathway for jailbreak attacks. This amplifies safety risks, creating new challenges for the reliable deployment of these models.

Early studies such as HateCheck[229], StereoSet[225], and CrowS-Pairs[230] primarily relied on predefined harmful scenarios and static test cases to evaluate model safety capabilities. However, these approaches suffered from an over-reliance on manually constructed datasets, resulting in limited coverage. Addressing this limitation, ToxiGen[226] leveraged large language models to generate large-scale adversarial and implicit harmful content (up to 274K samples), significantly enhancing the scale and complexity of test sets. This advancement facilitated the development of models with improved generalization capabilities in real-world scenarios. Expanding evaluation dimensions further, Do-Not-Answer[228] addressed the gap in safety assessment for Chinese-language contexts by establishing a standardized testing framework covering eight categories of sensitive topics, including healthcare and criminal activities. To confront emerging challenges in model safety defenses, JailbreakBench[227] systematically integrated over a hundred adversarial prompt techniques (e.g., role-playing and logic-exploiting prompts) to diagnose model vulnerabilities. Building upon this foundation, SG-Bench[233] introduced a cross-task safety generalization evaluation framework designed to test robustness against unseen attack patterns, shifting the evaluation paradigm from static testing toward complex, dynamic interactions. AnswerCarefully[234] further broadened the scope by focusing on Japanese-language contexts, offering 1,800 carefully curated question–answer pairs aligned with cultural norms, thus serving both instruction tuning and safety validation.

More recently, additional benchmarks have expanded the landscape of safety evaluations. HarmBench[231] introduced the first standardized framework for automated red-teaming and robustness refusal evaluation, covering 510 unique harmful behaviors across text and multimodal settings. It also proposed the R2D2 dynamic defense method and conducted the largest comparative study to date, evaluating 18 attack methods across 33 models. HEx-PHI[237] highlighted security risks introduced by fine-tuning, even in non-malicious contexts, and constructed a dataset of 330 red-team prompts grounded in usage policies from Meta and OpenAI to assess vulnerabilities across prohibited use categories. SimpleSafetyTests[238] provided a lightweight suite of 100 English prompts spanning five high-risk domains (e.g., self-harm, scams, child abuse) and demonstrated the effectiveness of lightweight safety filters, with GPT-4-based content moderation achieving the highest accuracy.

Other datasets focus on real-world user interactions. ToxicChat[232] introduced 10,166 toxicity-labeled samples collected from real user–AI conversations with Vicuna, including explicit "jailbreaking recall" metrics to capture hidden adversarial attempts. In-the-wild Jailbreak Prompts[239] systematically studied 1,405 jailbreak prompts gathered from Reddit, Discord, websites, and open datasets, showing their high attack success rates across major LLMs. Similarly, MaliciousInstruct[236] exposed the fragility of alignment in open-source LLMs by demonstrating that simple variations in decoding strategies could raise attack success rates from 0% to over 95%, revealing serious weaknesses in current alignment pipelines. Finally, SORRY-Bench[235] focused on safety refusal evaluation, providing a fine-grained taxonomy of 44 unsafe categories and 440 balanced prompts, along with 20 linguistic augmentations to test cross-linguistic and formatting robustness. Its human-in-the-loop design and efficient LLM-as-a-judge framework allow accurate, large-scale safety refusal benchmarking at lower computational cost.

Taken together, these benchmarks have significantly advanced the study of model safety by broadening the scope beyond purely static and English-only datasets toward more multilingual, adversarial, and fine-grained evaluations. Nonetheless, most current efforts remain grounded in static test suites, and truly dynamic evaluations—those that capture evolving attack strategies and interactive failure modes—are still underexplored. As such, developing systematic dynamic safety benchmarks remains an important direction for future work, and is essential for building safer and more reliable large language models.

5.1.2 Hallucination

Current large language models (LLMs) face the problem of hallucinations, which are primarily categorized into two types: **factual hallucinations**, where the model output contradicts verifiable facts, manifesting as factual inconsistency or fabrication, and **faithfulness hallucinations**, where outputs deviate from user instructions, input context, or lack internal logical consistency.

The causes of hallucinations span the entire model lifecycle. At the data level, misinformation, domain gaps, outdated knowledge, and deficiencies in rare knowledge recall and reasoning contribute significantly. At the training level, limitations such as unidirectional attention's poor contextual capture, the mismatch between autoregressive training and inference, and alignment phase issues like capability or belief misalignment exacerbate the problem. During inference, randomness introduced by decoding strategies (e.g., high-temperature sampling) and architectural constraints (e.g., attention locality and the Softmax bottleneck) further distort factual fidelity.

To address these challenges, primary solutions include data cleaning, retrieval-augmented generation (RAG), and knowledge editing for knowledge enhancement. At the model level, improvements in architecture and alignment data preparation are crucial. During inference, decoding strategies designed to enhance factuality and logical consistency can significantly reduce hallucinations.

To systematically evaluate these hallucinations, a diverse set of benchmarks has been developed. These benchmarks vary in scope, language coverage, task format, and annotation methodology:

TruthfulQA[240] identifies hallucinations where models mimic common human misconceptions across domains like science and history. FActScore[241] evaluates factual grounding in long-form generation by decomposing outputs into atomic facts (e.g., entities and events) and verifying them against external knowledge sources. REALTIMEQA[245] focuses on hallucinations arising from stale knowledge, testing LLMs' adaptability to dynamic, real-time information (e.g., sports, finance).

For faithfulness distortions, FaithBench[246] detects whether summarizations introduce information absent in source texts. DiaHalu[247] targets contextual contradictions in multi-turn dialogues, identifying issues like broken causality or entity inconsistency. FaithDial[53] further expands this by evaluating dialogue systems on their fidelity to input conversational context.

Several benchmarks focus on domain-specific or adversarial hallucinations. MedHallu[251] addresses hallucinations in medical generation tasks, ensuring alignment with trusted clinical knowledge. FreshQA[244] evaluates model freshness by probing with up-to-date world knowledge. FACTOR[250] introduces adversarial conditions (e.g., conflicting prompts) to test models' ability to resist and correct factual errors in real time.

Tools such as FELM[249] extend beyond detection by requiring traceable correction paths and explanation-based justifications. FactCheck-Bench[248] incorporates both model-generated and

manually curated samples to measure fact-checking accuracy. Large-scale datasets like HaluEval[242] and its improved variant HaluEval2.0[243] provide broad coverage across summarization, dialogue, and question answering, enabling benchmarking of hallucination rate at scale.

Despite the diversity of tools, the field still faces three persistent challenges: (1) the lack of a unified evaluation framework leads to fragmented coverage; (2) long-document coherence hallucinations remain difficult to detect; and (3) definitional ambiguity persists when differentiating between subjective judgments and verifiable facts.

5.1.3 Robustness

The rapid advancement of Large Language Models (LLMs) has significantly enhanced the capabilities of natural language processing systems. However, these models often exhibit vulnerabilities when exposed to adversarial inputs, distributional shifts, or subtle prompt variations. Such fragility can lead to erroneous or biased outputs, posing risks in critical applications. Consequently, evaluating and enhancing the robustness of LLMs has become a pivotal research focus.

Robustness in LLMs encompasses several dimensions. Adversarial robustness assesses model resilience against intentionally crafted inputs designed to mislead or deceive. Instruction-following robustness evaluates the consistency and accuracy of models in adhering to varied or complex instructions. Prompt robustness measures sensitivity to minor changes in prompt phrasing or structure. Tool-use robustness determines stability in scenarios requiring external tool integration or multi-step reasoning. These categories guide the development of benchmarks aimed at systematically evaluating LLM robustness across diverse challenges.

The progression of robustness benchmarks reflects a growing understanding of LLM vulnerabilities and the need for comprehensive evaluation tools. AdvGLUE[252], was among the first to systematically evaluate adversarial robustness by applying 14 textual adversarial attack methods to GLUE tasks, revealing significant performance drops in state-of-the-art models and highlighting the necessity for more resilient architectures. BOSS[255] addressed out-of-distribution robustness by evaluating how models trained on specific distributions perform when encountering data from different distributions, emphasizing the importance of generalization in LLMs. IFEval[253] focused on instruction-following capabilities, providing a suite of tasks requiring models to adhere to specific instructions and measuring their ability to follow complex directives accurately. CIF-Bench[256] extended this evaluation to multilingual contexts by assessing instruction-following in Chinese, testing models' zero-shot generalizability and highlighting challenges in cross-lingual understanding. PromptRobust[254] investigated the impact of prompt variations on model outputs, demonstrating that minor changes in prompt wording can significantly affect performance, thus underscoring the need for prompt-invariant models. RoTBench[257] explored robustness in tool-use scenarios by assessing how models perform in environments with varying levels of noise and complexity, thereby evaluating their adaptability in real-world applications. Collectively, these benchmarks have expanded the evaluation landscape, moving beyond traditional accuracy metrics to encompass robustness against a spectrum of perturbations and challenges. Future directions involve developing standardized evaluation protocols, creating benchmarks for additional languages and modalities, and integrating robustness assessments into the model development lifecycle to ensure the deployment of reliable and trustworthy LLMs.

5.1.4 Data Leak

The widespread deployment of Large Language Models (LLMs) has raised significant concerns regarding data leakage, particularly the inadvertent disclosure of sensitive information such as Personally Identifiable Information (PII). This issue stems from the extensive pretraining of LLMs on vast corpora, which may include sensitive data, leading to the models memorizing and potentially reproducing such information during inference. The problem is exacerbated when models are fine-tuned on domain-specific datasets containing confidential information, increasing the risk of privacy breaches. Consequently, evaluating and mitigating data leakage has become a critical area of research.

Recent efforts have introduced multiple benchmarks aimed at systematically measuring data leakage risks in LLMs. For example, WikiMIA[258] focuses on monolingual data leakage by assessing classification performance using the minimum-k% probability as a leakage indicator, operating over open datasets and including PII. KoLA[62] expands this analysis to bilingual contexts and evaluates generative models based on accuracy, relying on a hybrid dataset while excluding explicit

PII. C^2LEVA[18] provides a large-scale benchmark also in a bilingual setting, integrating both classification and generation tasks. It uses mean win rate as the primary metric to measure leakage and includes PII within its evaluation scope.

These benchmarks reflect a growing recognition of the multifaceted nature of privacy risks in LLMs and highlight the necessity of evaluating models beyond traditional performance metrics. Data leakage in these contexts can be categorized by dimensions such as leakage rate, the tendency of a model to expose PII, and the ability of a model to detect and manage sensitive data. These dimensions inform the development of specialized evaluation protocols tailored to different model behaviors and data sources.

Collectively, these benchmarks have expanded the evaluation landscape, moving beyond traditional accuracy-oriented assessments to encompass privacy concerns associated with LLMs. Future directions involve developing standardized evaluation protocols, extending benchmarks to support more languages and modalities, and integrating privacy assessments throughout the model development lifecycle to ensure the deployment of reliable and trustworthy LLMs.

5.1.5 Summary and Future Directions

As large language models (LLMs) transition from research prototypes to real-world deployment in high-stakes domains, their safety and reliability have emerged as core concerns. Modern LLM evaluation frameworks have moved beyond capability-centric metrics to focus on systematically identifying and quantifying risks such as hallucinations, bias, adversarial vulnerabilities, and data leakage. Through fine-grained benchmarks spanning safety, hallucination, robustness, and privacy, researchers have developed multi-task, multilingual, and scenario-rich tools to empirically assess whether a model is reliable, safe, and trustworthy enough for real-world applications. These efforts collectively form the empirical foundation for liability-aware and secure LLM deployment.

Future directions in LLM risk assessment will center on several key areas. First, the development of unified evaluation frameworks is essential to integrate diverse risk dimensions and ensure composability and comparability across benchmarks. Second, the field must expand beyond English-centric evaluations to support low-resource languages and diverse cultural contexts. Third, new challenges such as long-context coherence, multi-turn consistency, and up-to-date knowledge integration call for more sophisticated evaluation protocols. Fourth, the growing complexity of attack strategies highlights the need for interactive adversarial modeling and continuous red-teaming. Fifth, **dynamic real-time evaluation** is becoming increasingly important—LLMs must demonstrate their ability to respond accurately to time-sensitive and evolving information, particularly in domains like finance, healthcare, and current events. Finally, privacy auditing and redaction capabilities must be embedded across the training and inference pipeline, pushing LLM safety evaluations from the model level to the system level.

5.2 Agent

LLM agents are autonomous systems built upon foundation large language models, designed to transcend static prompt-response interactions and engage in goal-driven behaviors. By integrating components such as planning modules, tool-use capabilities, memory systems, and observation loops, these agents can decompose complex objectives into actionable steps, interact dynamically with external environments, and iteratively adapt their strategies until task completion. As LLM agents find increasing application in real-world scenarios, establishing systematic and comprehensive evaluation methodologies becomes essential. As summarized in Table 8, this survey organizes the evaluation framework of LLM agents into four key dimensions: (1) specific capability assessment, focusing on the fine-grained evaluation of individual functions (e.g., planning, reasoning, competition) and execution abilities (e.g., tool use, external control); (2) integrated capability assessment, emphasizing the coordination and synergy of multiple abilities in solving complex tasks; (3) domain proficiency evaluation, focusing on assessing the application of specialized knowledge and the effectiveness in performing tasks within specific professional domains; and (4) safety & risk evaluation, focusing on agents' resilience, susceptibility, and protective mechanisms in adversarial or unsafe scenarios.

Benchmark	Focus	Language	Source	Data Type	Eval.	Indicators	Amount	Method	Citations
Specific Capability Assessment									
FlowBench [259]	Workflow-Guided Planning	Monolingual	Hybrid	Generation	Hybrid	P; R; F1-score; Success Rate, etc.	5313	No	16
Robotouille [262]	Asynchronous Planning	Monolingual	Human	Generation	AE	Success Rate	300	No	5
LLF-Bench [264]	Learning from Language Feedback	Monolingual	Hybrid	Hybrid	AE	Reward; Success Rate	8 sets	No	17
Mobile-Bench [260]	SmartPhone Control	Monolingual	Hybrid	Generation	AE	CheckPoint; PassRate, etc.	832	No	36
Spa-Bench [265]	SmartPhone Control	Bilingual	Human	Generation	Hybrid	Success Rate; Step Ratio, etc.	340	Yes	20
BrowseComp [263]	Web Browse	Monolingual	Hybrid	Generation	AE	Accuracy; Calibration Error	1226	No	19
WebWalkerQA [261]	Web Browse	Bilingual	Hybrid	Generation	AE	Accuracy; Action Count	680	Yes	6
MultiAgentBench [266]	Collaboration & Competition	Monolingual	Hybrid	Generation	LLM	KPI; Planning Score, etc.	600	Yes	13
MAgIC [267]	Competition	Monolingual	Hybrid	Generation	AE	Win Rate; Judgement, etc.	103	Yes	53
ZSC-Eval [268]	Zero-shot Coordination	Monolingual	Model	Classification	Hybrid	BR-Prox; BR-Div, etc.	2 envs.	Yes	14
Integrated Capability Assessment									
SmartPlay [274]	Game	Monolingual	Open Data	Classification	AE	Completion Rate; Reward, et al.	6 games	No	80
BALROG [276]	Game	Monolingual	Open Data	Classification	AE	Average Progress	6 games	No	27
Embodied Agent Interface [277]	Embodied Decision Making	Monolingual	Open Data	Generation	AE	F1; Success Rate, etc.	438	Yes	67
τ-bench [278]	Tool-Agent-User Interaction	Monolingual	Hybrid	Generation	AE	Pass@k	165 tasks	No	58
TravelPlanner [273]	Travel Planning	Monolingual	Hybrid	Generation	AE	Delivery Rate; 3 Pass Rates	1225	No	170
GAIA [269]	General AI Assistants	Monolingual	Human	Generation	AE	Exact Match	466	No	185
AgentQuest [272]	Agent Improvement	Monolingual	Open Data	Generation	AE	Progress Rate; Repetition Rate	4 datasets	No	17
ColBench [279]	Backend Programming & Frontend Design	Monolingual	Hybrid	Generation	AE	Tests Passed; Success Rate, etc.	20K+	Yes	18
AgentBench [270]	Code, Game, Web	Monolingual	Hybrid	Generation	AE	Success Rate; F1; Reward, etc.	1360	No	215
AgentBoard [271]	Web, Tool, Embodied AI, Game	Monolingual	Hybrid	Generation	Hybrid	Progress Rate; Success Rate, etc.	1013	No	17
CharacterEval [275]	Role Playing	Monolingual	Open Data	Generation	LLM	Fluency; Coherency; Consistency, etc.	11376	No	71
Domain Proficiency Evaluation									
TheAgentCompany [287]	Digital Software Worker	Monolingual	Hybrid	Generation	Hybrid	Completion Score; Number of steps, etc.	175	No	39
OSWorld [280]	Computer Operation	Monolingual	Human	Generation	AE	Success Rate	369	No	184
Tapilot-Crossing [286]	Interactive Data Analysis	Monolingual	Hybrid	Hybrid	AE	Accuracy; AccR	1024	Yes	11
ScienceAgentBench [281]	Data-driven Scientific Discovery	Monolingual	Hybrid	Generation	Hybrid	Valid Execution Rate; Success Rate, etc.	102	No	53
SciReplicate-Bench [145]	Algorithmic Reproduction	Monolingual	Open Data	Generation	AE	Execution Accuracy; CodeBLEU, etc.	100	Yes	7
MLGym-Bench [283]	AI Research	Monolingual	Open Data	Generation	AE	AUP Score; Accuracy, etc.	13 tasks	Yes	21
InvestorBench [284]	Financial Decision Making	Monolingual	Open Data	Hybrid	AE	Cumulative Return; Sharpe Ratio, etc.	3 tasks	No	13
BixBench [139]	Biological Data Analysis	Monolingual	Hybrid	Hybrid	Hybrid	Accuracy	296	No	6
AgentClinic [282]	Clinical Decision Making	Monolingual	Open Data	Generation	LLM	Diagnostic Accuracy; Confidence, etc.	1544	No	69
CourtBench [285]	Legal Reasoning	Monolingual	Hybrid	MCQA	AE	Accuracy	124	Yes	11
Safety & Risk Evaluation									
ASB [291]	Attacks & Defenses	Monolingual	Hybrid	Generation	AE	Attack Success Rate; Refuse Rate, etc.	50 tasks	Yes	49
AgentHarm [288]	Jailbreak Attacks	Monolingual	Hybrid	Generation	Hybrid	Harm Score; Refusal Rate, etc.	440	No	67
SafeAgentBench [289]	Safe Task Planning	Monolingual	Model	Generation	Hybrid	Rejection Rate; Success Rate, etc.	750	No	6
R-Judge [290]	Safety Risk Awareness	Monolingual	Hybrid	Hybrid	Hybrid	Safety Judgment; Risk Identification	569	No	101

Table 8: Summary of representative benchmarks for LLM agent. Evaluation methods are abbreviated as AE (Automated Evaluation), ME (Mixed Evaluation). The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

5.2.1 Specific Capability Assessment

Assessing the specific capabilities of LLM-based agents is essential for understanding their functional reliability and task generalization limits. Instead of evaluating end-to-end task success alone, this line of research focuses on isolating and benchmarking core abilities such as planning, reasoning, tool use, and interactive behavior. Such evaluations enable more interpretable diagnostics and drive progress on targeted weaknesses in agent design.

A series of benchmarks center on planning and reasoning abilities. FlowBench [259] evaluates how LLM agents leverage workflow knowledge to perform structured, domain-specific planning. Robotouille [262] targets asynchronous planning, requiring agents to handle delayed effects and overlapping actions in long-horizon tasks. LLF-Bench [264] focuses on an agent's ability to improve

through iterative language feedback, testing whether agents can learn across turns using naturalistic supervision. WebWalkerQA [261] examines reasoning over hierarchical web structures, evaluating how agents navigate multi-layered pages to extract complex information.

Another major direction is evaluating external control and tool-use capabilities. SPA-Bench [265] and Mobile-Bench [260] assess LLM agents' performance on mobile device control, with tasks spanning single-app actions to multi-app collaboration. These benchmarks emphasize interface understanding, execution reliability, and reasoning under UI constraints. BrowseComp [263] extend this to web environments, measuring agents' effectiveness in retrieving hard-to-find facts through persistent, tool-mediated browsing.

Multi-agent scenarios test coordination, competition, and social reasoning. MultiAgentBench [266] examines agent collaboration across different topologies and tasks, while MAgIC [267] incorporates social deduction games and game-theoretic settings to probe deception, self-awareness, and judgment. ZSC-Eval [268] introduces the challenge of zero-shot coordination, where agents must generalize to novel partners in cooperative environments, with evaluation grounded in behavior diversity and robustness.

These benchmarks emphasize that evaluating isolated cognitive or interaction abilities is foundational for building trustworthy LLM agents. By pinpointing weaknesses in planning, reasoning, tool use, and multi-agent dynamics, specific capability assessment provides fine-grained insights that inform more robust and modular agent design.

5.2.2 Integrated Capability Assessment

Integrated capability assessment focuses on evaluating how well LLM agents coordinate multiple abilities—such as reasoning, planning, tool use, memory, and interaction—to solve complex, multi-step tasks. Unlike evaluations targeting isolated skills, this assessment emphasizes holistic competence in dynamic environments, where agents must sequence decisions, adapt to feedback, follow constraints, and operate across multiple modalities. It reflects the practical readiness of LLM agents to handle real-world challenges that demand cognitive integration and situational flexibility.

Several benchmarks assess agents' integrated reasoning and decision-making in game-like or embodied environments. SmartPlay [274] uses diverse games like Tower of Hanoi and Minecraft to decompose and assess nine essential agent abilities, including object dependency reasoning and learning from history. BALROG [276] evaluates both LLMs and VLMs in complex games requiring planning, spatial reasoning, and exploration across various difficulty levels. Embodied Agent Interface [277] provides a unified framework for embodied tasks and evaluates agents across subtasks like goal interpretation, action sequencing, and transition modeling.

Some benchmarks simulate real-world task environments requiring tool usage, constraint tracking, and user interaction. τ-bench [278] evaluates agents in multi-turn tool-augmented dialogues with domain-specific rules, measuring goal satisfaction and behavior consistency. TravelPlanner [273] tests complex real-world planning through a large-scale travel sandbox, assessing tool use, information integration, and constraint handling. GAIA [269] offers questions requiring general web search, multimodal perception, and robust reasoning, targeting human-level generalist performance. AgentQuest [272] introduces a modular benchmarking framework with extensible metrics to diagnose agent progress and failure modes over multi-step tasks.

Multi-turn interaction and collaborative reasoning are also critical to integrated capability assessment. SWEET-RL [279] focuses on multi-turn collaboration tasks like frontend design, using a step-wise reward system to optimize agent behavior. ColBench [279], built alongside it, facilitates realistic back-and-forth problem solving with human collaborators. AgentBench [270] and AgentBoard [271] provide large-scale evaluation environments across multiple domains, incorporating multi-turn planning, decision-making, and error tracking to reveal bottlenecks in agent performance. CharacterEval [275] adds a role-playing dimension to agent evaluation in Chinese, testing coherence, personality consistency, and long-term conversation management.

Together, these benchmarks emphasize the necessity of evaluating LLM agents' coordination and synergy across multiple core abilities to solve complex, multi-step problems. By focusing on holistic competence in dynamic and interactive environments, integrated capability assessments provide deeper insights into agents' real-world readiness, exposing challenges in adaptability, multi-modal reasoning, and sustained interaction.

5.2.3 Domain Proficiency Evaluation

As language agents transition from general-purpose tools to specialized assistants, evaluating their domain proficiency becomes essential. Unlike general reasoning tasks, these evaluations focus on the agent's ability to apply expert knowledge, follow domain-specific procedures, and complete professional tasks with precision. Such tasks often involve high-stakes decision-making, intricate workflows, multimodal inputs, or specialized tools, requiring agents to go beyond generic language understanding.

To this end, a wide range of domain-specific benchmarks have been proposed. In workplace and productivity scenarios, TheAgentCompany [287] assesses agents on realistic office tasks such as browsing, coding, and intra-team communication, while OSWorld [280] provides a scalable, interactive operating environment that evaluates agents' ability to perform open-ended computer tasks across platforms. In data science and scientific research, Tapilot-Crossing [286] benchmarks interactive data analysis capabilities, ScienceAgentBench [281] and SciReplicate-Bench [145] evaluate agents' ability to generate and reproduce scientific code, and MLGym-Bench [283] focuses on end-to-end AI research tasks from hypothesis generation to model evaluation.

The financial domain is covered by InvestorBench [284], which evaluates agents across diverse financial instruments and market scenarios, while the biomedical and healthcare domains are represented by BixBench [139] and AgentClinic [282], which test agents on long-horizon bioinformatics tasks and clinical decision-making under multimodal constraints, respectively. In legal contexts, CourtBench [285] serves as the evaluation suite within AgentCourt, measuring legal reasoning, cognitive agility, and argumentative rigor of adversarially evolved lawyer agents in simulated courtroom trials. For software engineering and algorithm comprehension, SciReplicate-Bench [145] further challenges agents to extract, understand, and implement methods from recent NLP papers.

Collectively, these benchmarks underscore the gap between general language proficiency and domain-specific expertise. They highlight that while current agents demonstrate partial competence, substantial limitations remain in reliability, depth of understanding, and tool integration across professional settings. As such, domain proficiency evaluation plays a critical role in driving the development of trustworthy, capable, and specialized LLM agents.

5.2.4 Safety & Risk Evaluation

Safety and risk evaluation addresses the robustness of LLM agents under adversarial, malicious, or failure-prone conditions. As agents move beyond static text generation to tool use, memory manipulation, and autonomous decision-making, the surface for potential misuse and failure grows significantly. This evaluation dimension focuses on whether agents can maintain aligned, reliable behavior while resisting manipulation, detecting hazards, and handling unsafe instructions. It plays a crucial role in ensuring trustworthy deployment, especially in sensitive or high-stakes domains.

One category of benchmarks targets adversarial vulnerabilities and attack-resistance. Agent Security Bench (ASB) [291] provides a comprehensive framework covering a wide range of real-world scenarios, agents, and attack types. It reveals significant weaknesses across agent components, including prompt processing, memory, and tool interfaces. AgentHarm [288] focuses on agent misuse by introducing harmful task prompts across domains such as fraud and cybercrime. It finds that many leading agents comply with malicious requests, and simple jailbreak templates can induce harmful multi-step behaviors.

A second line of work explores safety in embodied and task-execution environments. SafeAgentBench [289] evaluates whether agents can recognize and avoid hazardous instructions in interactive simulations. Results show that current agents generally lack the ability to reject unsafe plans, even when task understanding is adequate. R-Judge [290] shifts the focus to risk judgment, assessing whether agents can identify safety issues from interaction records. While some frontier models show moderate awareness, overall performance remains far from robust, highlighting the difficulty of safety reasoning in open-ended settings.

Collectively, these benchmarks reveal that LLM agents remain highly vulnerable to both adversarial manipulation and operational hazards. Improving their safety requires deeper integration of risk modeling, hazard detection, and behavioral safeguards, moving beyond surface-level refusals toward genuine situational awareness and robustness.

5.2.5 Summary and Future Directions

The evaluation of LLM agents has evolved into a multi-dimensional endeavor encompassing specific capabilities, integrated competencies, domain proficiency, and safety robustness. Recent benchmarks have illuminated the growing sophistication of agents across planning, reasoning, tool use, and interaction, while also highlighting key limitations in cross-skill coordination, domain-specific expertise, and adversarial resilience. Specific capability assessments provide fine-grained diagnostics of isolated functions; integrated evaluations reflect holistic problem-solving in dynamic environments; domain benchmarks underscore the gap between general language ability and professional-grade performance; and safety evaluations expose vulnerabilities in real-world scenarios. Overall, this emerging ecosystem of benchmarks offers a systematic lens through which to assess the functional maturity and deployment readiness of LLM agents.

Looking ahead, future research may prioritize three interlinked directions. First, there is a pressing need for evaluation compositionality, namely developing unified frameworks that assess how well agents orchestrate diverse skills across evolving contexts, rather than excelling in isolated tests. Second, grounded and continuous evaluation must be strengthened, where agent behavior is tested in realistic, long-term deployments involving dynamic feedback loops, imperfect tools, and human collaboration. Third, robustness and safety must move beyond binary refusal detection toward deeper modeling of intent, risk context, and adaptive safeguards. Achieving trustworthy autonomy will require benchmarks that are not only broader in coverage but also richer in interpretability and diagnostic precision, enabling the next generation of LLM agents to act reliably in complex, open-ended, and high-stakes environments.

5.3 Others

Benchmark	Focus	Language	Source	Data type	Eval.	Indicators	Amount	Method	Citations
PET-BENCH [292]	Virtual Pet Companionship	Monolingual	Hybrid	Hybrid	AE	BLEU, ROUGE, Accuracy	7,815	No	0
TP-RAG [293]	Travel Planning	Monolingual	Hybrid	Generation	AE	FR, RR, DMR, STR	2,348	No	1
FLUB [294]	Fallacy Understanding	Monolingual	Hybrid	Hybrid	AE	Accuracy, F-1 Score GPT-4 Score	834	No	15
CDEval [295]	Cultural Dimensions	Multilingual	Hybrid	MCQA	AE	Average Likelihood	2,953	No	22
NORMAD-ETI [296]	Cultural Adaptability	Monolingual	Hybrid	Classification	AE	Accuracy	2.6k	No	6
EmotionQueen [298]	Emotional Intelligence	Monolingual	Model	Generation	AE	Pass Rate Win Rate	10000	No	69
OR-Bench [299]	Over-refusal	Monolingual	Hybrid	Generation	AE	Over-refusal Rate, Acceptance Rate	80,000	No	79
SocialStigmaQA [300]	Stigma Amplification	Monolingual	Hybrid	MCQA	AE	Biased Answer Percent	10k	No	22
Shopping MMLU [301]	Online Shopping	Multilingual	Web	Hybrid	AE	Hit Rate@3, NDCG, Micro F1 ROUGE-L, BLEU	14683	No	17
JudgeBench [297]	LLM-based Judges	Monolingual	Hybrid	MCQA	AE	Accuracy	350	No	71
LLM-Evolve [302]	Evolving Capability	Monolingual	Opendata Extend	MCQA, Generation	AE	Success Rate, F1, Reward ,Game Progress, Step Success Rate, Accuracy	23k	No	9
SUC [305]	Structural Understanding Capabilities	Monolingual	Opendata	Generation	AE	Accuracy	255k	No	206
DOCBENCH [303]	Document reading	Monolingual	Hybrid	MCQA	AE	Accuracy	1102	No	13
ROUTERBENCH [306]	Routers Hinders Progress	Monolingual	Hybrid	Generation	AE	Exact Match, GPT-4, Accuracy, Total Cost (Dollars)	405,467	Yes	81
GAME COMPETITIONS [307]	Game Simulation	Monolingual	Hybrid	Generation	AE	Win Rate, Draw Rates, Disqualification Rate	2310	No	16
RTLLM 2.0 [308]	Design RTL generation	Monolingual	Human	Generation	AE	Pass@k	50	No	16
AD-LLM [309]	Anomaly Detection	Monolingual	Opendata	Classification	AE	AUROC, AUPRC	190k	Yes	16
ZIQI-Eval [310]	Musical Abilities	Monolingual	Human	MCQA	AE	Precision, Recall, Accuracy, F1	14,000	No	15
VisEval [304]	Tabular Data Vsualization	Monolingual	Hybrid	Generation	AE	Validity, Legality, Readability Rating	2,524	No	44

Table 9: Summary of other representative others benchmarks. The 'Method' column indicates if the paper proposed a new methodology (Yes/No).

In recent years, an increasing body of research has shifted toward evaluating large language models (LLMs) from unique, human-centered perspectives, moving beyond conventional NLP tasks such as question answering, summarization, and translation. These evaluations extend into more complex domains including cultural adaptability, emotional intelligence, value alignment, real-world task execution, and multimodal technical capabilities.

In the realm of cultural intelligence and social adaptability, CDEval [295] conducts multilingual multiple-choice testing across six cultural dimensions and seven domains with 2,953 samples, while NORMAD-ETI [296] uses 2.6k classification tasks from 75 cultural contexts to assess narrative adaptability. SocialStigmaQA [300] examines bias amplification on 93 social stigma topics through 10k multiple-choice samples, revealing that models may still exhibit systematic tendencies toward certain groups even under explicit prompting.

Emotional understanding and interpersonal interaction form another core dimension. EmotionQueen [298] evaluates the recognition of implicit emotions and appropriateness of affective responses in 10k generation tasks, using PASS and WIN rates as metrics. Similarly, PET-Bench [292] tests emotional support and memory-consistent dialogue through 7,815 virtual pet companionship scenarios, combining BLEU, ROUGE, and accuracy. Complementary to these are studies on value alignment, reasoning robustness, and safety balance. OR-Bench [299] analyzes 80k over-refusal

cases to measure the balance between refusals and acceptances; FLUB [294] tests logical fallacy recognition over 834 samples with accuracy, F1, and GPT-4 scores; JudgeBench [297] assesses the consistency of LLM-based judges across 350 multiple-choice questions.

Simulated real-world tasks provide more application-oriented evaluation settings. Shopping MMLU [301] measures reasoning, behavior modeling, and recommendation performance in e-commerce scenarios using 14,683 multilingual hybrid tasks, with metrics including hit rate@3, NDCG, micro-F1, ROUGE-L, and BLEU. TP-RAG [293] evaluates 2,348 travel planning tasks on itinerary design, destination matching, and scheduling accuracy. DOCBENCH [303] contains 1,102 document reading comprehension tasks; LLM-Evolve [302] tracks capability evolution over 23k extended tasks; SUC [305] features 255k structured data understanding tasks assessed by accuracy; and VisEval [304] evaluates 2,524 table visualization outputs for validity, legality, and readability.

Several benchmarks explore technical and multimodal extensions. ROUTERBENCH [306] evaluates routing strategies over 405,467 tasks with exact match, GPT-4 scores, accuracy, and economic cost. GAME COMPETITIONS [307] measure strategic reasoning through win, draw, and disqualification rates in 2,310 game simulations. RTLLM 2.0 [308] targets RTL code generation for hardware design with pass@k evaluation over 50 hand-crafted tasks. AD-LLM [309] tests anomaly detection on 190k classification tasks using AUROC and AUPRC. ZIQI-Eval [310] assesses musical understanding and generation on 14k multiple-choice tasks, measuring precision, recall, accuracy, and F1.

Overall, these diverse benchmarks reflect a clear trend: LLM evaluation is transitioning from purely linguistic competence assessment toward multidimensional, human-centered, socially-aware, and application-driven evaluation. By encompassing cultural adaptability, emotional perception, value alignment, real-world task complexity, and technical extension, these works establish a methodological foundation for developing next-generation language models that are not only linguistically capable but also socially responsible and practically valuable.

6 Conclusion

This survey maps the evolving landscape of LLM evaluation benchmarks, drawing on a detailed analysis of 283 benchmarks organized into three core categories: general capabilities, domain-specific expertise, and target-specific functionalities. Tracing their progression from task-specific leaderboards to multidimensional frameworks, we highlight how these benchmarks have both reflected and driven advances in LLM capabilities, from foundational linguistics to domain mastery, and from standalone performance to safety-critical reliability. Our taxonomy reveals key tensions: general benchmarks often sacrifice depth for breadth; domain-specific assessments risk over-specialization; and target-specific metrics struggle to balance technical rigor with real-world relevance. These challenges intensify as LLMs operate in dynamic, multi-agent, high-stakes environments, where static datasets and single-turn metrics fail to capture emergent behaviors or societal impacts. As LLMs integrate into sociotechnical systems, evaluation must shift from measuring "what models can do" to "how they should perform responsibly." Future benchmarks require dynamism (to match model evolution), causality (to explain outcomes), inclusion (to avoid bias), and robustness (to anticipate risks). Achieving this requires cross-disciplinary collaboration to align technical rigor with societal values.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[6] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[9] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[10] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[11] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[12] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

[14] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[15] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

[16] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[17] Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24948–24956, 2025.

[18] Yanyang Li, Tin Long Wong, Cheung To Hung, Jianqiao Zhao, Duo Zheng, Ka Wai Liu, Michael R. Lyu, and Liwei Wang. C^2leva: Toward comprehensive and contamination-free language model evaluation, 2025.

[19] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[20] C. E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64, 1951.

[21] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1998.

[22] Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002, 2002.

[23] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[24] Stefan Kombrink, Tomás Mikolov, Martin Karafiát, and Lukás Burget. Recurrent neural network based language modeling in meeting recognition. In *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, Florence, Italy, August 27-31, 2011*, pages 2877–2880. ISCA, 2011.

[25] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 932–938. MIT Press, 2000.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[33] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[34] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[36] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.

[37] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[38] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[39] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[40] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[41] Michael Boratko, Xiang Lorraine Li, Rajarshi Das, Tim O'Gorman, Dan Le, and Andrew McCallum. Protoqa: A question answering dataset for prototypical common-sense reasoning, 2020.

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[43] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277, 2021.

[44] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7239–7252, Online, July 2020. Association for Computational Linguistics.

[45] Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *arXiv preprint arXiv:2406.05761*, 2024.

[46] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[47] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[50] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024.

[51] Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. Dynaeval: Unifying turn and dialogue level evaluation. *arXiv preprint arXiv:2106.01112*, 2021.

[52] Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. Mdia: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint arXiv:2208.13078*, 2022.

[53] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. Faithdial: A faithful benchmark for information-seeking dialogue, 2022.

[54] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR, 2020.

[55] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[56] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. *URL https://arxiv. org/abs/2304.06364*.

[57] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark, 2023j. *Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Zhaohui Wy, Dawei He, Peng Cheng, Zhonghao Wang, and Haiying Deng. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation*, 2023.

[58] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023.

[59] Yi Zong and Xipeng Qiu. Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. *arXiv preprint arXiv:2402.15745*, 2024.

[60] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023.

[61] Zhang Ge, D Xinrun, C Bei, L Yiming, L Tongxu, Z Tianyu, Z Kang, C Yuyang, X Chunpu, G Shuyue, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.20847*, 2024.

[62] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023.

[63] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language, 2020.

[64] Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language, 2021.

[65] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[66] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020.

[67] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[68] Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[69] Anjiang Wei, Yuheng Wu, Yingjia Wan, Tarun Suresh, Huanmi Tan, Zhanke Zhou, Sanmi Koyejo, Ke Wang, and Alex Aiken. Satbench: Benchmarking llms' logical reasoning via automated puzzle generation from sat formulas, 2025.

[70] Jin Jiang, Yuchen Yan, Yang Liu, Yonggang Jin, Shuai Peng, Mengdi Zhang, Xunliang Cai, Yixin Cao, Liangcai Gao, and Zhi Tang. Logicpro: Improving complex logical reasoning via program-guided learning, 2025.

[71] Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[72] Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2853, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[73] Simeng Han, Aaron Yu, Rui Shen, Zhenting Qi, Martin Riddell, Wenfei Zhou, Yujie Qiao, Yilun Zhao, Semih Yavuz, Ye Liu, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Dragomir Radev, Rex Ying, and Arman Cohan. P-FOLIO: Evaluating and improving logical reasoning with abundant human-written reasoning chains. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16553–16565, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[74] Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20856–20879, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[75] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning, 2025.

[76] Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, and Xiangliang Zhang. Dissecting logical reasoning in llms: A fine-grained evaluation and supervision study, 2025.

[77] Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen tse Huang, Pinjia He, Wenxiang Jiao, and Michael R. Lyu. Logicasker: Evaluating and improving the logical reasoning ability of large language models, 2024.

[78] Santiago Ontanon, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. Logicinference: A new dataset for teaching logical inference to seq2seq models, 2022.

[79] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data, 2022.

[80] Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models, 2025.

[81] Leyan Pan, Vijay Ganesh, Jacob Abernethy, Chris Esposo, and Wenke Lee. Can transformers reason logically? a study in sat solving, 2025.

[82] Utkarsh Tiwari, Aryan Seth, Adi Mukherjee, Kaavya Mer, Kavish, and Dhruv Kumar. Debatebench: A challenging long context reasoning benchmark for large language models, 2025.

[83] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, 2023.

[84] Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples, 2023.

[85] Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability, 2024.

[86] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, 2021.

[87] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning, 2020.

[88] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.

[89] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics.

[90] Hailin Chen, Fangkai Jiao, Mathieu Ravaut, Nawshad Farruque, Xuan Phi Nguyen, Chengwei Qin, Manan Dey, Bosheng Ding, Caiming Xiong, Shafiq Joty, and Yingbo Zhou. Structtest: Benchmarking llms' reasoning through compositional structured outputs, 2025.

[91] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024.

[92] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024.

[93] Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore, December 2023. Association for Computational Linguistics.

[94] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms, 2019.

[95] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024.

[96] Tianqiao Liu, Zui Chen, Zhensheng Fang, Weiqi Luo, Mi Tian, and Zitao Liu. Matheval: A comprehensive benchmark for evaluating large language models on mathematical reasoning capabilities. *Frontiers of Digital Education*, 2(2):16, May 2025.

[97] Eldar Kurtic, Amir Moeini, and Dan Alistarh. Mathador-LM: A dynamic benchmark for mathematical reasoning on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17020–17027, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[98] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[99] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[100] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark, 2025.

[101] Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. Sysbench: Can large language models follow system messages?, 2024.

[102] Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. From passive to active reasoning: Can large language models ask the right questions under incomplete information?, 2025.

[103] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[104] Satyam Goyal and Soham Dan. Iolbench: Benchmarking llms on linguistic reasoning, 2025.

[105] Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. Analogical – a novel benchmark for long text analogy evaluation in large language models, 2023.

[106] Frederikus Hudi, Genta Indra Winata, Ruochen Zhang, and Alham Fikri Aji. Textgames: Learning to self-play text-based puzzle games via language model reasoning, 2025.

[107] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[108] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[109] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

[110] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

[111] Daman Arora, Himanshu Gaurav Singh, et al. Have llms advanced enough? a challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074*, 2023.

[112] Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*, 2024.

[113] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

[114] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, et al. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025.

[115] Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024.

[116] Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*, 2024.

[117] Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022.

[118] Albert S Yue, Lovish Madaan, Ted Moskovitz, DJ Strouse, and Aaditya K Singh. Harp: A challenging human-annotated math reasoning benchmark. *arXiv preprint arXiv:2412.08819*, 2024.

[119] Michael Shalyt, Rotem Elimelech, and Ido Kaminer. Asymob: Algebraic symbolic mathematical operations benchmark. *arXiv preprint arXiv:2505.23851*, 2025.

[120] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[121] Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025.

[122] Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, et al. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*, 2025.

[123] Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, et al. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*, 2025.

[124] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

[125] Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, et al. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*, 2025.

[126] Wei Li, Xin Zhang, Zhongxin Guo, Shaoguang Mao, Wen Luo, Guangyue Peng, Yangyu Huang, Houfeng Wang, and Scarlett Li. Fea-bench: A benchmark for evaluating repository-level code generation for feature implementation. *arXiv preprint arXiv:2503.06680*, 2025.

[127] Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*, 2025.

[128] Daniel JH Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

[129] Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, Yi Li, et al. Chemeval: A comprehensive multi-level chemical evaluation for large language models. *arXiv preprint arXiv:2409.13989*, 2024.

[130] Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. *arXiv preprint arXiv:2403.08192*, 2024.

[131] Haochen Zhao, Xiangru Tang, Ziran Yang, Xiao Han, Xuanzhi Feng, Yueqing Fan, Senhao Cheng, Di Jin, Yilun Zhao, Arman Cohan, et al. Chemsafetybench: Benchmarking llm safety on chemistry domain. *arXiv preprint arXiv:2411.16736*, 2024.

[132] Zhuoyu Wei, Wei Ji, Xiubo Geng, Yining Chen, Baihua Chen, Tao Qin, and Daxin Jiang. Chemistryqa: A complex question answering dataset from chemistry. 2020.

[133] Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. Scholarchemqa: Unveiling the power of language models in chemical research question answering. *arXiv preprint arXiv:2407.16931*, 2024.

[134] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

[135] Guangyong Chen, Pengfei Chen, Chang-Yu Hsieh, Chee-Kong Lee, Benben Liao, Renjie Liao, Weiwen Liu, Jiezhong Qiu, Qiming Sun, Jie Tang, et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427*, 2019.

[136] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[137] Haiteng Zhao, Chang Ma, FangZhi Xu, Lingpeng Kong, and Zhi-Hong Deng. Biomaze: Benchmarking and enhancing large language models for biological pathway reasoning. *arXiv preprint arXiv:2502.16660*, 2025.

[138] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, et al. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976*, 2024.

[139] Ludovico Mitchener, Jon M. Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew D. White, Lorenzo Sani, and Samuel G. Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *CoRR*, abs/2503.00096, 2025.

[140] Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

[141] Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.

[142] Yuyang Liu, Liuzhenghao Lv, Xiancheng Zhang, Li Yuan, and Yonghong Tian. Bioprobench: Comprehensive dataset and benchmark in biological protocol understanding and reasoning. *arXiv preprint arXiv:2505.07889*, 2025.

[143] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.

[144] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*, 2024.

[145] Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers. *CoRR*, abs/2504.00255, 2025.

[146] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061, 2024.

[147] Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

[148] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[149] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[150] Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. *arXiv preprint arXiv:2503.13517*, 2025.

[151] Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

[152] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

[153] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

[154] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, 2024.

[155] Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551, 2022.

[156] Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. Lexeval: a comprehensive chinese legal benchmark for evaluating large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc.

[157] Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. Laiw: A chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, 2025.

[158] Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. Citalaw: Enhancing llm with citations in legal domain. *arXiv preprint arXiv:2412.14556*, 2024.

[159] Haitao Li, Jiaying Ye, Yiran Hu, Jia Chen, Qingyao Ai, Yueyue Wu, Junjie Chen, Yifan Chen, Cheng Luo, Quan Zhou, et al. Casegen: A benchmark for multi-stage legal case documents generation. *arXiv preprint arXiv:2502.17943*, 2025.

[160] You Zuo, Kim Gerdes, Eric Villemonte de La Clergerie, and Benoît Sagot. Patenteval: Understanding errors in patent generation. In *NAACL2024-2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

[161] Shiwen Ni, Minghuan Tan, Yuelin Bai, Fuqiang Niu, Min Yang, Bowen Zhang, Ruifeng Xu, Xiaojun Chen, Chengming Li, and Xiping Hu. Mozip: A multilingual benchmark to evaluate large language models in intellectual property. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11658–11668, 2024.

[162] Qiyao Wang, Guhong Chen, Hongbo Wang, Huaren Liu, Minghui Zhu, Zhifei Qin, Linwei Li, Yilin Yue, Shiqiang Wang, Jiayan Li, et al. Ipbench: Benchmarking the knowledge of large language models in intellectual property. *arXiv preprint arXiv:2504.15524*, 2025.

[163] Qiyao Wang, Jianguo Huang, Shule Lu, Yuan Lin, Kan Xu, Liang Yang, and Hongfei Lin. Ipeval: A bilingual intellectual property agency consultation evaluation benchmark for large language models. *arXiv preprint arXiv:2406.12386*, 2024.

[164] Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, et al. Autopatent: A multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*, 2024.

[165] Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, et al. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7753–7774, 2024.

[166] Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*, 2025.

[167] Jiahao Zhao, Jingwei Zhu, Minghuan Tan, Min Yang, Renhao Li, Yang Di, Chenhao Zhang, Guancheng Ye, Chengming Li, Xiping Hu, et al. Cpsyexam: A chinese benchmark for evaluating psychology using examinations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11248–11260, 2025.

[168] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13947–13966, 2024.

[169] Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 2024.

[170] Shuyu Liu, Ruoxi Wang, Ling Zhang, Xuequan Zhu, Rui Yang, Xinzhu Zhou, Fei Wu, Zhi Yang, Cheng Jin, and Gang Wang. Psychbench: A comprehensive and professional benchmark for evaluating the performance of llm-assisted psychiatric clinical practice. *arXiv preprint arXiv:2503.01903*, 2025.

[171] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*, 2024.

[172] Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*, 2023.

[173] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 33469–33484, 2023.

[174] Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*, 2023.

[175] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[176] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[177] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021.

[178] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *ArXiv*, abs/2403.07974, 2024.

[179] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiao ke Hong, Wen-Ding Li, Jean Kaddour, Minglian Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiao-Nan Du, Harm de Vries, and Leandro von Werra. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *ArXiv*, abs/2406.15877, 2024.

[180] Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Yih, Daniel Fried, Si yi Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, 2022.

[181] Xiangru Tang, Bill Qian, Rick Gao, Jiakang Chen, Xinyun Chen, and Mark B. Gerstein. Biocoder: A benchmark for bioinformatics code generation with contextual pragmatic knowledge. *ArXiv*, abs/2308.16458, 2023.

[182] Quan Shi, Michael Tang, Karthik Narasimhan, and Shunyu Yao. Can language models solve olympiad programming? *ArXiv*, abs/2404.10952, 2024.

[183] Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, et al. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? *arXiv preprint arXiv:2506.11928*, 2025.

[184] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *ArXiv*, abs/2308.01861, 2023.

[185] Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, and Jing Ma. Mmcode: Benchmarking multimodal large language models for code generation with visually rich programming problems. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

[186] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[187] André Silva and Martin Monperrus. Repairbench: Leaderboard of frontier models for program repair. In *2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code)*, pages 9–16. IEEE, 2025.

[188] Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Yinxu Pan, Yesai Wu, Hui Haotian, Liu Weichuan, Zhiyuan Liu, et al. Debugbench: Evaluating debugging capability of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4173–4198, 2024.

[189] Yonghao Wu, Zheng Li, Jie M. Zhang, and Yong Liu. Condefects: A complementary dataset to address the data leakage concern for llm-based fault localization and program repair. *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024.

[190] Federico Cassano, Luisa Li, Akul Sethi, Noah Shinn, Abby Brennan-Jones, Jacob Ginesin, Edward Berman, George Chakhnashvili, Anton Lozhkov, Carolyn Jane Anderson, et al. Can it edit? evaluating the ability of large language models to follow code editing instructions. In *First Conference on Language Modeling*, 2024.

[191] Jiawei Guo, Ziming Li, Xueling Liu, Kaijing Ma, Tianyu Zheng, Zhouliang Yu, Ding Pan, Yizhi Li, Ruibo Liu, Yue Wang, Shuyue Guo, Xingwei Qu, Xiang Yue, Ge Zhang, Wenhu Chen, and Jie Fu. Codeeditorbench: Evaluating code editing capability of large language models. *ArXiv*, abs/2404.03543, 2024.

[192] Yun Peng, Jun Wan, Yichen Li, and Xiaoxue Ren. Coffe: A code efficiency benchmark for code generation. *ArXiv*, abs/2502.02827, 2025.

[193] Dong Huang, Jie M. Zhang, Yuhao Qing, and Heming Cui. Effibench: Benchmarking the efficiency of automatically generated code. *ArXiv*, abs/2402.02037, 2024.

[194] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021.

[195] Mohammad Abdullah Matin Khan, M Saiful Bari, Do Xuan Long, Weishi Wang, Md. Rizwan Parvez, and Shafiq R. Joty. xcodeeval: A large scale multilingual multitask benchmark for code understanding, generation, translation and retrieval. *ArXiv*, abs/2303.03004, 2023.

[196] Chenxiao Liu and Xiaojun Wan. Codeqa: A question answering dataset for source code comprehension. *ArXiv*, abs/2109.08365, 2021.

[197] Junjie Huang, Duyu Tang, Linjun Shou, Ming Gong, Ke Xu, Daxin Jiang, Ming Zhou, and Nan Duan. Cosqa: 20,000+ web queries for code search and question answering. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[198] Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. *ArXiv*, abs/2306.03091, 2023.

[199] Hong Yi Lin, Chunhua Liu, Haoyu Gao, Patanamon Thongtanunam, and Christoph Treude. Codereviewqa: The code review comprehension assessment for large language models. *ArXiv*, abs/2503.16167, 2025.

[200] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.

[201] Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei Lan, Sheng Zhang, Jiarong Jiang, Joseph Lilien, Steve Ash, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, and Bing Xiang. Dr. Spider: A diagnostic evaluation benchmark towards text-to-SQL robustness. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

[202] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357, 2023.

[203] Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[204] Patrick T Kon, Jiachen Liu, Yiming Qiu, Weijun Fan, Ting He, Lei Lin, Haoran Zhang, Owen M Park, George S Elengikal, Yuxin Kang, et al. Iac-eval: A code generation benchmark for cloud infrastructure-as-code programs. *Advances in Neural Information Processing Systems*, 37:134488–134506, 2024.

[205] Yuhe Liu, Changhua Pei, Longlong Xu, Bohan Chen, Mingze Sun, Zhirui Zhang, Yongqian Sun, Shenglin Zhang, Kun Wang, Haiming Zhang, et al. Opseval: A comprehensive task-oriented aiops benchmark for large language models. *arXiv preprint arXiv:2310.07637*, 2023.

[206] Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin SU, ZHAOQING SUO, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, et al. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *The Thirteenth International Conference on Learning Representations*, 2025.

[207] Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, 2019.

[208] Lijie Wang, Ao Zhang, Kun Wu, Ke Sun, Zhenghua Li, Hua Wu, Min Zhang, and Haifeng Wang. DuSQL: A large-scale and pragmatic Chinese text-to-SQL dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6923–6935, 2020.

[209] Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D Ernst. Nl2bash: A corpus and semantic parser for natural language interface to the linux operating system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[210] Hongcheng Guo, Jian Yang, Jiaheng Liu, Liqun Yang, Linzheng Chai, Jiaqi Bai, Junran Peng, Xiaorong Hu, Chao Chen, Dongfeng Zhang, et al. Owl: A large language model for it operations. In *ICLR*, 2024.

[211] Hongda Zhu, Jingzhe Ding, Dandan Wang, Yiwen Zhang, Siyao Liu, Yanan Liu, Bing Zhao, Tong Liu, and Zhaojian Li. Frontendbench: A benchmark for evaluating llms on front-end development via automatic evaluation, 2025.

[212] Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. Verilogeval: Evaluating large language models for verilog code generation. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–8. IEEE, 2023.

[213] Yao Lu, Shang Liu, Qijun Zhang, and Zhiyao Xie. Rtllm: An open-source benchmark for design rtl generation with large language model. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 722–727. IEEE, 2024.

[214] Lejla Skelic, Yan Xu, Matthew Cox, Wenjie Lu, Tao Yu, and Ruonan Han. CIRCUIT: A benchmark for circuit interpretation and reasoning capabilities of LLMs. *arXiv preprint arXiv:2502.07980*, 2025.

[215] Gwok-Waa Wan, Shengchu Su, Ruihu Wang, Qixiang Chen, Sam-Zaak Wong, Mengnv Xing, Hefei Feng, Yubo Wang, Yinan Zhu, Jingyi Zhang, Jianmin Ye, Xinlai Wan, Tao Ni, Qiang Xu, Nan Guan, Zhe Jiang, Xi Wang, and Yang Jun. Fixme: Towards end-to-end benchmarking of llm-aided design verification. 2025.

[216] Yuhao Du, Shunian Chen, Wenbo Zan, Peizhao Li, Mingxuan Wang, Dingjie Song, Bo Li, Yan Hu, and Benyou Wang. Blenderllm: Training large language models for computer-aided design with self-improvement. *ArXiv*, abs/2412.14203, 2024.

[217] Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6, 2024.

[218] The MITRE Corporation. Aviationgpt: A large language model for the aviation domain, 2023. Approved for Public Release, Distribution Unlimited. PRS Case 23-4011.

[219] Ce Guo and Tong Zhao. Resbench: A resource-aware benchmark for llm-generated fpga designs. *Proceedings of the 15th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies*, 2025.

[220] Yuchao Wu, Xiaofei Yu, Hao Chen, Yang Luo, Yeyu Tong, and Yuzhe Ma. Picbench: Benchmarking llms for photonic integrated circuits design. In *2025 Design, Automation & Test in Europe Conference (DATE)*, pages 1–6, 2025.

[221] Xiang Zhou et al. ElecBench: A power dispatch evaluation benchmark for large language models. *arXiv preprint arXiv:2407.05365*, 2024.

[222] Jerry Junyang Cheung, Shiyao Shen, Yuchen Zhuang, Yinghao Li, Rampi Ramprasad, and Chao Zhang. Msqa: Benchmarking llms on graduate-level materials science reasoning and knowledge. *ArXiv*, abs/2505.23982, 2025.

[223] Beiming Liu, Zhizhuo Cui, Siteng Hu, Xiaohua Li, Haifeng Lin, and Zhengxin Zhang. Llm evaluation based on aerospace manufacturing expertise: Automated generation and multi-model question answering, 2025.

[224] Rui He, Liang Zhang, Mengyao Lyu, Liangqing Lyu, and Changbin Xue. Using large language models for aerospace code generation: Methods, benchmarks, and potential values. *Aerospace*, 12(6):498, 2025.

[225] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[226] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

[227] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

[228] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023.

[229] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*, 2020.

[230] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[231] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. 2024.

[232] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation, 2023.

[233] Yutao Mou, Shikun Zhang, and Wei Ye. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *Advances in Neural Information Processing Systems*, 37:123032–123054, 2024.

[234] Hisami Suzuki, Satoru Katsumata, Takashi Kodama, Tetsuro Takahashi, Kouta Nakayama, and Satoshi Sekine. Answercarefully: A dataset for improving the safety of japanese llm output, 2025.

[235] Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*, 2025.

[236] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

[237] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

[238] Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

[239] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.

[240] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[241] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023.

[242] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

[243] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024.

[244] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023.

[245] Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[246] Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, et al. Faithbench: A diverse hallucination benchmark for summarization by modern llms. *arXiv preprint arXiv:2410.13210*, 2024.

[247] Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models, 2024.

[248] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers, 2024.

[249] Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523, 2023.

[250] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models, 2024.

[251] Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models, 2025.

[252] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

[253] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[254] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts, 2024.

[255] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

[256] Yizhi LI, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Zekun Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Stephen W. Huang, Chenghua Lin, and Jie Fu. Cif-bench: A chinese instruction-following benchmark for evaluating the generalizability of large language models, 2024.

[257] Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. Rotbench: A multi-level benchmark for evaluating the robustness of large language models in tool learning, 2024.

[258] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2023.

[259] Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. Flowbench: Revisiting and benchmarking workflow-guided planning for llm-based agents. In *EMNLP (Findings)*, pages 10883–10900. Association for Computational Linguistics, 2024.

[260] Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, Rui Yan, and Shuo Shang. Mobile-bench: An evaluation benchmark for llm-based mobile agents. In *ACL (1)*, pages 8813–8831. Association for Computational Linguistics, 2024.

[261] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal. *CoRR*, abs/2501.07572, 2025.

[262] Gonzalo Gonzalez-Pumariega, Leong Su Yean, Neha Sunkara, and Sanjiban Choudhury. Robotouille: An asynchronous planning benchmark for LLM agents. In *ICLR*. OpenReview.net, 2025.

[263] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516, 2025.

[264] Ching-An Cheng, Andrey Kolobov, Dipendra Misra, Allen Nie, and Adith Swaminathan. Llf-bench: Benchmark for interactive learning from language feedback. *CoRR*, abs/2312.06853, 2023.

[265] Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, Kaiwen Zhou, Rui Shao, Liqiang Nie, Yasheng Wang, Jianye Hao, Jun Wang, and Kun Shao. Spa-bench: a comprehensive benchmark for smartphone agent evaluation. In *ICLR*. OpenReview.net, 2025.

[266] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. Multiagentbench: Evaluating the collaboration and competition of LLM agents. *CoRR*, abs/2503.01935, 2025.

[267] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *EMNLP*, pages 7315–7332. Association for Computational Linguistics, 2024.

[268] Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. Zsc-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *NeurIPS*, 2024.

[269] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *ICLR*. OpenReview.net, 2024.

[270] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In *ICLR*. OpenReview.net, 2024.

[271] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn LLM agents. In *NeurIPS*, 2024.

[272] Luca Gioacchini, Giuseppe Siracusano, Davide Sanvito, Kiril Gashteovski, David Friede, Roberto Bifulco, and Carolin Lawrence. Agentquest: A modular benchmark framework to measure progress and improve LLM agents. In *NAACL (Demonstrations)*, pages 185–193. Association for Computational Linguistics, 2024.

[273] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. In *ICML*. OpenReview.net, 2024.

[274] Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. Smartplay : A benchmark for llms as intelligent agents. In *ICLR*. OpenReview.net, 2024.

[275] Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *ACL (1)*, pages 11836–11850. Association for Computational Linguistics, 2024.

[276] Davide Paglieri, Bartlomiej Cupial, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Lukasz Kucinski, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: benchmarking agentic LLM and VLM reasoning on games. In *ICLR*. OpenReview.net, 2025.

[277] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS*, 2024.

[278] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *CoRR*, abs/2406.12045, 2024.

[279] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. SWEET-RL: training multi-turn LLM agents on collaborative reasoning tasks. *CoRR*, abs/2503.15478, 2025.

[280] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS*, 2024.

[281] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *ICLR*. OpenReview.net, 2025.

[282] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *CoRR*, abs/2405.07960, 2024.

[283] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob N. Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. Mlgym: A new framework and benchmark for advancing AI research agents. *CoRR*, abs/2502.14499, 2025.

[284] Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, Jimin Huang, Lingfei Qian, Xueqing Peng, Qianqian Xie, and Jordan W. Suchow. INVESTORBENCH: A benchmark for financial decision-making tasks with llm-based agent. *CoRR*, abs/2412.18174, 2024.

[285] Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *CoRR*, abs/2408.08089, 2024.

[286] Jinyang Li, Nan Huo, Yan Gao, Jiayi Shi, Yingxiu Zhao, Ge Qu, Yurong Wu, Chenhao Ma, Jian-Guang Lou, and Reynold Cheng. Tapilot-crossing: Benchmarking and evolving llms towards interactive data analysis agents. *CoRR*, abs/2403.05307, 2024.

[287] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. *CoRR*, abs/2412.14161, 2024.

[288] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of LLM agents. In *ICLR*. OpenReview.net, 2025.

[289] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied LLM agents. *CoRR*, abs/2412.13178, 2024.

[290] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-judge: Benchmarking safety risk awareness for LLM agents. In *EMNLP (Findings)*, pages 1467–1490. Association for Computational Linguistics, 2024.

[291] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): formalizing and benchmarking attacks and defenses in llm-based agents. In *ICLR*. OpenReview.net, 2025.

[292] Hongcheng Guo, Zheyong Xie, Shaosheng Cao, Boyang Wang, Weiting Liu, Zheyu Ye, Zhoujun Li, and Zuozhu Liu. Act-as-pet: Benchmarking the abilities of large language models as e-pets in social network services. *CoRR*, abs/2506.03761, 2025.

[293] Hang Ni, Fan Liu, Xinyu Ma, Lixin Su, Shuaiqiang Wang, Dawei Yin, Hui Xiong, and Hao Liu. TP-RAG: benchmarking retrieval-augmented large language model agents for spatiotemporal-aware travel planning. *CoRR*, abs/2504.08694, 2025.

[294] Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. When llms meet cunning texts: A fallacy understanding benchmark for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[295] Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou, editors, *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[296] Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[297] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[298] Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. EmotionQueen: A benchmark for evaluating empathy of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[299] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *CoRR*, abs/2405.20947, 2024.

[300] Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21454–21462. AAAI Press, 2024.

[301] Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, Haodong Wang, Zhengyang Wang, Wenju Xu, Jingfeng Yang, Qingyu Yin, Xian Li, Priyanka Nigam, Yi Xu, Kai Chen, Qiang Yang, Meng Jiang, and Bing Yin. Shopping MMLU: A massive multi-task online shopping benchmark for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[302] Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. LLM-evolve: Evaluation for LLM's evolving capability on benchmarks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[303] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. DOCBENCH: A benchmark for evaluating llm-based document reading systems. *CoRR*, abs/2407.10701, 2024.

[304] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Trans. Vis. Comput. Graph.*, 31(1):1301–1311, 2025.

[305] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets LLM: can large language models understand structured table data? A benchmark and empirical study. In Luz Angelica Caudillo-Mata, Silvio Lattanzi, Andrés Muñoz Medina, Leman Akoglu, Aristides Gionis, and Sergei Vassilvitskii, editors, *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, pages 645–654. ACM, 2024.

[306] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *CoRR*, abs/2403.12031, 2024.

[307] Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. Evaluating large language models with grid-based game competitions: An extensible LLM benchmark and leaderboard. *CoRR*, abs/2407.07796, 2024.

[308] Shang Liu, Yao Lu, Wenji Fang, Mengming Li, and Zhiyao Xie. Openllm-rtl: Open dataset and benchmark for llm-aided design RTL generation. In Jinjun Xiong and Robert Wille, editors, *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, ICCAD 2024, Newark Liberty International Airport Marriott, NJ, USA, October 27-31, 2024*, pages 60:1–60:9. ACM, 2024.

[309] Tiankai Yang, Yi Nian, Li Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A. Rossi, Kaize Ding, Xia Hu, and Yue Zhao. AD-LLM: Benchmarking large language models for anomaly detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1524–1547, Vienna, Austria, July 2025. Association for Computational Linguistics.

[310] Jiajia Li, Lu Yang, Mingni Tang, Chenchong Chenchong, Zuchao Li, Ping Wang, and Hai Zhao. The music maestro or the musically challenged, A massive music evaluation benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3246–3257. Association for Computational Linguistics, 2024.

[311] Alejandro F Villaverde, David Henriques, Kieran Smallbone, Sophia Bongard, Joachim Schmid, Damjan Cicin-Sain, Anton Crombach, Julio Saez-Rodriguez, Klaus Mauch, Eva Balsa-Canto, et al. Biopredyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. *BMC systems biology*, 9(1):8, 2015.

[312] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*, 2023.

[313] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[314] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[315] Norman L Webb. Depth-of-knowledge levels for four content areas. *Language Arts*, 28(March):1–9, 2002.

[316] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie

Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025.

[317] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[318] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[319] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models, 2023.

[320] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024.