Article

Not peer-reviewed version

Rethinking Benchmark Comparability: A Survey of Reasoning Benchmarks for Large Language Models

Chenyuan Zhang , Simin Liu , Hanjing Li , Te Gao , Yidi Wang , Qiguang Chen , Xiachong Feng , Li Cai , Mengnan Du , Zhuotao Tian , Libo Qin * , Philip S. Yu , Min Zhang

Posted Date: 13 May 2026

doi: 10.20944/preprints202605.0806.v1

Keywords: large language models; reasoning; reasoning benchmarks; reasoning evaluation; natural language processing

 doi:10.20944/preprints202605.0806.v1

Article

Rethinking Benchmark Comparability: A Survey of Reasoning Benchmarks for Large Language Models

Chenyuan Zhang [1,2], Simin Liu [1], Hanjing Li [1], Te Gao [3], Yidi Wang [1], Qiguang Chen [3], Xiachong Feng [4], Li Cai [5], Mengnan Du [6], Zhuotao Tian [1], Libo Qin [1,3,5,*], Philip S. Yu [7] and Min Zhang [1]

[1] Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China
[2] Shanghai Innovation Institute, Shanghai 201210, China
[3] Central South University, Changsha 410083, China
[4] The University of Hong Kong, Hong Kong, China
[5] Text Computing and Cognitive Intelligence Ministry of Protect Education Engineering Research Center, Guizhou University, Guiyang 550025, China
[6] The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China
[7] University of Illinois Chicago, Chicago 60607, United States
* Correspondence: qinlibo@hit.edu.cn

Abstract

As reasoning becomes a defining capability of large language models, reasoning benchmarks have moved to the center of evaluation. However, despite the rapid growth in the number of benchmarks and reported scores, benchmark results are often not directly comparable. This is because benchmarks may differ not only in the reasoning capabilities they target, but also in the conditions under which models are evaluated and the criteria used to assess success. To address this challenge, we present the first survey of reasoning benchmarks for large language models across three dimensions: Object, Setting, and Evaluation. Object defines the reasoning capability under examination. Setting specifies the conditions that shape model behavior. Evaluation determines how success is measured. We further introduce extended scenarios to account for special conditions. Based on this analysis, we identify two major weaknesses in current practice, namely heterogeneous benchmark objects and weakly justified settings, and derive practical guidance for benchmark selection, construction, and reporting, along with future directions for benchmark development. We hope this survey will help advance reasoning evaluation beyond score comparison alone toward benchmarks that are more interpretable, better justified, and easier to implement. A repository for the related papers is available at https://github.com/chenyuanTKCY/Awesome-Benchmarks-for-LLM-Reasoning.

Keywords: large language models; reasoning; reasoning benchmarks; reasoning evaluation; natural language processing

1. Introduction

Reasoning has become one of the central axes of progress in large language model research [1–3]. As models are increasingly required to move beyond short response generation and support complex problem solving, the significance of reasoning has become progressively more pronounced [1,2]. Early reasoning in language models was largely confined to relatively short and task-specific intermediate computations, often lacking the depth required for complex inference [4]. The emergence of chain-of-thought prompting marked a major shift by showing that explicitly generated intermediate reasoning steps can unlock substantially stronger performance on complex tasks [5]. More recent research has pushed this paradigm toward longer and more exploratory reasoning trajectories, more deliberate System 2-style thinking, and even latent reasoning mechanisms beyond fully verbalized thought chains [6–11]. Meanwhile, reasoning abilities are expanding toward multi-step, multimodal, and cross-domain settings, enabling models to address increasingly diverse and complex problems [1,12–16].

 doi:10.20944/preprints202605.0806.v1

Under such a trend, the quality of benchmarks largely determines whether reasoning evaluation can be conducted in an objective and reliable manner [6,17–19]. The need has grown rapidly in recent years, accompanied by a corresponding expansion in the range and diversity of relevant works [20–25]. Existing reasoning benchmarks now span symbolic reasoning, mathematical reasoning, knowledge intensive reasoning, and agentic decision making [26–31]. Although these benchmarks are often grouped together under *the label of reasoning*, they frequently **differ substantially in the capabilities they target, the information by design, the protocol formulation they assume, and the logic by which outputs are evaluated** [32–34]. Moreover, scores that appear to belong to the same category are often treated as comparable, though they often reflect materially different task semantics [35,36]. Therefore, it is no longer sufficient to ask whether a model can produce a correct answer. Rather, we must establish a scientific foundation for rethinking this incomparability [20,37,38].

Motivated by these challenges, we define the comparability of reasoning benchmarks not merely as an organizational inconvenience, but as a **scientific problem in how reasoning benchmarks are produced, interpreted, and utilized**. As shown in Figure 1, **we present the first comprehensive survey of reasoning benchmarks for large language models** from the perspective of benchmark semantics, benchmark construction, and benchmark evaluation, reframing the comparability of reasoning benchmarks. Rather than treating benchmarks as a flat collection of datasets, we organize the space **through three closely connected dimensions, Object, Setting, and Evaluation**. **Object (What): Reasoning Capabilities of Reasoning Benchmarks** identifies the target reasoning capability that a benchmark is intended to probe, including logical reasoning, mathematical reasoning, knowledge reasoning and agentic reasoning. **Setting (How): Construction of Reasoning Benchmarks** specifies the execution conditions under which that capability is elicited, including data provenance and the protocol formulation.

Evaluation (How Well): Assessment of Reasoning Benchmarks specifies the outcomes to be evaluated and how they are measured, including evaluation units and evaluation dimensions. To supply and expand the methodology, we further instantiate it through an exploration of extended scenarios, including multilingual and localized, multimodal, vertical-domain, as well as agentic and interactive benchmarks. This methodology makes benchmark explicit, separating capability evidence from protocol effects, and provides a practical guideline for organizing existing benchmarks as well as designing new ones. Finally, we identify two central weaknesses in current benchmark practice, namely **excessive heterogeneity of benchmarks** and **insufficiently justified benchmark settings**, and discuss future directions of reasoning benchmarks.

Our contributions are threefold.

- To the best of our knowledge, we provide **the first comprehensive survey of reasoning benchmarks for large language models**, framing a scientific foundation for how reasoning benchmarks are constructed, interpreted, and ultimately utilized in practice.
- We introduce **a unified and operational perspective** that connects benchmark Object, Setting, and Evaluation, and further instantiate it through an exploration of extended scenarios to support our methodology, and accordingly offer **practical guidelines** for benchmark selection and construction.
- We identify **two central weaknesses** in current practice as well as analyze the potential impact of these weaknesses on the scientific validity. Additionally, we outline **future directions** for reasoning benchmarks, covering measurement mechanics, promising directions and essential governance considerations.

 doi:10.20944/preprints202605.0806.v1

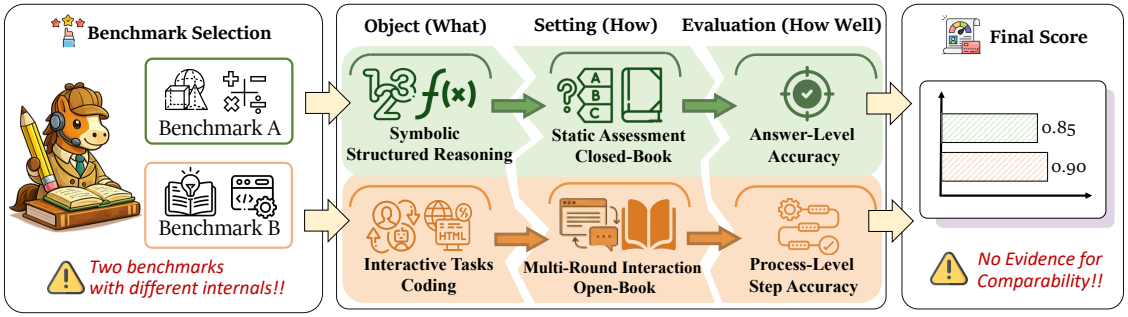

Figure 1. The Incomparability of Scores Arising from the Fragmented Landscape of Reasoning Benchmarks: Even under controlled conditions of backbone and task, reported scores exhibit incomparability, attributable to **internal structural incoherence** across benchmark constructions.

2. Background

2.1. The Reasoning in LLMs

In this survey, reasoning is viewed as the capability **to reach a correct final decision by maintaining logically dependent intermediate states under uncertainty and constraints** [6,37,39,40]. This aspect distinguishes reasoning from surface fluency: as failures in reasoning often stem not from linguistic incoherence, but from step-level inconsistencies and violations of cross-step constraints [3,26]. A model can be linguistically coherent yet fail to satisfy cross-step constraints, and such failure is precisely what reasoning benchmarks are expected to expose [18,39,41].

Thus, reasoning can appear in **multiple forms**, including symbolic derivation, mathematical problem solving, knowledge-intensive evidence, and sequential tool-mediated decision making [42–47]. Although these forms differ significantly in domain, interface modality, and task formulation, they share a **common reasoning structure**: success depends on whether the model can derive an answer or an action through a sequence of interdependent intermediate states maintained over time [39,41,47–53].

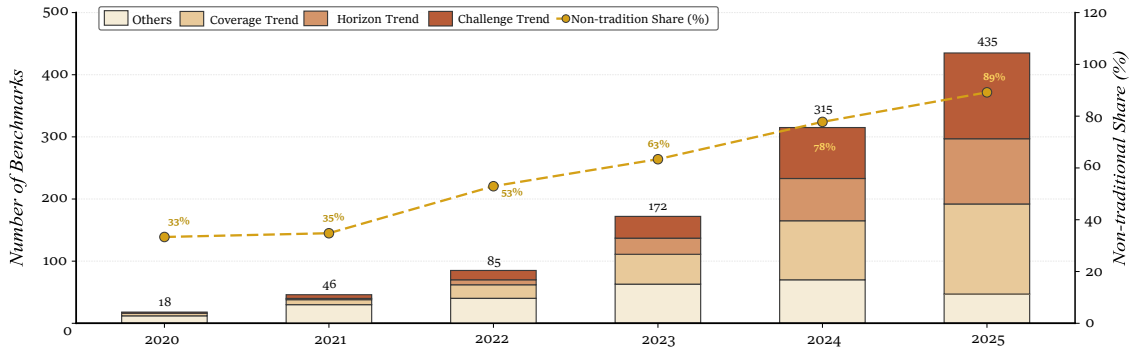

Figure 2. Trends in global LLM reasoning benchmarks from 2020 to 2025. The stacked bars show the annual benchmark counts by category, while the dashed line indicates the share of non-traditional benchmarks. The figure shows rapid growth after 2023 and a clear shift toward broader, longer-horizon, and more challenging reasoning benchmark designs.

2.2. Basic Elements of Benchmark Description

Following our definition of reasoning, a benchmark can be understood as a **structured evaluation instrument that turns a reasoning claim into an assessable form** [54–57]. More specifically, a benchmark specifies what capability is intended to be measured, under what conditions that capability is elicited, and how performance is judged. In this sense, benchmarks do not merely provide scores. They also determine what counts as evidence of reasoning and therefore shape how progress in reasoning is interpreted [51,58,59].

This role is particularly important in the study of LLM reasoning. Unlike tasks with relatively direct input–output mappings, reasoning performance is often sensitive to benchmark design choices such as task formulation, data provenance, inference-time protocol, and scoring criteria [60,61]. As

 doi:10.20944/preprints202605.0806.v1

a result, a reported score may reflect not only a model's underlying reasoning ability, but also the assumptions embedded in the benchmark itself [62–64]. For this reason, benchmark description is a necessary part of reasoning research, since benchmark results can only be interpreted properly when the capability, the setting, and the criteria are explicit [59,65,66].

2.3. The trend of reasoning benchmarks

In recent years, the development of reasoning benchmarks has been closely shaped by the rapid improvement of large language models reasoning abilities [67–69]. As shown in Figure 2, as these models have grown stronger, the landscape of reasoning benchmarks has expanded not only in scale, but also in difficulty, coverage, and task composition [70–72]. Early reasoning benchmarks were typically built around short-context settings and relatively simple question-answering tasks [73–75]. These benchmarks were easy to scale and compare across models, but they were limited in the range of reasoning behaviors they could capture[73,74]. As performance on such tasks improved, benchmark design gradually moved beyond narrow, static QA toward more challenging suites that emphasize longer reasoning chains, complex problem structures, and diverse task settings [34,67–69,76]. Thus, the evolution of reasoning benchmarks has been marked not only by rising difficulty and quantity, but also by a broader reconsideration of the scenarios they are expected to cover [72,77,78].

2.4. The growing incomparability of reasoning benchmarks

However, this shift toward more difficult, diverse, and application-oriented benchmarks has also introduced a serious comparability challenge [47,71,72,78–80]. As reasoning benchmarks expand across domains, formats, and evaluation settings, they become increasingly heterogeneous in different objects [47,68,69,77].

As a result, benchmark scores **are often no longer derived from comparable task assumptions**: they may reflect different input lengths, interaction protocols, tool access, environmental constraints, and success criteria rather than performance on a shared reasoning scale [47,72,80].

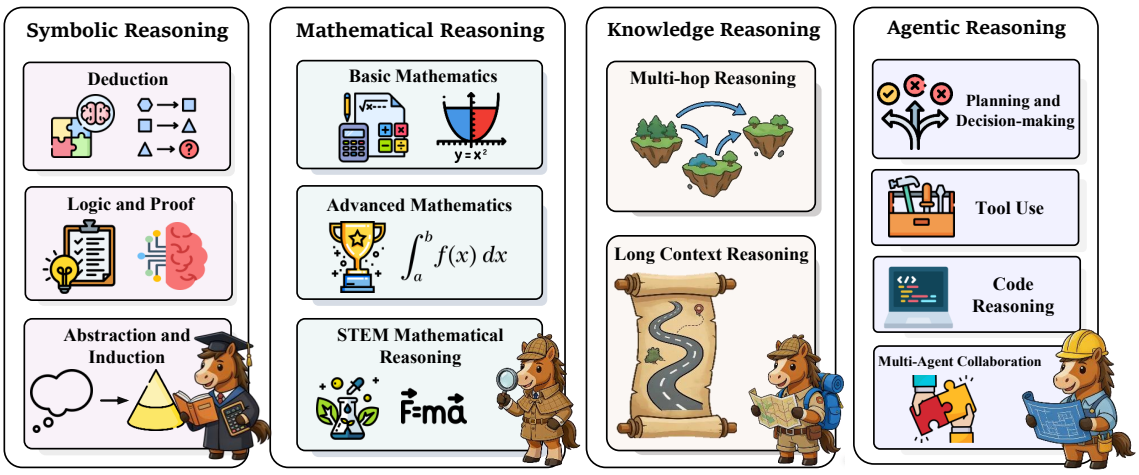

Figure 3. A High-Level Taxonomy of Benchmark Objects. Evaluation tasks are organized into four primary domains: symbolic, mathematical, knowledge, and agentic reasoning, with constituent subtasks detailed within each domain.

3. Object (What): Reasoning Capabilities of Reasoning Benchmarks

As shown in Figure 3, this section clarifies the core capabilities that reasoning benchmarks are intended to evaluate. Given the **multifaceted nature** of reasoning, it is essential to establish a structured taxonomy for these evaluation objects. According to Figure 4, we organize target capabilities into four dimensions: **Symbolic Reasoning**, **Mathematical Reasoning**, **Knowledge Reasoning**, and **Agentic Reasoning**. For each dimension, we define the scope and list the sub-capabilities that benchmarks commonly instantiate.

 doi:10.20944/preprints202605.0806.v1

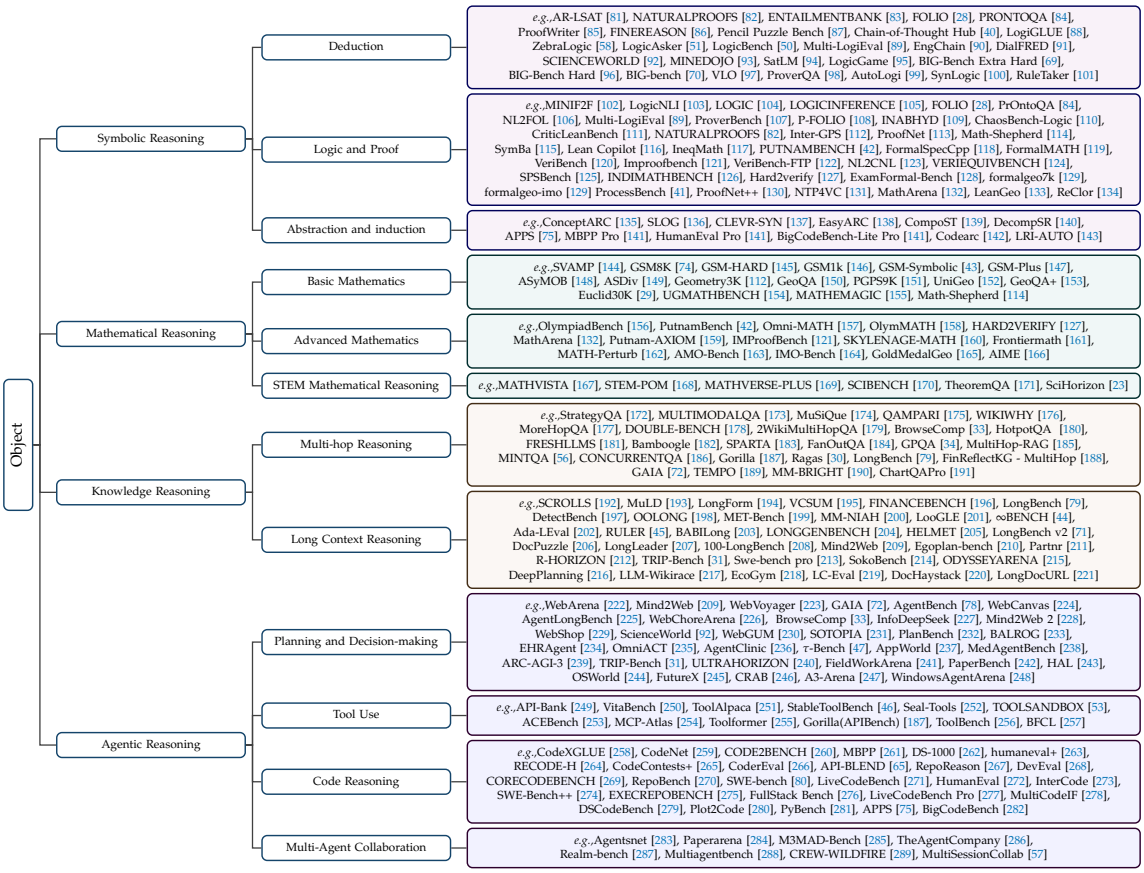

Figure 4. Taxonomy of reasoning benchmarks, covering symbolic, mathematical, knowledge, and agentic reasoning with their subcategories, providing a structured overview of the evaluation landscape and annotated with representative benchmarks.

3.1. Symbolic Reasoning

Symbolic reasoning can be divided into three benchmark sub-objects: **Deduction**, **Logic and Proof** and **Abstraction and Induction**. **Deduction** focuses on whether a model can derive valid consequences from explicit premises under a fixed rule system, while **Logic and Proof** emphasizes the construction, validation, and search of proof structures in semi-formal or formal settings. **Abstraction and Induction** further assesses the ability to infer general rules from examples and transfer them to new cases.

3.1.1. Deduction

Deduction is defined as controlling inference from explicit premises under a fixed rule system [84,85]. What makes it a distinct object is that benchmark success depends primarily on whether a model can compute consequence relations over stated facts and rules, rather than on background world knowledge, stylistic generation, or broad reading comprehension [28,85]. In this object, the benchmark oracle is usually exact, either as entailment labels, proof depth conditioned accuracy, or verifiable proof traces [84,85]. This object is instantiated in benchmarks that present a small theory, often in natural language but with tightly constrained semantics, and then ask whether a conclusion follows, contradicts the theory, or remains unknown [28,85].

Some benchmarks sharpen the same target by pairing synthetic worlds with proof chains that can be checked against an underlying formal semantics, which makes benchmark errors interpretable as failures of stepwise consequence computation rather than failures of retrieval or domain knowledge [84]. Other benchmarks extend the object toward open domain language while preserving first order logical grounding, so that the benchmark still centers on valid deduction instead of mere plausibility [28,113,290].

 doi:10.20944/preprints202605.0806.v1

3.1.2. Logic and Proof

Logic and proof is a benchmark object whose target capability is maintaining logical validity across structured arguments and proof obligations [102,113,290]. Unlike deduction benchmarks, which mainly test closure under explicit premises, this object is defined by benchmarks in which the central challenge is to recognize valid argument structure, assemble proof steps, or search for proofs in semi formal or formal systems [102,113,290].

One benchmark family employs standardized logical reasoning tasks that derive their complexity from constraint satisfaction, quantified statements, and logical consistency. ReClor and AR-LSAT package logic into exam style multiple choice formats while still targeting proof relevant structure rather than open ended knowledge recall. A second family makes the proof artifact explicit. EntailmentBank asks for multi-step entailment trees grounded in scientific facts, so the benchmark object is instantiated as proof graph construction rather than answer selection alone. A third family moves to formal mathematics and theorem proving [102,113,290]. MiniF2F, ProofNet, and LeanDojo benchmark proof search, premise selection, formalization, or theorem completion in systems where correctness is checked by proof assistants [102,113,130,131,133,290].

When a benchmark is dominated by formal proof obligations, machine checked derivations, or explicit proof artifacts, it fits logic and proof even if the underlying content is mathematical [102,113,290]. By contrast, natural language contest problems whose main difficulty is derivational mathematical problem solving belong more naturally under advanced mathematics [75,158].

3.1.3. Abstraction and Induction

Abstraction and induction is a benchmark object whose target capability is inferring latent rules or concepts from sparse evidence and then applying them to novel instances. It is distinct from symbolic deduction because the rules are not fully given, and it is distinct from generic perception because benchmark success depends on discovering the governing transformation, relation, or concept rather than recognizing familiar surface patterns. The object is defined by benchmarks that minimize opportunities for memorization and instead force the model to construct a task specific abstraction [135].

Canonical examples are ARC-style benchmarks, where models must generalize from a handful of demonstrations to a held-out case, forcing construction of task-specific abstractions instead of memorization [135,142,291]. The same benchmark object also appears beyond visual formats, including program synthesis and code-generation settings, where success depends on inducing reusable solution structure under novelty [75,141]. More broadly, benchmarks in this family are unified not by modality but by the requirement to discover the governing transformation, concept, or schema before applying it [135,143].

Key Incomparability: Symbolic Reasoning

- ***Why the distinction matters***: distinct facets of symbolic reasoning competence.
- *Deduction*: benchmarked as consequence computation from explicit premises.
- *Logic and proof*: benchmarked as validity-preserving proof construction or verification.
- *Abstraction-and-induction*: benchmarked as inferring latent rules from limited examples and systematically generalizing them to unseen cases.

3.2. Mathematical Reasoning

Mathematical reasoning can be divided into three benchmark objects: **Basic Mathematics**, **Advanced Mathematics**, and **STEM Mathematical Reasoning**. Basic mathematics centers on elementary quantitative problem solving that requires only short derivations; advanced mathematics emphasizes sustained reasoning and multi-step problem solving over challenging mathematical tasks; and

 doi:10.20944/preprints202605.0806.v1

STEM mathematical reasoning targets quantitative reasoning embedded within authentic scientific or engineering contexts.

3.2.1. Basic Mathematics

Basic mathematics benchmarks target the ability to solve short- to medium-horizon quantitative problems drawn from elementary arithmetic, introductory algebra, and basic geometry [114,144,145,149,155]. They are defined not simply by the presence of numbers, but by tasks that require mapping short verbal or diagram-grounded descriptions into elementary operations, equations, or intermediate quantities with exact answers [74,112,149,150,152]. The emphasis is usually on procedural correctness over familiar school-level content, rather than advanced theory or long formal derivations [74,114,145].

Canonical examples include arithmetic and algebra word-problem benchmarks such as SVAMP, GSM8K, GSM-HARD and ASDiv [43,74,144–149]. Related datasets extend the same target to elementary geometry and mixed-format reasoning, including Geometry3K, GeoQA, PGPS9K, UniGeo, GeoQA+, Euclid30K, UGMATHBENCH, and MATHEMAGIC [29,112,150–155]. Some benchmarks further stress robustness by perturbing wording, quantities, or symbolic templates, testing whether models capture underlying quantitative structure rather than brittle lexical patterns [43,144,147,148]. Across these variants, the core target remains elementary mathematical composition with relatively short derivational chains [43,74,144,147]. Once tasks require competition-level ideas, substantial theorem use, or sustained symbolic derivations, the focus shifts toward advanced mathematics or broader STEM reasoning benchmarks [75,158,170,171].

3.2.2. Advanced Mathematics

Advanced mathematics benchmarks target sustained mathematical problem solving beyond elementary curricula. Their difficulty lies not only in long solution chains, but also in selecting the right mathematical ideas, maintaining symbolic consistency, and producing coherent derivations or proof sketches under limited prompt scaffolding [75,132,158]. In this benchmark object, the emphasis is natural-language mathematical reasoning rather than routine calculation or short-form answer retrieval [75,132,158].

This object includes several closely related benchmark families. Foundational static corpora include MATH and a growing set of contest-style collections such as OlympiadBench, PutnamBench, and Putnam-AXIOM, which span olympiad, Putnam-level, and other advanced competition problems across algebra, number theory, geometry, and combinatorics [42,75,156–159,163–165,292]. A second line of work focuses on harder or less saturated evaluation, as in FrontierMath and MathArena, which aim to better distinguish genuine reasoning from benchmark memorization or contamination [132,161]. A third line examines model robustness, formal verification, and proof-sensitive mathematical reasoning through benchmarks such as MATH-Perturb, HARD2VERIFY, IMProofBench, and SKYLENAGE-MATH [121,127,160,162].

3.2.3. STEM Mathematical Reasoning

STEM mathematical reasoning benchmarks target quantitative reasoning embedded in scientific or engineering contexts. Unlike advanced mathematics benchmarks, their focus is not mathematical derivation in isolation, but the coordinated use of equations, units, laws, diagrams, and domain assumptions to model and solve problems in fields such as physics, chemistry, and engineering [156,170,171,293–295]. In these benchmarks, successful reasoning depends not only on correct calculation, but also on choosing the right scientific formalism and respecting domain constraints throughout the solution process.

This object encompasses several closely related benchmark families, each targeting distinct yet interconnected aspects of reasoning. Cross-domain suites such as SciBench and TheoremQA evaluate whether models can identify and apply the appropriate principle across mathematics, physics, chemistry, finance, and computing, rather than merely execute symbolic manipulation [170,171,296]. A second family emphasizes physics-centered quantitative reasoning, including UGPhysics, PHYSICS, and

 doi:10.20944/preprints202605.0806.v1

ABench-Physics, where benchmark design stresses formula selection, symbolic consistency, physical interpretation, and stronger leakage control on advanced undergraduate problems [293–295]. A third family extends the object into multimodal and visually grounded settings, such as MathVista, STEM-POM, MathVerse-Plus, and OlympiadBench, where diagrams, spatial structure, or scientific visual context are integral to the reasoning process rather than peripheral presentation features [156,167–169].

Key Incomparability: Mathematical Reasoning

- ***Why the distinction matters***: distinct facets of mathematical reasoning competence.
- *Basic mathematics* benchmarks test short-chain elementary quantitative problem solving.
- *Advanced mathematics* benchmarks test sustained reasoning on advanced problems.
- *STEM mathematical reasoning* benchmarks test quantitative reasoning integrated with scientific or engineering structure.

3.3. Knowledge Reasoning

Knowledge reasoning can be divided into two primary benchmark reasoning objects: **Multi-hop Reasoning** and **Long Context Reasoning**. Multi-hop reasoning focuses on composing several separated pieces of evidence into a coherent inference, while long context reasoning focuses on locating, retaining, and integrating relevant information across very large context windows.

3.3.1. Multi-hop Reasoning

Multi-hop reasoning benchmarks evaluate whether a system can connect several separated pieces of evidence and use them jointly to reach one answer or conclusion [172–174,179,180]. What makes this object distinct from single-passage question answering is that no single fact is sufficient on its own: the model must identify intermediate links and compose them across documents, modalities, or retrieval steps [173,174,179,180]. Accordingly, the core benchmark target is not knowledge access by itself, but evidence composition under conditions where the reasoning chain is frequently rendered observable, systematically verifiable, or explicitly safeguarded against single-hop shortcuts [174,177–180].

Several benchmark families fall under this object. Classic multi-document QA benchmarks require models to connect evidence distributed across multiple sources and, in many cases, recover an implicit or annotated reasoning path [174,177–180]. Other benchmarks place more weight on decomposition, explanation, or fact aggregation. StrategyQA, QAMPARI, WikiWhy, and MINTQA are representative here, since answering them depends less on extracting one decisive span than on assembling several supporting statements into a coherent inference [56,172,175]. The same benchmark object also extends beyond plain text: MultiModalQA, FanOutQA, MM-BRIGHT, and ChartQAPro distribute relevant evidence across tables, images, charts, webpages, or other modalities, so successful reasoning requires cross-modal evidence integration rather than textual chaining alone [173,190,191]. More recent benchmarks further combine multi-hop reasoning with retrieval-intensive or agentic settings, including Bamboogle, BrowseComp, and FinReflectKG-MultiHop [33,72,181–183,185,186,188,189].

3.3.2. Long Context Reasoning

Long context reasoning is a benchmark object whose target capability is maintaining and integrating relevant information when the evidence is distributed across very large context windows. Its defining benchmark demand is the ability to perform scale-sensitive reasoning under memory pressure, not simply the existence of multiple discrete facts. The object is instantiated in benchmarks where accuracy depends on locating, retaining, and combining information as context length grows, often under conditions where retrieval by superficial salience is unreliable [44,45,71,79,297–299].

Realistic multitask suites such as LongBench and LongBench v2 benchmark this object by collecting summarization, question answering, retrieval, and reasoning settings that span long documents and multi document corpora [71,79]. Their contribution is to shift the benchmark target from isolated

 doi:10.20944/preprints202605.0806.v1

needle retrieval to mixed long context workloads with more natural task formats [71,79]. A second family exploits synthetic controllability to target and isolate specific long-context failure modes for fine-grained analysis [45,297]. RULER and LongReason vary context length, distractor structure, and reasoning pattern so that one can separate failures of memory, localization, and aggregation [45,297]. A third family stresses extreme length or uniform evidence relevance [44,298]. InfiniteBench pushes toward very large contexts, while Loong constructs settings where many documents matter and no single easy retrieval shortcut suffices [44,220,298].

Key Incomparability: Knowledge Reasoning

- ***Why the distinction matters***: distinct facets of knowledge reasoning competence.
- *Multi-hop reasoning*: benchmarked as the process of assembling multiple separated pieces of evidence into a single, well-supported answer or conclusion.
- *Long context reasoning*: benchmarked as identifying and integrating pertinent information distributed across very large contexts.

3.4. Agentic Reasoning

Agentic reasoning can be divided into four benchmark objects: **Planning and Decision-making**, **Tool Use**, **Code Reasoning**, and **Multi-Agent Collaboration**. Planning and decision-making focuses on sequential action selection in evolving environments, tool use focuses on the correct selection and invocation of external interfaces, code reasoning focuses on reasoning over program semantics and executable correctness, and multi-agent collaboration focuses on coordinated problem solving across multiple agents.

3.4.1. Planning and Decision-making

Planning and decision-making benchmarks evaluate the ability to select actions over time in an evolving environment rather than solve a static input–output problem [78,92,222,300]. Their target is sequential competence under feedback, partial observability, and state changes induced by previous actions, so benchmark success depends on trajectory quality rather than local response correctness alone [47,78,92,222,244,300].

One benchmark family uses interactive text environments [92,300]. ALFWorld operationalizes this capability through long-horizon household tasks requiring navigation, subgoal ordering, and object-state tracking [300], while ScienceWorld extends the target to scientific experimentation, where agents must plan action sequences that reveal, manipulate, and reason about environmental state to complete a final objective [92].

A second family evaluates the same capability in realistic digital environments. WebArena, Mind2Web, Mind2Web 2, and related benchmarks target long-horizon web interaction, where success requires action sequencing, interface interpretation, and recovery from earlier mistakes under real website constraints [33,209,222–224,227,228]. A parallel line extends sequential decision-making to shopping, operating systems, mobile apps, and heterogeneous digital tasks, including WebShop, OSWorld, WindowsAgentArena, and others [229,237,244–248]. AgentBench broadens the scope further, shifting the evaluation object from domain-specific success to the generality of sequential decision-making competence across multiple environments [78]. A growing body of work also instantiates planning and decision-making in specialized domains [31,234,236,238,241,242].

3.4.2. Tool Use

Tool use is a benchmark object whose target capability is deciding when external tools are needed, selecting the appropriate interface, and producing valid calls that advance the task state. What makes it a distinct benchmark object is that correctness is grounded in executable interaction with APIs, functions, databases, or simulators, rather than in free form language alone [46,249,256].

 doi:10.20944/preprints202605.0806.v1

The benchmark target is often compositional because a system must choose tools, supply arguments, interpret outputs, and sometimes abstain when no tool should be called [47,257,301].

Runnable API benchmarks form one core family [249]. API-Bank frames tasks as dialogues with executable APIs, so the benchmark directly measures whether tool invocation resolves the user request under realistic interface constraints [249]. A second family scales the interface space [46,256]. ToolBench exposed models to large collections of real APIs, and StableToolBench refined that design by stabilizing the underlying execution environment so that benchmark results depend less on external service volatility [46,256]. Another family focuses on tool awareness and function calling [257,301]. MetaTool benchmarks whether the model knows when to use a tool and which one to choose, while BFCL benchmarks structured function calling, including multi turn and multilingual settings where correctness can be checked against argument structure [257,301]. τ-bench adds stateful conversations with domain APIs and policy constraints, making the benchmark target reliable tool mediated task completion [47].

3.4.3. Code Reasoning

Code reasoning is a benchmark object whose target capability lies in understanding, generating, executing, or modifying programs in ways that are faithful to program semantics and intended behavior [80,261,271,272,302–304]. It is distinct from generic tool use because code is not merely an external aid, but the object of reasoning itself, with correctness typically grounded in test execution, runtime behavior, or repository level constraints [80,261,271,272,302,304]. The benchmark target therefore concerns semantic alignment between specification, program, and observed execution [80,261,271,272,302,304]. Function level synthesis benchmarks are the most established family [261,272]. HumanEval and MBPP instantiate the object by pairing natural language specifications with hidden unit tests, so benchmark success depends on producing code that generalizes beyond the visible prompt and passes executable correctness checks [261,272]. A second family targets code understanding rather than synthesis alone [302]. CRUXEval uses input and output prediction style tasks to benchmark whether the system can reason about program behavior, edge cases, and execution traces [302]. A third family addresses contamination and temporal freshness [271]. LiveCodeBench continuously updates coding tasks so that benchmark results better reflect current reasoning ability on previously unseen problems [271]. Finally, repository scale benchmarks such as SWE-bench, SWE-bench Verified, and SWE-bench-Live move the object from isolated functions to real software maintenance, where success requires navigating codebases, editing multiple files, and satisfying regression tests tied to authentic issue reports [80,280,281,303,304].

3.4.4. Multi-Agent Collaboration

Multi-agent collaboration is a benchmark object whose target capability lies in solving tasks through structured coordination and communication among multiple reasoning entities. It is distinct from single agent planning because benchmark success depends not only on local decision quality, but also on communication, role allocation, information sharing, and adaptation to the actions of teammates. The object is defined by benchmarks where no single agent view or policy is sufficient, either because information is distributed, workloads must be partitioned, or synchronized action is required [288,305–307].

Embodied or simulated team environments form one important family [305,306,308]. VillagerBench instantiates the object through collaborative tasks in a Minecraft like world, where agents must divide labor, coordinate temporally, and react to changing task dependencies [305]. Collab-Overcooked provides a complementary family centered on tightly coupled coordination with natural language communication, which makes benchmark success sensitive to both action timing and communication usefulness [306]. A second family benchmarks asymmetric information and communication explicitly. COMMA uses multimodal puzzles in which agents hold different pieces of evidence, so the benchmark target becomes the quality of inter agent message passing and joint inference [307]. A third family broadens coverage across settings [288]. MultiAgentBench benchmarks collaboration and, in

 doi:10.20944/preprints202605.0806.v1

certain settings, competition across a wide range of scenarios, positioning the benchmark less as a test of performance within a single environment and more as an assessment of broad collaborative competence.

Another case is with social interaction benchmarks that measure persuasion, persona consistency, or conversational realism without requiring joint task completion. Those settings are adjacent, but they do not define this object unless successful benchmarking depends on coordinated problem solving across agents. Multi-agent collaboration is therefore best understood as a benchmark object for collective reasoning under communication and coordination constraints [288,307,309].

Key Incomparability: Agentic Reasoning

- ***Why the distinction matters***: distinct facets of agentic reasoning competence.
- *Planning and decision-making*: benchmarked as selecting actions over time in an evolving environment.
- *Tool use*: benchmarked as choosing and correctly invoking external tools through executable interfaces.
- *Code reasoning*: benchmarked as reasoning about program semantics via the generation, comprehension, or transformation of code.
- *Multi-agent collaboration*: benchmarked as accomplishing complex tasks through structured coordination and information exchange among multiple agents.

4. Setting (How): Construction of Reasoning Benchmarks

A reasoning benchmark does not evaluate reasoning in the abstract. What it measures depends both on **how benchmark instances are constructed and on how the evaluation protocol is assessed**. We therefore organize benchmark construction along two complementary dimensions. The first is **Data Provenance**, which concerns where benchmark instances come from and how they are collected, synthesized, or curated. The second is **Protocol Formulation**, which specifies the setting under which reasoning is evaluated. Together, these dimensions clarify what evidence is available, what actions are permitted, what outputs are expected, and what counts as success, making benchmark comparisons more faithful and diagnostically meaningful.

 doi:10.20944/preprints202605.0806.v1

4.1. Data Provenance

Table 1. Representative benchmark construction works by data provenance, categorized into naturalistic real-world-derived, interaction-derived, expert-curated, model-generated, and human-AI collaborative benchmarks, each with specific tasks and domains.

Category	Benchmark	Description
	Naturalistic Data	
Real-World-Derived Benchmarks	ReClor [134]	Standardized logical reading exams.
	MATH [75]	Competition math problems.
	NaturalProofs [82]	Theorems and proofs in natural language.
	SCROLLS [192]	Naturally long documents from many domains.
	SWE-bench [80]	GitHub issues, PRs, and repository states.
	SWE-Bench Pro [213]	Harder engineering tasks from active repositories.
	LiveBench [310]	Newly released competitions, papers, and news.
	LongBench v2 [71]	Documents, dialogues, repositories, and tables.
	DeepScholar-Bench [311]	Related-work synthesis from arXiv papers.
	TEMPO [189]	Time-evolving evidence collections.
	XCR-Bench [312]	Culturally grounded parallel corpora.
Interaction-Derived Benchmarks	WebShop [229]	Product catalog with instructions and demonstrations.
	Mind2Web [209]	Real website action trajectories.
	WebLINX [313]	Conversational web navigation traces.
	OSWorld [244]	Real OS and application tasks.
	MCP-Bench [314]	Tasks built on live MCP servers.
	MCP-Atlas [254]	Larger MCP server and workflow collection.
	Constructed Data	
Expert-Curated Benchmarks	GPQA [34]	Expert-authored frontier questions.
	FrontierMath [161]	Expert-written advanced math problems.
	Humanity's Last Exam [59]	Specialist-authored hard questions.
	FOLIO [28]	Natural language with first-order logic.
	ProcessBench [41]	Expert-labeled reasoning error steps.
Model-Generated Benchmarks	AutoLogi [99]	Generated logic puzzles with verification.
	MPBench [315]	Generated multimodal process-error data.
Human–AI Collaborative Benchmarks	MMLU-ProX [316]	LLM translation plus expert review.
	SuperGPQA [317]	Expert writing with human–LLM filtering.
	DocPuzzle [206]	Human–AI annotation and validation loop.

Focusing specifically on this dimension, we systematically decompose **Data Provenance** into two distinct aspects, as illustrated in Figure 5: **Naturalistic Data** and **Constructed Data**.

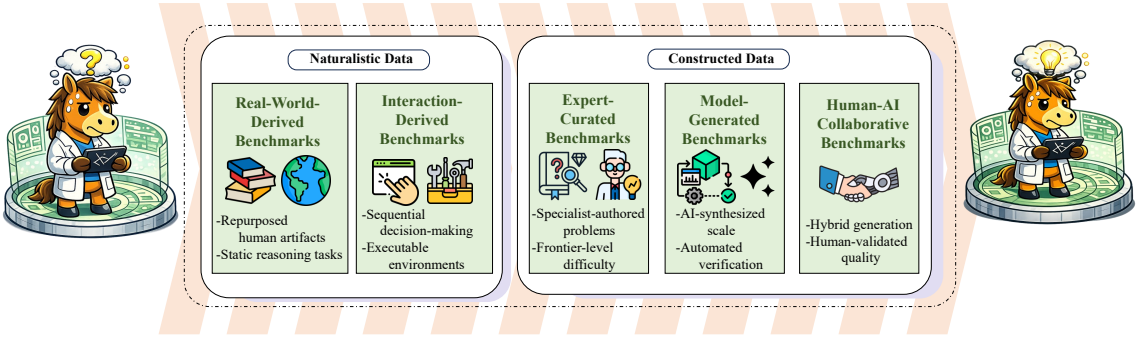

Figure 5. Benchmarks data provenance is structured around two primary data paradigms: Naturalistic Data (Real-World-Derived and Interaction-Derived) and Constructed Data (Expert-Curated, Model-Generated, and Human-AI Collaborative).

4.1.1. Naturalistic Data

Naturalistic data is harvested from real-world sources and categorized into **Real-World-Derived Benchmarks** and **Interaction-Derived Benchmarks**. Once collected and refined, these datasets serve as the primary foundation for benchmark construction.

Real-World-Derived Benchmarks.

A first line of work constructs reasoning benchmarks by mining naturally occurring artifacts originally created for human use rather than for model testing. Early examples repurpose standardized examination materials: ReClor converts graduate-level logical reading questions into benchmark instances [134], while MATH draws on competition mathematics to yield challenging step-

 doi:10.20944/preprints202605.0806.v1

based reasoning problems [75]. NaturalProofs extends this approach to formal knowledge resources, organizing theorem statements and proofs in natural mathematical language into benchmarkable reasoning units [82].

This paradigm later moves beyond isolated questions to richer and more structurally complex real-world artifacts. SCROLLS demonstrates that naturally long documents spanning multiple domains can be standardized into long-context reasoning tasks without artificial context extension [192]. SWE-bench shows how GitHub issue reports, pull requests, and repository snapshots can be converted into executable software-reasoning instances [80], and SWE-Bench Pro continues this trajectory by curating longer-horizon engineering problems from actively maintained repositories [213]. LiveBench pushes toward freshness by continuously sourcing problems from newly released competitions, papers, and news [310], while LongBench v2 broadens the naturalistic scope by packaging documents, dialogue histories, repositories, and structured data into realistic long-context tasks [71]. Recent work further specializes naturalistic corpora for emerging reasoning needs: DeepScholar-Bench derives related-work synthesis tasks from recent arXiv papers [311], TEMPO organizes temporally evolving evidence into cross-period retrieval tasks [189], and XCR-Bench turns culturally grounded parallel corpora into benchmark instances [312].

Interaction-Derived Benchmarks.

A second line of work constructs benchmarks from realistic interaction data, where reasoning is embedded in sequential decision-making. WebShop couples a large real-product catalog with crowdsourced instructions and demonstrations, transforming shopping interactions into benchmark tasks that require grounded multi-step reasoning [229]. Mind2Web advances this by collecting action sequences on real websites, preserving authentic page structure, user intent, and action dependencies within each benchmark instance [209]. WebLINX further scales interaction harvesting to multi-turn conversational web navigation with expert demonstrations, making dialogue history itself part of the benchmark state [313].

More recent work compiles interaction data into executable environments rather than static trajectories. OSWorld instantiates open-ended computer tasks over real operating systems and applications, equipping each benchmark instance with an initial machine state and an execution-based verifier [244]. MCP-Bench extends this construction to tool-use settings by building tasks on live MCP servers, where agents must coordinate tool discovery, parameter grounding, and multi-step execution [314]. MCP-Atlas scales the design to a broader collection of real MCP servers and cross-tool workflows, illustrating a broader shift in interaction-derived benchmarks toward executable state, realistic affordances, and workflow-level dependencies [254].

4.1.2. Constructed Data

Conversely, constructed data is synthesized by either humans or AI models and is further partitioned into **Expert-Curated Benchmarks**, **Model-Generated Data** and **Human–AI Collaborative Benchmarks**. These datasets are typically engineered to evaluate specific reasoning capabilities or to address the inherent limitations of naturalistic data, offering a complementary foundation for benchmark development.

Expert-Curated Benchmarks.

Expert-curated benchmarks emphasize deliberate problem authoring, difficulty control, and contamination resistance during dataset construction. One prominent line of work builds frontier-level reasoning benchmarks by commissioning domain specialists to write original questions that are difficult to solve through memorization or shallow retrieval, as exemplified by GPQA, FrontierMath, and Humanity's Last Exam [34,59,161]. Another line focuses on structural rigor by coupling natural-language instances with formal representations or process-level annotations. For example, FOLIO grounds deductive reasoning examples in first-order logic, while ProcessBench operationalizes reasoning-quality assessment through expert-annotated error locations in step-by-step mathematical

 doi:10.20944/preprints202605.0806.v1

solutions [28,41]. In these benchmarks, the central contribution is not merely a test set, but a carefully designed construction protocol that encodes the targeted reasoning skill into the data itself.

Model-Generated Benchmarks.

A complementary paradigm uses LLMs or programmatic generators to synthesize benchmark instances at scale. Rather than manually authoring every example, these benchmarks define controllable generation procedures and then rely on automatic verification, rejection sampling, or post-filtering to preserve quality. AutoLogi, for instance, transforms logic reasoning evaluation from multiple-choice questions into open-ended logic puzzles with controllable difficulty [99]. Similarly, MPBench constructs multimodal reasoning data for process-error identification, extending constructed benchmarks from answer-level supervision to step-aware and process-aware assessment [315]. Such model-generated benchmarks are especially useful when the goal is to cover large combinatorial spaces, create adversarial variants, or attach fine-grained intermediate supervision that would be prohibitively expensive to obtain fully by hand.

Human–AI Collaborative Benchmarks.

Between fully manual and fully synthetic construction lies a hybrid paradigm in which models draft, translate, perturb, or filter candidate items and humans subsequently validate them. This design has become increasingly common in recent reasoning benchmarks. MMLU-ProX uses LLM-assisted translation followed by expert review to build parallel multilingual reasoning questions, enabling controlled cross-lingual comparison without sacrificing conceptual fidelity [316]. SuperGPQA combines large-scale expert question writing with human–LLM collaborative filtering to eliminate trivial or ambiguous items across a broad range of graduate-level disciplines [317]. DocPuzzle similarly adopts an annotation-validation loop to construct realistic long-context reasoning problems with explicit attention to process complexity [206]. Overall, recent benchmark construction has shifted from static answer-only datasets toward data pipelines that explicitly encode frontier difficulty, process supervision, multilingual transfer, and scalable but human-grounded quality control.

Key Incomparability: Data Provenance

- ***Why the distinction matters***: data provenance shapes benchmark realism, controllability, and contamination risk, so benchmark results are not fully comparable across different construction sources.
- *Naturalistic data*: constructed from real-world artifacts or interaction traces.
- *Constructed data*: constructed from intentionally designed instances, whether human-authored, model-generated, or human–AI co-created.

4.2. Protocol Formulation

Focusing specifically on this latter dimension, we systematically decompose **Protocol Formulation** into four distinct layers, as illustrated in Figure 6: **Task Interface**, **Knowledge Access**, **Tooling and Environment**, and **Constraints and Controls**.

 doi:10.20944/preprints202605.0806.v1

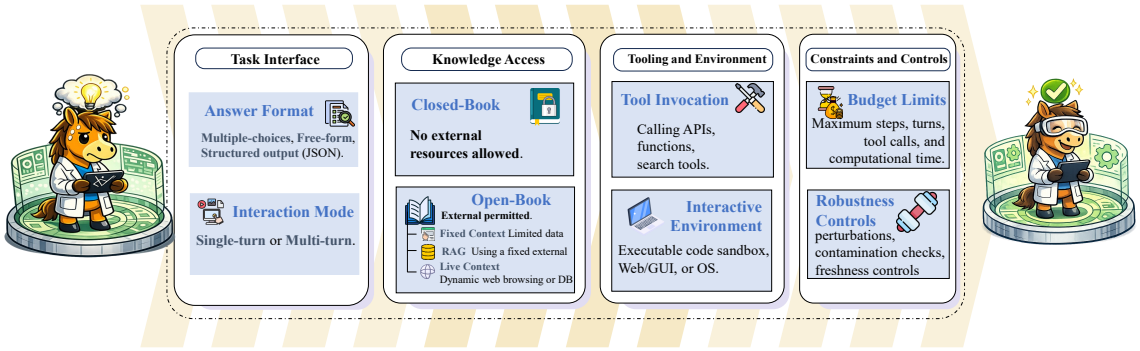

Figure 6. Setting layers that define benchmark protocol formulation: a four-dimensional framework comprising Task Interface, Knowledge Access, Tooling and Environment, and Constraints and Controls.

4.2.1. Task Interface

Answer Format

Answer format defines the set of outputs a benchmark accepts as valid. It determines how tightly models are supervised, how ambiguous scoring can be, and whether evaluation primarily rewards reasoning alone or reasoning plus faithful realization. We analyse this dimension along three common formats: fixed-choice, free-form, and structured outputs.

Fixed-choice formats constrain outputs to a closed candidate set, improving reproducibility and scoring stability by minimizing surface-form variance. This design forms the basis of many multiple-choice benchmarks, from RiddleSense [318] and MedMCQA [319] to more recent datasets such as MME-RealWorld [320], SATBench [321], and CogToM [322]. Some benchmarks deliberately mix fixed-choice with other formats for broader diagnosis: LogicBench combines binary QA and multiple choice [50], SceMQA mixes multiple-choice and free-response questions [323], and OlympiadBench includes standardized open-ended answers alongside proof problems [156].

Free-form formats allow unconstrained textual answers and are better suited to end-to-end problem solving, but they require normalization, objective references, or judge-based evaluation. Short-answer settings emphasize deterministic checking, as in GSM8K [74], SVAMP [144], and objectively scored tasks in LiveBench [310]. Other benchmarks rely on richer open-ended generation or preference judgments, including AlpacaEval-LC [324], U-MATH [325], and TYPED-RAG [326].

Structured formats require schema-compliant or executable outputs, enabling verification beyond string matching. Formal reasoning benchmarks such as miniF2F [102] and miniF2F-v2s [327] treat valid outputs as formal statements or proofs. This requirement is also common in tool-use benchmarks, including API-Bank [249], ToolBench [256], and TOOLSANDBOX [53], which require structured API calls, typed arguments, or replayable action traces. Software engineering benchmarks push this logic further by evaluating executable patches against repository-level tests, as in SWE-bench [80] and SWE-Bench Pro [213].

Interaction Mode

Interaction mode specifies whether a benchmark evaluates a single response under a fixed input or an interactive policy that must act over multiple steps with feedback. It therefore distinguishes one-shot inference from temporally extended behavior. We analyse this dimension along two common modes: **single-turn benchmarks** and **multi-turn benchmarks**.

Single-turn benchmarks provide the full context upfront and score a single final response, making them well suited to controlled comparison of reasoning, knowledge use, long-context understanding, and structured generation. Representative examples include GPQA [34], LongBench [79], and MMLU-Pro [68]. **Multi-turn benchmarks** instead evaluate policies that repeatedly observe, act, and revise before termination. They are necessary when success depends on exploration, tool use, social interaction, or recovery from intermediate errors over long horizons, as in benchmarks such as WebShop [229], WebArena [222], Mind2Web 2 [228], and TERMINAL-BENCH [328].

 doi:10.20944/preprints202605.0806.v1

4.2.2. Knowledge Access

Knowledge access defines the extent to which a model can retrieve, ground, and reason over information that may lie beyond its parametric knowledge. Benchmarks in this space are organized along a single axis: whether external retrieval is permitted at inference time. Rather than treating knowledge as a fixed property of the model, this dimension asks how reliably a model can locate relevant evidence, whether stored internally or sourced externally, and integrate it into coherent, accurate responses. We analyze this dimension along two common protocols: **closed-book protocol** and **open-book protocol**.

Closed-book Protocol

Closed-book Protocol forbids external retrieval at inference time, so the model must answer from parametric knowledge and the information already contained in the prompt. This setting isolates internal knowledge and reasoning from retrieval ability [67,68,146].

A useful variant is given-context closed-book evaluation, where all supporting evidence is embedded in a long input rather than retrieved externally. This setting tests evidence localization, aggregation, and long-range reasoning under fixed inputs. Representative benchmarks include L-EVAL [329], ∞BENCH [44], RULER [45], MMLongBench-Doc [330], NeedleBench [200], and LongDocURL [221], along with other relevant evaluation settings [79,202,203,322].

Open-book Protocol

Open-book Protocol allows access to external evidence at inference time, so performance depends on evidence acquisition, selection, grounding, and temporal freshness in addition to reasoning. One major branch of existing work relies on fixed corpora and standardized retrieval pipelines, encompassing curated benchmark datasets, comprehensive evaluation suites, and well-defined document-level RAG settings such as MultiHop-RAG [185], RAGAS [30], ChatQA [331], RAGBench [332], and MMRAG-DocQA [333].

Another line of research focuses on evaluating live or continuously updated knowledge access, with particular emphasis on information freshness, web browsing capabilities, citation accuracy, and the ability to adapt to real-time changes. Representative examples include LiveBench [310], Chatbot Arena [334], and FINDEEPFORECAST [335].

4.2.3. Tooling and Environment

Tooling and environment evaluate a model's capacity to act beyond text generation by interacting with external systems, executing operations, and adapting to feedback from a changing world. We analyze this dimension along two common settings: tool invocation and interactive environment.

Tool Invocation

Tool invocation evaluates whether a model can decide when to call tools, select appropriate APIs, and incorporate returned observations into subsequent reasoning. This line of work spans early tool-use paradigms and training setups such as Toolformer [255], API-Bank [249] and ReAct [13], as well as dedicated evaluation suites such as ToolBench [256], ComplexFuncBench [336], and BFCL [257]. Recent benchmarks go beyond isolated function calls, focusing on multi-tool orchestration, user–agent interaction, and coordination, as illustrated by WebArena, AgentBench, and M^3-Bench [222,314,337–339].

Single-tool settings isolate call or no-call decisions, tool retrieval or selection, and argument grounding under relatively fixed schemas, making failure modes such as malformed arguments, irrelevant calls, and schema violations easier to diagnose [187,249,257,301,336]. Multi-tool settings evaluate routing across heterogeneous tools, intermediate-state tracking, and recovery over longer workflows or conversations, where success depends on correct sequencing, dependency management, and robust use of tool outputs [288,338,339].

 doi:10.20944/preprints202605.0806.v1

Interactive Environment

Interactive environments place reasoning in a stateful world whose state evolves after each action. They span execution-grounded coding tasks, web and GUI interaction, OS-level control, terminal use, and user-interactive settings, enabling evaluation of long-horizon control, recovery, and environment-grounded verification [248,272,328,340].

Code execution environments let models write or modify code and receive runtime signals such as compiler feedback, unit-test results, repository-level regression outcomes, or execution traces. They are well suited for evaluating iterative debugging and execution-grounded reasoning [80,271–274,279,341]. While GUI-based environments require perception and action grounding over webpages, screenshots, desktop or mobile interfaces, terminals, and evolving user interactions. They therefore test whether language-level plans can be reliably translated into interface actions under partial observability and procedural constraints [72,209,222,229,248,328,342].

4.2.4. Constraints and Controls

Constraints and controls examine whether evaluation outcomes remain meaningful when the conditions of inference are made more realistic or more adversarial. A model that performs well under unconstrained, clean-input settings may degrade substantially when resources are capped, inputs are perturbed, or test items have leaked into pretraining, yet standard benchmarks rarely make these failures visible. We analyze this dimension along three common settings: budget limits, robustness controls, and protocol formulations.

Budget Limits

Budget limits impose explicit caps on actions, tool use, compute, or reasoning length, forcing models to allocate exploration, verification, and recovery under fixed resources. They make evaluation more deployment-relevant by separating genuine decision quality from gains that come purely from unconstrained computation. In practice, such controls appear as step limits, tool-call limits, cost-aware planning objectives, and token-efficiency mechanisms in recent evaluation suites and methods, including HELM [343], LiveCodeBench [271] and Dynamic Thinking-Token Selection [344].

Robustness Controls

Robustness controls stress models under controlled variation and guard evaluation against leakage or stale knowledge, making failures attributable to specific instability modes rather than to underspecified test conditions. Rather than treating accuracy as a single static quantity, they make failure modes more diagnostic by tying errors to identifiable forms of instability. One common strategy is perturbation-based stress testing, where semantics-preserving changes such as paraphrases, distractors, rewrites, or format shifts are introduced to check whether predictions remain stable [345,346]. Another strategy emphasizes contamination resistance and temporal validity through benchmark design choices such as rolling refresh, temporal cutoffs, and controlled answer exposure [59,271,310].

 doi:10.20944/preprints202605.0806.v1

Key Incomparability: Protocol Formulations

- ***Why the distinction matters***: benchmark settings shape what performance actually reflects, so results are not fully comparable across different interfaces, access conditions, and control regimes.
- *Answer format*: benchmarked through fixed-choice, free-form, or structured outputs.
- *Interaction mode*: benchmarked as single-turn response or multi-turn policy execution.
- *Knowledge access*: benchmarked under closed-book or open-book conditions.
- *Tooling and environment*: benchmarked with or without tool invocation and interactive environments.
- *Constraints and controls*: benchmarked under explicit budget limits and robustness controls.

5. Evaluation (How Well): Assessment of Reasoning Benchmarks

Evaluation determines what a benchmark rewards, what it ignores, and how its scores should be interpreted. A single score can report progress while hiding whether gains come from better reasoning, easier outputs, looser matching rules, or heavier compute. As shown in FigureS 7 and 8, to make metric choices explicit, we structure assessment into two components: **Evaluation Unit** and **Evaluation Dimensions**.

These dimensions jointly **depict not only raw performance, but also the credibility, robustness, and real-world deployability of reasoning systems** [41,78,343,347,348].

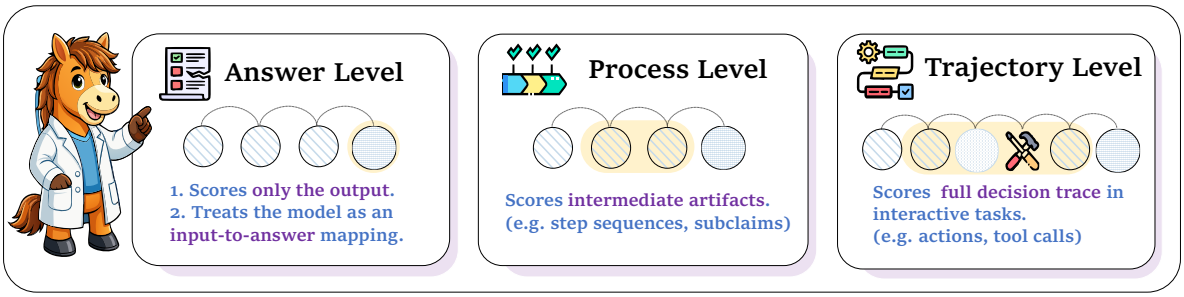

Figure 7. The conceptual architecture of the evaluation framework adopted throughout the survey, organized into three hierarchical levels that reflect increasing degrees of observational granularity: Answer Level, Process Level, and Trajectory Level.

5.1. Evaluation Unit

Evaluation Unit specifies which object is scored, ranging from a final answer to an interaction trace. Thus, we categorize evaluation units into three levels of granularity: **answer-level, which focuses on the final output; process-level, which examines intermediate reasoning artifacts; and trajectory-level, which considers the full decision trace in interactive settings**.

5.1.1. Answer-level

Answer-level evaluation focuses solely on the final output, treating the model as a mapping from input to answer. It remains the most widely used evaluation unit because it **scales easily across large benchmark suites**, **supports straightforward aggregation**, and **facilitates simple leaderboard comparisons**. This form of evaluation works best *when benchmarks specify a clear target and a stable matching criterion*, such as a unique option, a normalized short answer, or a numeric value with an allowed tolerance. Its limitation, however, is equally clear: final-answer correctness alone reveals little about how the result was obtained. A correct answer may reflect genuine reasoning or merely exploitation of superficial cues. For that reason, benchmark construction at this level often emphasizes tighter task design and stronger control of spurious shortcuts.

 doi:10.20944/preprints202605.0806.v1

In practice, answer-level scoring appears in several recurring benchmark formats. Broad multi-task suites unify heterogeneous tasks under standardized interfaces and summarize model performance with aggregate scores, making cross-model comparison easy and reproducible [67,70,310,343]. Reasoning-oriented benchmarks, by contrast, rely on expert-written or exam-style problems with uniquely verifiable answers, preserving the usefulness of final-answer accuracy even as difficulty rises [34,69,74,75]. Code benchmarks take a different route: they regard the generated program itself as the answer and judge it through execution, tests, or functional checks, thereby maintaining objectivity despite wide variation in surface form [261,271,272,349]. Some recent benchmarks further strengthen answer-level evaluation by refreshing instances over time and combining them with automated verification, which helps leaderboards remain informative even under rapid model iteration [69,271,304,310].

5.1.2. Process-level

Process-level evaluation examines the intermediate reasoning artifacts produced along the way, including step sequences, intermediate variables, proofs, subclaims, and structured rationales. Rather than treating reasoning as valuable only when it leads to the right endpoint, this level evaluates the quality of the process itself. Its advantage is most evident when intermediate steps can be checked automatically through symbolic validation, constraint checking, proof checking, or execution-based verification.

Another strength is diagnostic precision: instead of merely observing that a final answer is wrong, process-level scoring can identify where the reasoning first breaks down and whether the failure arises from conceptual misunderstanding or faulty rule application.

Benchmarks in this category generally follow two main directions. One line of work makes intermediate reasoning explicitly verifiable by representing it as objects such as entailment trees, natural-language proofs, or machine-checkable proof states; correctness can then be assessed incrementally rather than only at the endpoint [83–85,102]. Another line of research emphasizes diagnostic evaluation, providing step-level annotations that identify the earliest erroneous step or categorize the specific type of reasoning error within a trace [41,48,49,350].

5.1.3. Trajectory-level

Trajectory-level evaluation considers the entire decision trace in tasks that unfold over time, including actions, tool calls, observations, intermediate states, and termination choices. Here the system is evaluated not simply as an answer generator but as a policy operating through a sequence of decisions. This perspective is especially important in agentic and interactive settings, where identical final answers may arise from trajectories that differ sharply in efficiency, risk, and robustness.

Benchmarks at the trajectory level usually place the model inside an interactive environment and assess the resulting trace from multiple angles. Some emphasize long-horizon agent behavior, focusing on coherence and recovery under observational feedback while reporting both task success and trajectory-sensitive diagnostics [78,222,229,351]. Others center on tool use or API interaction, where actions are grounded in executable calls and structured state transitions, making it possible to evaluate traces through database state, argument correctness, and rule compliance [46,47,72,249]. In software engineering benchmarks, trajectories often consist of iterative code edits coupled with test feedback, which allows evaluation to penalize wasted iterations while rewarding fixes that remain robust beyond a single patch [46,78,80,304].

 doi:10.20944/preprints202605.0806.v1

Key Incomparability: Evaluation Unit

- ***Why the distinction matters***: evaluation units determine whether a benchmark measures only final correctness, intermediate reasoning quality, or full sequential behavior.
- *Answer-level*: benchmarked through the correctness of the final output alone.
- *Process-level*: benchmarked through the validity of intermediate reasoning steps or artifacts.
- *Trajectory-level*: benchmarked through the quality of the full action and observation trace over time.

5.2. Evaluation Dimensions

Evaluation dimensions determine **which aspects of performance a benchmark prioritizes and rewards**, as well as **how the resulting scores are to be interpreted**. In this survey, we organized evaluation dimensions into **three pillars: correctness, reliability, and efficiency**.

5.2.1. Correctness

Correctness evaluates whether model outputs satisfy the success criteria defined by a task, typically operationalized as either **Final-answer Accuracy** or **Step-level Scoring**.

Final-answer Accuracy

As for final-answer accuracy, evaluation checks whether the output matches the benchmark target under a predefined matching rule, such as exact match, normalized match, option selection, tolerance-based numeric match, or test-based acceptance. These rules are typically implemented through reusable evaluation templates. In exam-style and knowledge-intensive benchmarks, correctness is usually defined as selecting the keyed option under fixed decoding, as in MMLU, C-Eval, GPQA, and AGIEval [34,67,73,352]. Other difficult reasoning benchmarks retain the same answer-level notion of success while increasing task difficulty [96]. In free-form reasoning and math tasks, evaluation often extracts a short final answer and applies exact or normalized matching [70,343]. Some math benchmarks further strengthen this rule with symbolic equivalence or numeric tolerance in order to reduce sensitivity to superficial formatting differences while preserving strict correctness [74,75,96]. For executable tasks, correctness is instead determined by an external verifier such as unit tests or sandbox execution, turning natural-language or code outputs into behavior-level acceptance [75,261,271,272,353]. Under this final-answer view, *acc* reports the fraction of instances for which the top output is correct. It is therefore the standard metric in deterministic settings where each instance has a unique gold target and decoding is fixed, including multiple-choice exams and short-answer math word-problem benchmarks [34,73–75,352]. By contrast, *pass@k* measures whether at least one of the top k sampled candidates is correct. This metric is more appropriate in settings where stochastic sampling constitutes an integral part of the intended inference procedure and correctness is adjudicated by an external verifier, especially in functional-correctness evaluation for code synthesis and competition-style programming [41,52,75,261,271,272,353,354].

Step-level Scoring

As for step-level scoring, the correctness is assessed at the process level rather than at the endpoint. Recent process-oriented benchmarks introduce intermediate supervision or evaluation hooks by labeling step validity or requiring explicit step-wise outputs [41,52,354]. Such setups make it possible to evaluate critics and process-aware reward models directly, instead of inferring process quality only from final answers [102,127]. In multi-step mathematics, intermediate correctness concerns whether each step satisfies arithmetic and reasoning constraints, and benchmarks often operationalize this by asking models or verifiers to identify the earliest incorrect step or assign validity labels to each step [41,52,127,354]. This form of scoring supports finer-grained error diagnosis and objectives related

 doi:10.20944/preprints202605.0806.v1

to early detection of solution failure [74]. In deduction and proof, step validity can be grounded in machine-checkable proof obligations and tactic traces, so intermediate scoring reduces to whether each step preserves formal validity under a proof assistant or logic checker [28,102,113,290,355]. A related but distinct aspect is **calibration**, which asks whether the model assigns reliable confidence to answers or intermediate decisions. This matters in selective answering, abstention, and risk-aware deployment, where uncertainty quality is important alongside correctness itself. Calibration-oriented benchmarks therefore examine whether confidence aligns with empirical correctness and whether models can appropriately abstain or hedge under uncertainty, using reliability metrics and selective risk–coverage analyses in both short-form and long-form settings [343,356–359].

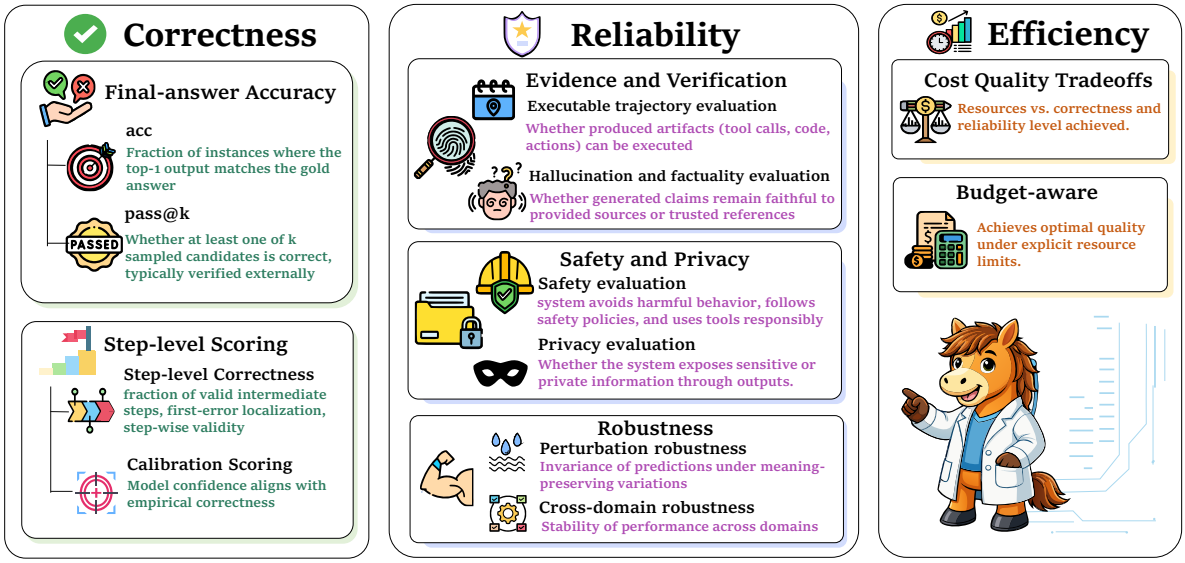

Figure 8. Framework overview of the evaluation protocol used throughout the survey, spanning three key pillars: Correctness (Final-answer Accuracy, Step-level Scoring, Calibration), Reliability (Evidence & Verification, Safety & Privacy, Robustness), and Efficiency (Cost-Quality Tradeoffs, Budget-aware Evaluation).

5.2.2. Reliability

Reliability concerns whether the reasoning is confident, consistent, and robust under varying conditions. [343,356–359]. To better depict this evaluation dimension, we categorize reliability into **three aspects: evidence and verification, safety and privacy, and robustness**.

 doi:10.20944/preprints202605.0806.v1

Table 2. Representative benchmark works by evaluation dimension, spanning correctness (e.g., reasoning, knowledge), reliability (e.g., safety, robustness), and efficiency (e.g., latency, cost), with details on each benchmark's design and focus.

Category	Benchmark	Description
	Correctness	
Final-Answer Accuracy	MMLU [73]	Standardized multi-subject multiple-choice exams under fixed decoding.
	C-Eval [352]	Chinese exam-style multiple-choice evaluation.
	GPQA [34]	Expert-authored frontier knowledge questions with answer-key matching.
	AGIEval [67]	Human exam and admission-test style academic evaluation.
Step-Level Scoring	Let's Verify Step by Step [354]	Step-wise mathematical reasoning supervision and verification.
	ProcessBench [41]	Expert-labeled reasoning steps and localized process errors.
	ProcBench [52]	Intermediate-step validity assessment in multi-step solutions.
	Hard2Verify [127]	Detection and evaluation of incorrect intermediate reasoning steps.
	Reliability	
Evidence and Verification	FactScore [360]	Claim-level factual support verification against trusted references.
	KALMA [361]	Attribution and citation-grounded verification of generated claims.
	FEVEROUS [362]	Fact verification with explicit evidence selection from documents and tables.
	FreshQA [181]	Time-sensitive question answering with date-appropriate evidence.
	APIBank [249]	Tool-use outputs validated through executable API calls and states.
	ToolEmu [363]	Safety-aware tool-use trajectories checked in runnable environments.
	TruthfulQA [364]	Truthfulness under misconception-inducing prompts.
	HaluEval [365]	Detection and avoidance of unsupported or contradictory generations.
Safety and Privacy	SafetyBench [366]	Policy compliance and safe response behavior across risk categories.
	AI Safety Benchmark [367]	Risk-aware behavior under realistic and adversarial prompting.
	HarmBench [368]	Harmful request compliance and refusal robustness.
	JailbreakBench [369]	Jailbreak resistance under adversarial attack prompts.
	ToolEmu [363]	Safe decision-making in interactive tool-using environments.
	ProPILE [370]	Memorization and private-string leakage under probing.
	PromptExtractionBench [371]	Hidden-prompt and secret extraction in application-like settings.
	PrivacyBench [372]	Long-horizon privacy preservation and selective disclosure.
Robustness	Dynabench [373]	Meaning-preserving adversarial and dynamically collected perturbations.
	Robustness Gym [374]	Controlled perturbation suites for invariance and failure analysis.
	PromptRobust [375]	Sensitivity to paraphrase, formatting, distractors, and prompt variation.
	HaluEval [365]	Robustness to misleading context and unsupported continuations.
	WebArena [222]	Interactive robustness under layout, state, and trajectory changes.
	Mind2Web [209]	Real-web task transfer under interface and action variation.
	Efficiency	
Cost–Quality Tradeoffs	HELM [343]	Joint reporting of accuracy, calibration, robustness, and efficiency signals.
	Bench360 [347]	Quality evaluation tied to latency and deployment-oriented system cost.
	FlexBench [376]	Flexible benchmarking of performance under practical serving constraints.
	BRACE [377]	Latency-sensitive evaluation in interactive reasoning settings.
	StableToolBench [46]	Task success evaluated together with tool-call efficiency.
	TAU [47]	Agent performance under tool budgets and interaction constraints.
	ToolRET [378]	Retrieval- and tool-augmented reasoning with explicit call efficiency.
	WebArena [222]	Interactive web-agent success under step and tool-use cost.
Budget-Aware Optimality	ESC [379]	Early stopping once sufficient evidence is gathered.
	Reasoning Token Evaluation [380]	Performance under explicit token-budget constraints.
	CogniLoad [381]	Benchmarking quality under controlled reasoning-load limits.
	TAU [47]	Budget-matched agent evaluation with step and query caps.
	HELM [343]	Holistic reporting across correctness, calibration, robustness, and efficiency.
	MLPerf Power [382]	Joint evaluation of model quality, speed, and energy usage.
	BRACE [377]	Tradeoff analysis between response quality and interactive latency.

Evidence and Verification

This concerns two closely related issues: whether a claimed solution is backed by inspectable support, and whether the benchmark can validate that support in a reliable way [181,360,361,383,384]. To make reliability operational, many benchmarks require support to be explicit and machine-checkable [343,360,361,383]. One line of work studies attribution and verifiable generation: models must answer with citations, and evaluation verifies whether the cited spans fully support, only partially support, or even contradict the associated claims [360,361,364,383]. Another line focuses on fact verification with evidence, scoring both the correctness of the verdict and the adequacy of the selected support, including dialogue-grounded checking and multimodal evidence verification [362,385–387]. Time-sensitive benchmarks add a further constraint: supporting evidence must be correct for the reference date, which helps expose stale factual recall and evidence drift in retrieval- or browsing-based systems [181,388]. Executable trajectory evaluation shifts the focus from stated answers to whether the produced artifact can actually run and be reproduced, including code, tool calls, intermediate computations, and environment actions; this perspective is central to tool-use and agent benchmarks [80,222,249,272,363]. Some benchmarks execute generated code or patches in sandboxes and score unit-test outcomes, emphasizing reproducibility, determinism, and failure localization in realistic repositories [80,261,272,302,389].

Hallucination evaluation asks whether generated content remains faithful to provided sources or other trusted references, a concern that is especially salient in open-book, retrieval-based, and long-context settings where fluent but unsupported content is a common failure mode [360,361,364,365,390].

 doi:10.20944/preprints202605.0806.v1

Some benchmarks probe truthfulness under misconception pressure by using questions that invite plausible but false answers and rewarding systems that avoid common falsehoods and misleading continuations [181,364]. Others construct hallucination-detection datasets with human annotations of unsupported or contradictory generations, then test whether systems can avoid or identify such content across diverse prompts and contexts [365,383]. A further strand decomposes model outputs into atomic claims and verifies each claim against trusted references, enabling claim-level auditing rather than relying only on final-answer accuracy [360,361,383].

Safety and Privacy

Safety and privacy assess whether a system remains appropriate, secure, and non-disclosive when confronted with risky prompts, sensitive contexts, or high-impact actions.[363,366,367,372]. Recent benchmarks focus on risk-aware behavior under adversarial prompting and realistic deployment conditions, often using runnable harnesses and standardized taxonomies to distinguish justified refusal, unsafe compliance, and tool-mediated downstream harm [343,363,366,367].

Safety evaluation asks whether a model avoids harmful behavior, adheres to safety policies, and handles tools responsibly when actions can create downstream consequences, especially in interactive or agentic settings[363,366–369]. Some benchmarks center on policy knowledge and compliance across categorized scenarios, measuring safe completions, refusal quality, and calibration across languages and domains [366,367]. Others probe adversarial misuse and jailbreak robustness with curated harmful goals and attack prompts, tracking harmful completion rates, refusal robustness, and over-refusal on benign inputs [368,369,391,392]. Agent-oriented evaluations instead place models in tool-using environments and test whether they recognize high-impact operations, avoid unsafe actions, and maintain safe execution over interactive trajectories [363,393].

Privacy evaluation asks whether a model leaks sensitive information, reproduces memorized private content, or reveals confidential context, making it especially relevant for private-document tasks, personal prompts, and contamination-sensitive setups[370–372,394]. One line of work studies memorization and training-data leakage by probing whether models can be induced to reproduce rare private strings or personally identifiable information through adaptive querying and prompt engineering [370]. Another examines prompt and secret exfiltration in application-like settings, asking whether hidden system prompts or confidential strings can be extracted through injection-style interactions that resemble deployed workflows [371,395]. Long-horizon conversational benchmarks, by contrast, simulate user-specific profiles and private memories to test selective disclosure, access control, and leakage over multi-turn interactions [372,394].

Robustness

Robustness asks whether a system maintains performance under intent-preserving variations and under distribution shifts that resemble real use [222,373–375,396]. Benchmarks usually instantiate this goal through controlled perturbations and shifted settings, then inspect invariance, degradation patterns, and failure modes to diagnose reliance on superficial cues or brittle heuristics [373–375,397]. Recent work broadens the scope to prompt injection, adversarial reasoning pressure, social interaction, and environment-level failure localization, especially in interactive settings [39,92,324,343,396,398–404].

Perturbation robustness tests invariance to paraphrases, formatting changes, reordered context, noise, and distractors, making it useful for exposing shortcut reliance and other brittle heuristics [373–375]. One line of work uses adversarial or dynamic data collection to generate meaning-preserving variants and pair them with diagnostics of what changed and why the model failed [373–375]. Another line builds structured perturbation suites, such as paraphrase, distractor, and format edits at scale, to support controlled ablations over wording, presentation, and spurious cues [181,365,375]. Interactive benchmarks extend this logic to trajectories: small changes in layout, action order, or tool outputs can alter the full interaction path, so robustness is evaluated beyond the final answer alone [78,209,222]. Cross-domain robustness measures whether performance transfers

 doi:10.20944/preprints202605.0806.v1

across domains, languages, modalities, and task templates, which is essential for benchmarks claiming broad reasoning rather than narrow adaptation to a single distribution [77,352,397,405]. Broad subject-transfer benchmarks probe stability across disciplines, difficulty levels, and professional areas while holding the evaluation protocol fixed [96,352,405].

5.2.3. Efficiency

Efficiency asks whether a system can achieve high-quality results while consuming minimal computational resources, latency, or external dependencies. [343,347,348,376,380]. We categorize efficiency evaluation into two aspects: cost–quality tradeoffs and budget-aware optimality.

Cost–Quality Tradeoffs

Cost quality tradeoffs measure how much compute, time, or external resource is required to reach a given level of correctness and reliability, making evaluation more meaningful for deployment under practical constraints [343,347,348,376,380]. Benchmarks usually operationalize this trade-off in three ways: *reporting quality alongside explicit cost signals*, *measuring latency and throughput, or energy under standardized serving settings*, and *tracing quality across token or query budgets* [343,347,348,376,380].

Latency-score ties quality to end-to-end response time, establishing a direct relationship between performance and inference speed. This is especially relevant in interactive settings where long deliberation reduces usability [347,376,377,406]. Latency-oriented benchmarks therefore standardize measures such as time to first token and total completion time under streaming or batch settings, then relate them to task quality [347,376,377,382,406]. Tool-call-score links quality to external tool usage, such as the number of calls or total call cost, which is critical in agentic and open-book settings where retrieval and execution dominate expense [46,47,222,244,378]. Accordingly, tool-augmented benchmarks score task success together with tool-use efficiency; examples range from controlled API environments to realistic interactive systems with auxiliary tools [46,47,222,244,378].

Budget-aware Optimality

Budget-aware optimality evaluates whether a system achieves the best possible quality under explicit resource constraints rather than relying on unconstrained computation [47,348,380,381,406]. These benchmarks impose fixed token, step, or query budgets and evaluate systems under matched constraints, employing token-economy tests, controllable synthetic workloads, and agent tasks characterized by step caps and human reference trajectories.

Early stopping evaluation tests whether a system can terminate once it has gathered sufficient evidence, which is especially important in long-horizon tasks where extra steps increase both cost and error risk [222,379]. Typical protocols evaluate adaptive stopping in decoding or interaction, ending when agreement stabilizes or when further reasoning is unlikely to improve outcomes within the remaining budget.

Multi-objective evaluation measures performance across simultaneous goals such as correctness, reliability, and cost, making it well suited to agentic settings where gains on one objective may degrade another [343,347,377,382]. These benchmarks typically report a metric vector and may use Pareto-style selection over accuracy, reliability, latency, and energy, covering holistic evaluation suites, system benchmarks, and energy-centered tests.

 doi:10.20944/preprints202605.0806.v1

Key Incomparability: Evaluation Dimensions

- ***Why the distinction matters***: evaluation dimensions determine whether a benchmark prioritizes task success alone, dependable behavior, or resource-efficient performance.
- *Correctness*: benchmarked as whether outputs satisfy task-defined success criteria.
- *Reliability*: benchmarked as whether outputs remain verifiable, safe, private, and robust.
- *Efficiency*: benchmarked as how much time, compute, or external resources are required to achieve a given level of performance.

6. Scenario-Based Extensions of Reasoning Benchmarks Analysis

This section conducts a scenario-based extension of reasoning benchmarks. As shown in Figure 9, rather than proposing a second analytic methodology beyond Object, Setting, and Evaluation, it organizes **compact discussions around four recurrent deployment contexts** that increasingly shape benchmark design in practice: **Multilingual and Cultural Benchmarks**, **Multimodal Benchmarks**, **Vertical-Domain Benchmarks**, and **Agentic and Interactive Benchmarks**. Here, scenario refers to **deployment-oriented context**, including *language coverage, modality, work domain, and interaction environment*. These extensions complement Section 3– 5 by reviewing recent advancements in these widely concerned scenarios, and demonstrating **how the structure of these benchmarks can be interpreted within the framework introduced above**.

6.1. Multilingual and Cultural Benchmarks

Along the language dimension, benchmark development has progressively expanded along two closely related yet analytically distinct directions: **cultural benchmarks** and **multilingual benchmarks** [407–410].

6.1.1. Cultural Benchmarks

Cultural benchmarks are grounded in local entities, institutions, conventions, and social norms, particularly in domains where linguistic form, background knowledge, and culturally situated reasoning are tightly coupled [352,408,410–412]. Early work was relatively sparse and domain-specific, demonstrating how evaluation could be adapted to specialized local settings [413,414]. This line later expanded through exam- and knowledge-centered benchmarks, initially concentrated in Chinese and other regional settings, and gradually broadened to more languages, regions, dialects, and low-resource contexts [55,67,415–426]. This expansion improved local realism and relevance, but it also made strict cross-lingual comparison harder, as benchmark content became increasingly tied to specific cultural, educational, and sociolinguistic contexts.

 doi:10.20944/preprints202605.0806.v1

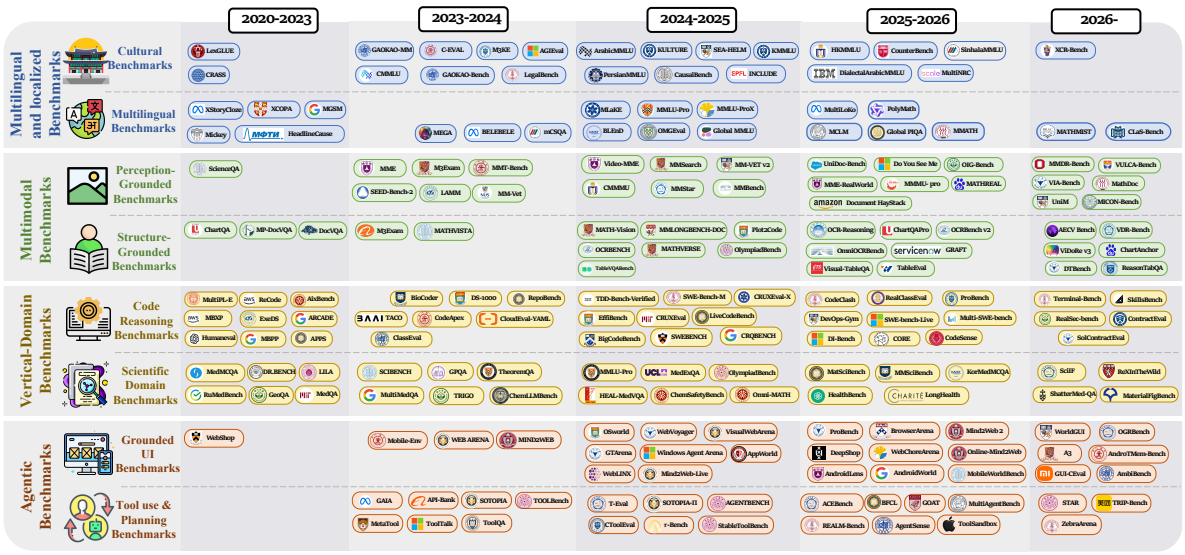

Figure 9. Scenario-Based Extensions of Reasoning Benchmarks: A Developmental Timeline. Spanning four benchmark families (Multilingual, Multimodal, Vertical-Domain, and Agentic) across five stages, this framework maps progression from static tasks to dynamic, tool-augmented, and real-world evaluation.

6.1.2. Multilingual Benchmarks

Multilingual benchmark construction initially remained largely translation-based, with design priorities centered on prompt matching, answer-space alignment, and item-level comparability across languages [64,427–430]. Although this improved cross-lingual symmetry, it also introduced translation artifacts, culturally flattened prompts, and continued dependence on English-centered task formulations [407,409]. Subsequent work moved beyond fully parallel designs to encompass mathematical and causal reasoning, shared-item resources, culturally grounded datasets, and expert-verified multilingual extensions [408,428,431–435]. From there, multilingual benchmarks further broadened toward locally grounded knowledge, culturally situated commonsense, cross-lingual alignment, and more interactive settings involving multi-turn interaction, retrieval, tool use, and agentic capabilities [401,409,410,436–443]. As a result, coverage expanded from language inclusion alone to language-conditioned task settings, though benchmark comparability became more fragile once full parallelism and difficulty control were harder to maintain across languages [316,407,410,443].

6.2. Multimodal Benchmarks

Multimodal benchmarks can be organized along two parallel lines: **perception-grounded benchmarks** that target broad coverage across diverse scenes, documents, and videos, and **structure-grounded benchmarks** that focus on inputs, tables, charts, diagrams, and forms, where relational structure is explicit and therefore more amenable to fine-grained verification [167,320,444–447].

6.2.1. Perception-Grounded Benchmarks

Perception-grounded benchmarks have gradually moved from relatively clean image–text inference settings toward document-heavy, high-resolution, and real-world scenarios, and more recently toward richer settings involving tool use, multi-step evidence access, long-context inputs, and interactive constraints [62,244,320,444,445,448–450]. The central evaluation object is evidence-bound inference after perceptual extraction, assessed through answer accuracy, grounding fidelity, robustness to perturbation, and calibration under uncertainty [320,450,451]. A key diagnostic challenge in this branch is separating failures of perception from failures of reasoning under explicit evidence conditions [450,452].

 doi:10.20944/preprints202605.0806.v1

6.2.2. Structure-Grounded Benchmarks

Structure-grounded benchmarks follow a different logic: inputs are not only multimodal but also structurally organized in ways that expose relations, quantities, alignments, and dependencies, enabling constraint-aware interpretation and finer-grained verification [167,445,453]. The evaluation object shifts accordingly to structural interpretation coupled with reasoning over explicit constraints, and metrics often extend beyond final-answer correctness to include quantitative tolerance, format or unit consistency, and agreement with intermediate reasoning states where annotations permit [167,452,453]. This distinction between the two benchmark families is best understood as an analytic framework rather than a fixed taxonomy; in practice, benchmarks that claim to measure process-level competence should specify explicit process-supervision schemes so that intermediate correctness is observable and testable rather than left implicit [444,448,452].

6.3. Vertical-Domain Benchmarks

Vertical-domain benchmarks fall into two parallel lines: **code reasoning benchmarks**, which center on programming artifacts, execution environments, and repository-scale intervention, and **scientific domain benchmarks**, which center on scientific evidence, expert knowledge, multimodal observations, and constraint-sensitive reasoning. Across both lines, benchmark development has moved away from thin answer-only tasks toward richer evaluation objects, more explicit settings, and broader criteria [343,373,454].

6.3.1. Code Reasoning Benchmarks

Code reasoning benchmarks did not evolve simply by making programming questions harder. Early work, including HumanEval, MBPP, and APPS, evaluated function synthesis and short-form code generation; subsequent benchmarks such as MultiPL-E, MBXP, ReCode, and ARCADE broadened language coverage, execution conditions, and task formats without fully leaving snippet-scale evaluation behind [75,261,272,455–460].

The evaluated object grew richer once coding was embedded in more specialized and structured contexts. BioCoder and DS-1000 introduced biomedical and data-science programming, while RepoBench, ClassEval, and CloudEval-YAML brought in repository evidence, class-level structure, and configuration-sensitive tasks, making evaluation less reducible to isolated snippets [24,262,270,461–464]. Later benchmarks sharpened this further by emphasizing verified development, efficiency, repository-centered repair, and live development workflows [24,80,282,302,304,465–475]. Terminal-Bench and SkillsBench extend evaluation to longer tool-mediated trajectories, while RealSec-Bench, ContractEval, and SolContractEval show that in security-sensitive and contract-specific settings, ordinary completion metrics are no longer sufficient [328,476–479].

6.3.2. Scientific Domain Benchmarks

Scientific domain benchmarks followed a parallel but distinct trajectory. Early benchmarks, including DisKnE, MedQA, GeoQA, MedMCQA, and others, framed scientific competence primarily as question answering or short-form expert response, expanding topical coverage across subdomains, languages, and regional contexts while keeping the dominant format close to expert QA [66,150,319,480–483].

Subsequent benchmarks introduced stronger reasoning demands and richer evidential structure. SCIBENCH, GPQA, and TheoremQA raised the difficulty bar through scientific problem solving, graduate-level expert questioning, and theorem-centered reasoning, while MultiMedQA, TRIGO, and ChemLLMBench widened the evidential base through multi-source medical evidence, mathematically structured tasks, and chemistry-specific judgment [34,54,170,171,484,485]. More recent work diversified the field further, extending evaluation to medically grounded visual evidence, safety-sensitive chemical reasoning, materials science, longitudinal health contexts, and figure-grounded scientific understanding [68,156,157,486–497]. Across these developments, the central shift is that domain benchmarks no longer test subject knowledge alone, but increasingly ask whether reasoning remains valid under domain-specific evidence, constraints, risks, and verification standards.

 doi:10.20944/preprints202605.0806.v1

6.4. Agentic and Interactive Benchmarks

Agentic and interactive benchmarks mark a shift from answer-only evaluation toward explicit action trajectories, where benchmark conclusions depend on how interaction is grounded, what external operations are permitted, and how multi-step behavior is assessed. This subsection covers two closely related lines: **grounded UI interaction**, which requires models to perceive and act within changing interfaces, and **tool use with multi-step planning**, which requires models to select, invoke, and sequence external operations toward a longer-horizon goal.

6.4.1. Grounded UI Interaction

Grounded UI interaction benchmarks require models not only to follow a fixed tool schema, but also to perceive a changing interface, connect visible elements to the user's goal, and act through clicking, typing, scrolling, or navigation. Early environments such as WebShop made interface actions central to task completion; subsequent benchmarks extended this toward richer interfaces, longer trajectories, and more realistic web and mobile interaction traces [209,222,229,313,498]. This line later broadened toward open-web interaction, stronger visual grounding, live environments, and workflow-shaped objectives, where success increasingly depends on staying grounded across multiple interface states rather than solving a single isolated step [223,224,226,228,470,499–502].

A parallel branch extends grounded interaction beyond the browser to desktop, app, and mobile environments with longer horizons, evolving state, and stronger memory demands [237,244,248,340, 503–505]. Further benchmarks widen the space through more heterogeneous interface settings and evaluation goals [247,506–509]. Across these settings, the object is no longer a single correct output but grounded action under a perceptual state, making interface versioning, environment control, and trajectory logging increasingly important for meaningful comparison.

6.4.2. Tool Use with Multi-Step Planning

Early broad evaluations offered wide behavioral coverage, but most tasks still terminated at a final answer rather than an executable trajectory, leaving tool choice, intermediate decisions, and recovery after failure largely invisible [70,343]. Once external action interfaces became part of the model stack, benchmark design shifted toward making action itself observable. GAIA framed real-world problem solving in ways requiring decomposition and external operations; API-Bank turned structured API invocation into the benchmark object; MetaTool, ToolTalk, and ToolQA scored the full sequence of tool-mediated decisions rather than only the final response; and ToolBench enlarged the space further by increasing tool diversity and making tool selection a substantial part of the task [72,249,256,301,510,511].

As this line matured, benchmark design became more differentiated. Some benchmarks focused on tool-call quality, function-calling reliability, and reproducibility [25,46,257,512]. Others shifted toward execution under more realistic constraints, including transactional long-horizon interaction, richer traces, and stronger environment grounding [47,53,253,287,513]. At the same time, the benchmark object widened beyond single-agent tool invocation to include social interaction, strategic behavior, coordination, and multi-agent collaboration, reinforcing trajectory-level evaluation as the dominant assessment form in this area [31,78,231,288,399,514,515].

Across both lines, agentic and interactive benchmarks are best understood not as a simple extension of answer-only testing, but as settings in which object, protocol, and metric become tightly coupled, where changes in interface state, tool schema, execution constraints, or trajectory scoring can all materially alter what a reported score means.

 doi:10.20944/preprints202605.0806.v1

Takeaway of Scenario-Based Extensions of Reasoning Benchmarks Analysis

- **Scenario expansion:** Reasoning benchmarks are increasingly extended along deployment-oriented scenarios, especially multilingual/cultural, multimodal, vertical-domain, and agentic/interactive scenarios.
- **Evaluation shift:** Benchmark design is moving from static, answer-only tasks toward more context-sensitive, structured, and realistic forms of assessment.
- **Broader capability coverage:** Evaluation now goes beyond final-task correctness to include grounding, external tool use, domain constraints, interaction, and process-level behavior.
- **Comparability challenge:** As benchmarks become more scenario-aware, cross-benchmark comparability becomes harder to preserve, making object, setting, and evaluation protocol more tightly coupled.

7. Current Threats to Benchmark Comparability

The preceding analysis suggests that benchmark comparability cannot be determined from benchmark names alone. Once a benchmark is decomposed into its **Object, Setting, and Evaluation**, it becomes easier to see why superficially similar results are often not directly comparable. Benchmarks that are all described as evaluating reasoning may differ in the capability they actually target, the conditions under which that capability is elicited, and the metric semantics used to define success.

Accordingly, we identify current threats to benchmark comparability in two broad categories: excessive heterogeneity of benchmark and insufficiently scientific benchmark settings.

7.1. Excessive heterogeneity of benchmarks

Heterogeneity from Temporal and Environmental Drift

The first source of heterogeneity is **temporal drift**. This problem is especially salient in tasks grounded in live knowledge, recent events, web content, or competitive programming streams. In such settings, the target world changes faster than benchmark publication and maintenance cycles, so a benchmark may become partially outdated even when its original design was sound. Under these conditions, lower model performance should not be interpreted too quickly as weaker reasoning. Instead, it may indicate that the benchmark gold label, supporting evidence, or assumed world state no longer matches the current world. A second source of heterogeneity is **environmental drift**. If temporal drift changes the world that a benchmark refers to, environmental drift changes the conditions under which that benchmark is executed. This issue is especially serious in interactive environments and software-based tasks, where even small changes in dependencies, interfaces, operating systems, packages, or execution settings can alter which action paths are available to a model and which outcomes count as success. Consequently, two evaluations conducted under the same benchmark name may still produce different outcomes because the effective environment is no longer the same [80,222,244].

Heterogeneity from Reporting

A further source is **reporting heterogeneity**. Even when the target world and execution environment remain relatively stable, comparisons can still break down if studies differ in how the benchmark is operationalized and reported. In practice, papers often use the same broad benchmark label while differing in stopping rules, retry policies, budget limits, parser assumptions, retrieval settings, tool permissions, or evaluation filters. At first glance, these choices may seem secondary. In fact, they shape what the model is allowed to observe, what actions it may take, how long it may continue, and how its output is interpreted. For that reason, they are not incidental implementation details but part of the benchmark itself. Once such differences are left implicit, benchmark comparisons can become misleading, and reported rankings may shift without reflecting a meaningful change in underlying reasoning ability [271,310,516].

 doi:10.20944/preprints202605.0806.v1

Taken together, temporal drift, environmental drift, and reporting heterogeneity do more than introduce noise into evaluation. More importantly, they show why benchmark scores are difficult to interpret unless object, setting, and evaluation remain aligned. A reported score is scientifically meaningful only when the benchmark object is stable, the evaluation setting is sufficiently controlled, and the metric preserves a consistent notion of success. Once any of these conditions breaks, the score no longer reflects model capability alone; it also reflects changes in benchmark configuration. This is precisely why benchmark names are often poor comparison units.

7.2. Insufficiently scientific benchmark settings

The problem becomes even more serious when evaluation is **diagnostically weak**. Many reasoning benchmarks still summarize performance with a single end-to-end score. Such a score is convenient for ranking, but it is often too coarse for interpretation, because it does not reveal whether failure arose from retrieval, reasoning, grounding, attribution, action selection, or evidence selection.

This limitation is especially serious in evidence-intensive tasks, where a system may fail for different reasons even when the final answer is equally wrong. A benchmark that separates **answer quality** from **support quality** therefore provides clearer diagnostic value than one reporting final correctness alone, since it distinguishes failure in producing an answer from failure in establishing valid support for that answer [360,383,517,518]. For these reasons, current reasoning benchmarks should not be compared at the level of benchmark names alone. What must be compared instead is the full benchmark configuration: the capability under test, the conditions under which it is elicited, and the semantics of the metric used to score it. Only under that stronger notion of alignment can benchmark results function as reliable evidence rather than as superficially comparable numbers.

Takeaway of Current Threats to Benchmark Comparability

- **Excessive heterogeneity:** Temporal drift, environmental drift, and reporting heterogeneity make benchmark scores unstable and reduce direct comparability across studies.
- **Insufficiently scientific settings:** Coarse end-to-end evaluation provides limited diagnostic value, making it difficult to interpret whether benchmark results truly reflect reasoning ability.

8. Practical Guidelines

Having established that benchmark results are meaningful only insofar as they provide decision-relevant evidence about model behavior, the discussion must now move from evaluative use to evaluative design. This shift is necessary because the evidential value of a benchmark is not determined at the moment of score interpretation alone, but is already structured by the way the benchmark is chosen and built. In other words, benchmark selection and benchmark construction are not separable stages, but two aspects of the same scientific problem: the former asks what kind of evidence is needed for a deployment claim, while the latter asks how such evidence can be generated in a valid, interpretable, and reproducible form. As shown in Figure 10, we therefore address these two questions in sequence, first by examining **how to choose an appropriate benchmark (Benchmark Usage)**, and then by specifying **how to build a scientifically sound benchmark (Benchmark Construction)**.

8.1. How to Choose an Appropriate Benchmark

Selecting an appropriate benchmark is fundamentally an exercise in matching evaluation evidence to deployment need, not in following the most visible leaderboard. The primary question is not which benchmark is popular or convenient to run, but which deployment failure the benchmark is designed to rule out [20,36]. A benchmark is appropriate only when its task content, operational setting, and score semantics expose that failure directly, rather than through a nearby proxy [20,519]. In this sense, benchmark selection is neither a matter of convenience nor convention, but of whether the chosen

 doi:10.20944/preprints202605.0806.v1

benchmark can furnish evidence genuinely relevant to the claim being made about model behavior in deployment.

Benchmark selection should therefore begin with the core capabilities. Practitioners should ask which dependency structure governs success in the target use case–whether that involves symbolic derivation, quantitative constraint tracking, long-context evidence aggregation, grounded tool use, or long-horizon action coherence [18,47,50,71]. A benchmark that targets the wrong dependency structure may still yield strong scores, yet those scores offer little practical reassurance about deployment behavior [20,37,520]. **The operational setting should then be treated as constitutive of benchmark identity rather than as an ancillary implementation detail**. Tool permissions, retrieval access, context length, interaction horizon, budget constraints, and environment dynamics all shape what kind of competence is actually being measured [45,47,222,340,521]. Two benchmarks bearing similar task labels may thus yield very different evidence once these conditions diverge [47,53,71,79]. What matters is not surface-level categorical similarity, but whether the benchmark faithfully reproduces the operative conditions.

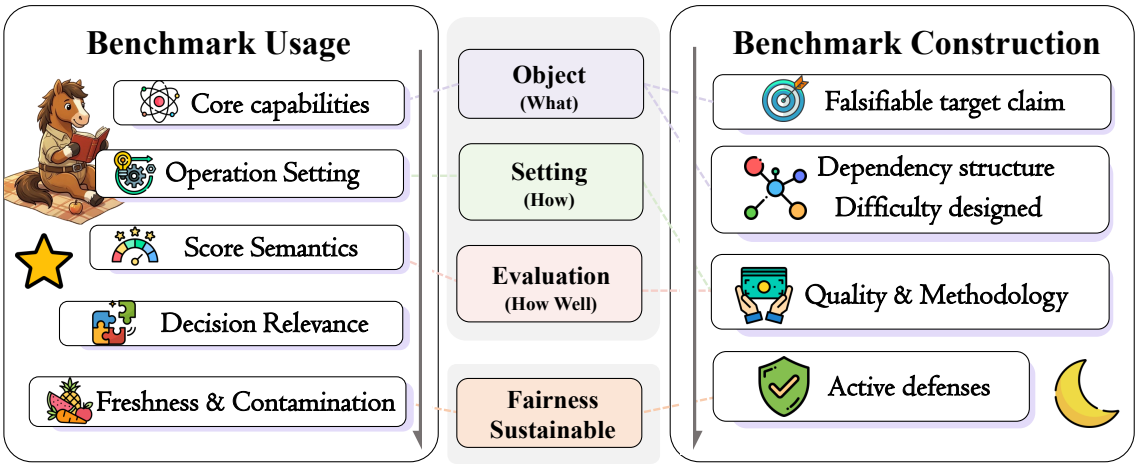

Figure 10. A practical, stepwise guideline flow for benchmark selection and construction, structured around two interdependent pillars: Benchmark Usage and Benchmark Construction–unified through four cross-cutting axes: Object, Setting, Evaluation, and Fairness.

Score semantics must likewise be grounded in decision relevance. A benchmark proves more useful when its reported outcomes align with the downstream cost profile of failure [20,343,519]. In some use cases, final answer correctness is the overriding concern; in others, evidence support, robustness under perturbation, or resource efficiency are equally consequential. Benchmark choice should also reflect this asymmetry rather than assume that a single headline figure suffices for all model-selection decisions [343,345,360,366,521]. This consideration explains why coverage and realism must be balanced deliberately: broad benchmarks are valuable for initial screening, but they simplify operational conditions and compress heterogeneous failure modes into aggregate scores [32,343]. Narrower benchmarks may span fewer tasks, yet provide stronger evidential weight in high-stakes domains because they preserve more of the true operational structure [222,238,340]. The right benchmark is therefore not simply the largest, the most difficult, or the one with the most recognized leaderboard; it is the one whose reasoning object, setting, and score semantics best match the deployment claim under scrutiny [20,519].

Freshness and contamination risk deserve explicit attention before any benchmark is adopted as decision evidence. Static benchmarks can remain valuable as historical baselines, but they grow less informative when models have likely encountered overlapping data distributions, or when the environment the benchmark represents has materially changed [271,310,454,516]. Practitioners should therefore treat benchmark age, refresh policy, and contamination control as first-order selection criteria rather than afterthoughts [271,310,522]. A well-chosen benchmark ultimately supports a bounded claim: it tells the user what kind of reasoning behavior has been tested, under which conditions, and

 doi:10.20944/preprints202605.0806.v1

with what practical relevance [20,519]. When such linkages are explicitly established, benchmark outcomes constitute meaningful, interpretable evidence; in their absence, benchmark comparisons reduce to little more than superficial score ornamentation [20,36].

8.2. How to Build a Scientifically Sound Benchmark

A scientifically sound benchmark begins with a **falsifiable target claim**: Benchmark construction should start from a concrete capability question or deployment risk, then define evidence of success or failure [20,36,519]. This anchors the benchmark in a theoretically grounded reasoning object, thereby preventing the drift toward tasks that are merely convenient to collect yet fail to provide meaningful diagnostic insight into model capabilities [20,36].

Task design must satisfy two conditions. First, it should preserve the **dependency structure** that the benchmark intends to measure. If the goal is to test multi-step reasoning, success should depend on maintaining valid intermediate dependencies rather than exploiting shallow cues or annotation artifacts [41,114,520]. If the goal is to test grounded decision making, the benchmark should require state tracking, constraint satisfaction, or evidence integration that cannot be bypassed by surface matching alone [53,222,229,340]. Then, control the **difficulty designed**: benchmark builders must distinguish genuine reasoning difficulty from accidental difficulty caused by ambiguous phrasing, unstable grading, noisy annotation, or missing context, since a good benchmark is challenging for principled reasons and does not rely on confusion as a substitute for rigor [36,523,524].

Meanwhile, benchmark construction should also account for **both the quality of the underlying data and the methodology used** to evaluate model behavior. The former requires explicit **specification of data provenance and protocol formulation**, since these elements shape the observable behavior being measured [20,53,257,273,510]. The latter should likewise preserve structure rather than collapse it prematurely: benchmark builders should **report supported evaluation units and dimensions**, since metric panels are more scientifically informative than a single aggregate score and reveal trade-offs that would otherwise remain hidden [20,41,343,345,360,366,519,521].

Scientific benchmark construction further requires **active defenses** against misleading progress claims and a commitment to reproducibility and interpretive honesty. **Contamination checks, controlled perturbations, freshness management, and version governance should be built into the benchmark design** from the beginning, since without these safeguards score gains may reflect memorization, formatting sensitivity, or stale distributions rather than transferable capability improvement [162,271,310,373,454,516,525]. A reusable benchmark package should include immutable data snapshots, prompt templates, parsing and matching code, runtime specifications, and reference scripts reproducing published aggregates, with environment-based benchmarks additionally requiring deterministic logging and artifact storage [53,222,237,257,273,340,519,526]. Finally, benchmark builders should explicitly state the interpretation boundary: a strong benchmark specifies what it measures, what it omits, and why the chosen evidence is sufficient for the target use case, which makes results cumulative across papers, labs, and model generations [20,519,527].

Takeaway of Practical Guidelines

- **How to Choose an Appropriate Benchmark:** Benchmark selection should be driven by deployment relevance, requiring alignment between the benchmark's reasoning object, operational setting, evaluation semantics, and the concrete failure modes of the target use case.
- **How to Build a Scientifically Sound Benchmark:** Benchmark construction should start from a falsifiable target claim and ensure dependency-preserving task design, controlled difficulty, transparent protocols, multidimensional evaluation, and reproducible implementation.

 doi:10.20944/preprints202605.0806.v1

9. Future Directions

As shown in Figure 11, the future of reasoning benchmarks will likely be shaped by six closely connected directions: **integrated measurement**, **temporal and verifiable evaluation**, **comprehensive agentic scenarios**, **embodied scenarios**, **benchmark lifecycle stewardship**, and **benchmark ecosystem**. Taken together, these directions point toward a benchmark paradigm that is more firmly grounded in scientific principles, more reflective of real-world deployment conditions, and more sustainable over time.

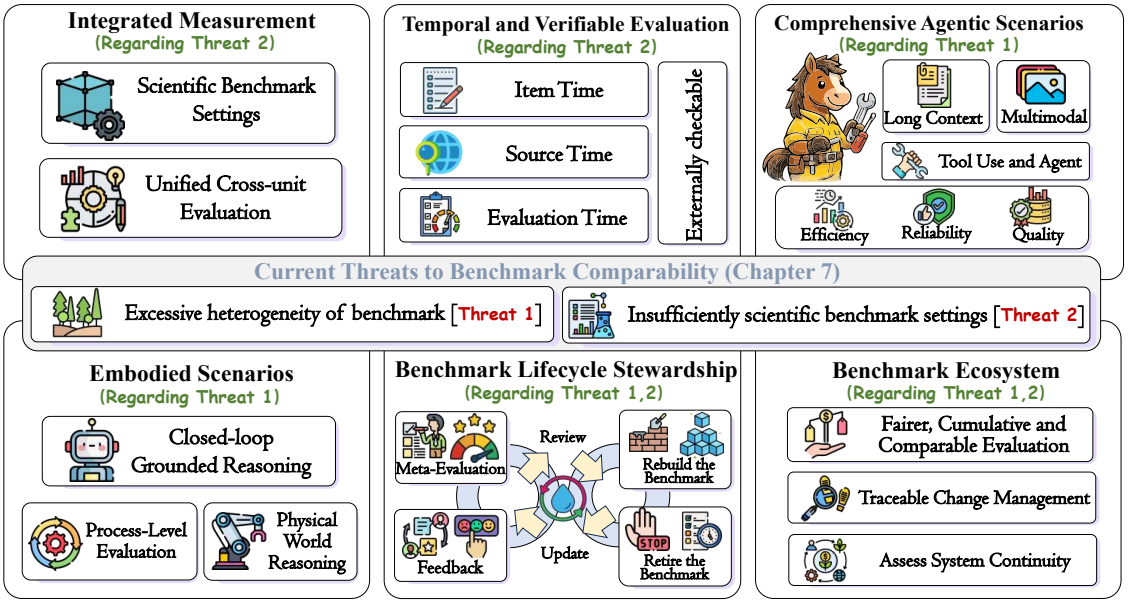

Figure 11. Future Directions for Reasoning Benchmarks. Proposed directions–integrated measurement frameworks, temporally verifiable evaluation, comprehensive agentic scenarios and embodied scenarios–respond to threats of benchmark heterogeneity and unscientific settings (Chapter 7), with governance for lifecycle stewardship and benchmark ecosystem.

9.1. Integrated Measurement

The next phase of reasoning benchmark research will be **defined less by larger test collections and more by stronger measurement architectures**. Current progress has revealed a persistent gap between benchmark visibility and benchmark comparability [528–530]. New suites appear quickly, yet their results often remain difficult to align because they encode different assumptions due to **insufficiently justified benchmark settings**. Closing this gap requires infrastructure that treats protocol semantics as a primary research focus under the same reasoning object, not as appendix material [271,310,343].

A fundamental necessity is the establishment of unified cross-unit evaluation, as answer-level, process-level, and trajectory-level assessments are still often built in parallel and reported in isolation [32,41,531,532]. **In this way, a promising direction is to design benchmarks in which these units are connected by scientific benchmark settings**, so a final score can be traced back to intermediate reasoning validity and policy level behavior. Such designs would allow researchers to distinguish genuine reasoning improvements from compensatory heuristics that only work at one observational layer [41,222,244].

9.2. Temporal and Verifiable Evaluation

As reasoning systems are increasingly deployed in environments where facts, sources, and interfaces change over time, **temporal grounding should become a standard benchmark dimension rather than an optional add-on** [388,516,533]. Static evaluation is no longer sufficient for measuring performance in settings where models must reason over evolving knowledge, retrieve up-to-date evidence, and produce claims that remain checkable after the fact [388,534].

 doi:10.20944/preprints202605.0806.v1

Therefore, a promising direction is to **build temporally explicit and verifiable benchmarks that represent time at multiple levels, including item time, source time, and evaluation time**, while also attaching answers to auditable evidence or executable verification procedures. Such benchmarks would make it possible to evaluate whether a system can distinguish durable knowledge from time-sensitive knowledge, update its conclusions when evidence changes, and support its outputs with **claims that remain externally checkable**. They would also encourage benchmark designs that age more gracefully through refreshable data pipelines, objective verification, and contamination-resistant update cycles. The significance of this shift is twofold. Methodologically, it would allow the field to **separate durable reasoning competence from freshness-dependent performance**, which would produce cleaner comparisons across models and evaluation rounds. Practically, it would **make reported progress more trustworthy by reducing conflation between better reasoning, broader retrieval coverage, and more recent data exposure**, while also moving benchmark design toward more auditable and realistic evaluation for dynamic knowledge environments.

9.3. Comprehensive Agentic Scenarios

Agentic reasoning evaluation opens a closely related frontier where safety, efficiency, and reliability must be co-measured rather than treated as separate leaderboards. In long horizon environments, a system can reach success while spending excessive resources, issuing risky actions, or relying on unstable recovery behavior. Future benchmarks should encode these tradeoffs directly in their reporting interfaces, allowing deployment teams to choose operating points with explicit risk awareness. This direction is especially urgent for real tool-based ecosystems, where small action errors can have irreversible downstream impacts [47,222,244,363].

A further opportunity emerges in long context and multimodal integration [221,535]. Many current suites still evaluate long context retrieval, multimodal grounding, and agentic planning in separate silos, while real deployments increasingly combine all three [77,216]. Future benchmark design should therefore move toward coupled tasks where systems must retrieve sparse evidence in long inputs, ground decisions in heterogeneous modalities, and execute coherent actions under bounded budgets. This coupling will likely reduce headline scores in the short term, but it will produce a more faithful picture of system readiness [44,45,79,244], a trend already visible in long-form or domain-specific settings and web or medical agent evaluations [192,194–196,205,227,228,235–238,536].

9.4. Embodied Scenarios

Embodied scenarios have become an important direction for next-generation reasoning benchmarks, as reasoning in the physical world is not only about **producing plausible textual plans, but also about grounding decisions in perception, affordances, and action consequences**. Recent embodied reasoning benchmarks have gradually moved from instruction following in simulated household environments to more realistic, long-horizon, and interaction-intensive settings [537–539]. Correspondingly, current methods mainly follow two technical routes. The first route is a modular planner–executor paradigm, where LLMs are used for high-level task decomposition, while external perception modules, value functions, or skill libraries are responsible for grounding and execution, as exemplified by SayCan [540]. The second route is end-to-end multimodal action modeling, which directly integrates visual observations, language, and robot trajectories into a unified model for action prediction, represented by PaLM-E and RT-2 [541,542]. Together, these lines of work suggest that embodied reasoning should be treated as a closed-loop process of observing, planning, acting, and revising, rather than as a static text-only reasoning problem.

Despite this progress, current embodied reasoning methods still face two major limitations. First, many existing settings **remain overly dependent on simulation priors, fixed skill vocabularies, or narrow action interfaces**, which means that strong benchmark performance does not necessarily indicate robust physical reasoning [539,543,544]. Second, current formulation protocols often emphasize final task success while under-evaluating process-level reasoning abilities, such as active exploration under partial observability, uncertainty awareness, task tracking, and recovery from execution er-

 doi:10.20944/preprints202605.0806.v1

rors [211,545]. Therefore, the key future direction is to design embodied reasoning benchmarks that explicitly evaluate *closed-loop grounded reasoning*: capabilities should be tested on whether they can update their beliefs from new observations, choose actions based on affordances and safety constraints, re-plan after failures, and coordinate with humans or agents in dynamic environments [211,544,545].

9.5. Benchmark Lifecycle Stewardship

Benchmark governance should move beyond a paper-centric release model toward full lifecycle stewardship. As reasoning continues to evolve in different forms and with shifting emphases, evaluation cannot rely on static benchmarks released once and then left unmanaged. Instead, benchmarks should be organized around a more controlled and comparable structure across three core dimensions: object, setting, and metrics. That is, the community should be explicit about what capability is being evaluated, under what conditions it is evaluated, and by what criteria performance is judged. Only with such structure can benchmark results remain interpretable and comparable across generations of tasks, models, and evaluation regimes.

Benchmark development should be treated as a lifecycle rather than a one-time artifact. **The process should begin with careful design**, including explicit construct definitions, scope decisions, contamination safeguards, and documentation of intended use [519,546]. **It should then be followed by meta-evaluation**: assessing not only model performance on the benchmark, but also the benchmark's own validity, sensitivity, robustness, and patterns of actual use within the community [547,548]. These signals should, in turn, inform **whether a benchmark should be incrementally updated, substantially rebuilt, or formally retired** [549,550]. Versioned artifacts, transparent patch logs, deprecation policies, and community-readable schema mappings are therefore not auxiliary governance features, but core infrastructure for maintaining continuity across benchmark generations.

9.6. Benchmark Ecosystem

The broader value of such a lifecycle is **to create an evaluative ecosystem that is fairer, more cumulative, and more comparable**. Without it, the field risks mistaking benchmark churn for scientific progress, while allowing hidden shifts in task design, settings, or metrics to undermine comparison. With it, reasoning evaluation can develop in a more orderly way: new benchmarks and new results refine shared instruments, preserve continuity where possible, and introduce change in a traceable form when necessary. In this sense, lifecycle stewardship is not merely about benchmark maintenance; it is a condition for building a healthier evaluative culture and for supporting the disciplined growth of the community [271,310,343].

 doi:10.20944/preprints202605.0806.v1

Takeaway of Future Directions

- **Integrated Measurement:** Future reasoning benchmarks should prioritize stronger measurement architectures that unify evaluation units under scientifically aligned settings.
- **Temporal and Verifiable Evaluation:** Benchmark design should make temporal grounding and verifiability explicit, so that evaluation can distinguish durable reasoning competence.
- **Comprehensive Agentic Scenarios:** Agentic benchmarks should move toward coupled evaluation of safety, efficiency, reliability, long-context reasoning, multimodal grounding, and bounded-budget action in realistic tool-mediated environments.
- **Embodied Scenarios:** Embodied reasoning benchmarks should transition from measuring static task success to evaluating closed-loop, grounded reasoning, placing emphasis on perception, affordance reasoning, adaptive replanning, and recovery within dynamic physical environments.
- **Benchmark Lifecycle Stewardship:** Benchmarks should be managed as evolving scientific instruments, with explicit design, meta-evaluation, versioning, updating, and retirement policies.
- **Benchmark Ecosystems:** A healthier benchmark ecosystem should support cumulative, fair, and traceable evaluation, so that new benchmarks extend shared measurement infrastructure.

10. Related Work

Related studies can be broadly grouped into two categories: surveys on **reasoning in large language models** and surveys on **benchmarks and evaluation**.

The first line of work focuses on reasoning itself, including its conceptualization, elicitation, and training in large language models. Early survey efforts such as Huang and Chang [551] provide a systematic overview of reasoning in LLMs, with a particular emphasis on informal deductive reasoning. In recent surveys, Sun et al. [552] review reasoning in foundation models from a more comprehensive angle, summarizing tasks, methods, benchmarks, and future directions, while also discussing multimodal reasoning, agents, and alignment-related issues. Plaat et al. [27] concentrate on multi-step reasoning and organize the literature around a generate-evaluate-control framework. Other work further synthesizes the development of reasoning LLMs from the perspective of System 1 and System 2 reasoning, highlighting architectural choices, training strategies, and evaluation practices [7].

The second line of work focuses on benchmarks and evaluation.

A representative example is the survey by Chang et al. [553], which reviews LLM evaluation from the perspectives of what to evaluate, where to evaluate, and how to evaluate. Moving further in this direction, Ni et al. [22] systematically summarize a large collection of LLM benchmarks and categorize them into general, domain-specific, and target-specific benchmarks. At the same time, several studies critically examine the reliability and validity of benchmarks themselves. For instance, Xu et al. [554] survey benchmark data contamination and show that training-test leakage can seriously undermine the interpretability of benchmark scores. In the multimodal setting, Li et al. [444] review benchmarks for multimodal large language models.

Similarly, Yang et al. [555] provide a critical review of causal reasoning benchmarks and argue that many existing datasets remain vulnerable to shallow pattern matching or knowledge retrieval. Taken together, existing benchmark-oriented surveys either cover evaluation too broadly or focus on a specific reasoning subdomain. In contrast, our work centers specifically on reasoning benchmarks, with particular attention to capability coverage, task setting, and the corresponding evaluation.

 doi:10.20944/preprints202605.0806.v1

11. Conclusions

In this work, we propose a comprehensive survey for reaoning benchmarks of large language models. This survey introduces a new perspective for better understanding the broad landscape of reasoning benchmarks, including the object, settings, evaluation as well as the extended scenarios. In addition, we extend our discussion to the current limitations of this area, practical guidelines for practitioners, and potential future directions for advancing reasoning benchmarks. To further support the community, we have also been maintaining an online survey repository that continuously tracks newly released benchmarks under our structured outline. We hope this effort can serve as a useful resource for researchers and practitioners, and contribute to the broader development of the community.

References

1. Zhao W X, Zhou K, Li J, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023, 1: 1–124
2. Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025
3. Song P, Han P, Goodman N. Large language model reasoning failures. arXiv preprint arXiv:2602.06176, 2026
4. Nye M, Andreassen A J, Gur-Ari G, et al. Show your work: Scratchpads for intermediate computation with language models. 2021
5. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 2022, 35: 24824–24837
6. Chen Q, Qin L, Liu J, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567, 2025
7. Li Z Z, Zhang D, Zhang M L, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419, 2025
8. Hao S, Sukhbaatar S, Su D, et al. Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769, 2024
9. Chen X, Zhao A, Xia H, et al. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. arXiv preprint arXiv:2505.16782, 2025
10. Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models. Patterns, 2025, 6
11. Chen Q, Qin L, Zhang J, et al. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 8199–8221
12. Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 2023, 36: 11809–11822
13. Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models. In: Proceedings of The eleventh international conference on learning representations, 2023
14. Zhang Z, Zhang A, Li M, et al. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023
15. Chen Q, Yang M, Qin L, et al. Ai4research: A survey of artificial intelligence for scientific research. arXiv preprint arXiv:2507.01903, 2025
16. Zhang C, Chen Q, Chen X, et al. Less languages, less tokens: An efficient unified logic cross-lingual chain-of-thought reasoning framework. arXiv preprint arXiv:2604.20090, 2026
17. Qin L, Chen Q, Feng X, et al. Large language models meet nlp: A survey. arXiv preprint arXiv:2405.12819, 2024
18. Wang Z, Wu F, Wang H, et al. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in llm agents. arXiv preprint arXiv:2601.22311, 2026
19. Lu C, Chen Z, Zhao H, et al. Lore: A large generative model for search relevance. arXiv preprint arXiv:2512.03025, 2025
20. Bean A M, Kearns R O, Romanou A, et al. Measuring what matters: Construct validity in large language model benchmarks. arXiv preprint arXiv:2511.04703, 2025
21. Kargupta P, Li S S, Wang H, et al. Cognitive foundations for reasoning and their manifestation in llms. arXiv preprint arXiv:2511.16660, 2025

 doi:10.20944/preprints202605.0806.v1

22. Ni S, Chen G, Li S, et al. A survey on large language model benchmarks. arXiv preprint arXiv:2508.15361, 2025
23. Qin C, Chen X, Wang C, et al. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025. 5754–5765
24. Du X, Liu M, Wang K, et al. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. arXiv preprint arXiv:2308.01861, 2023
25. Guo Z, Huang Y, Xiong D. Ctooleval: A chinese benchmark for llm-powered agent evaluation in real-world api interactions. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 15711–15724
26. Lin Z, Gou Z, Liang T, et al. Criticbench: Benchmarking llms for critique-correct reasoning. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 1552–1587
27. Plaat A, Wong A, Verberne S, et al. Multi-step reasoning with large language models, a survey. ACM Computing Surveys, 2025, 58: 1–35
28. Han S, Schoelkopf H, Zhao Y, et al. Folio: Natural language reasoning with first-order logic. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 22017–22031
29. Lian S, Wu C, Yang L T, et al. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. arXiv preprint arXiv:2509.24473, 2025
30. Es S, James J, Anke L E, et al. Ragas: Automated evaluation of retrieval augmented generation. 2024, pages 150–158
31. Shen Y, Huang Z, Wang Z, et al. Trip-bench: A benchmark for long-horizon interactive agents in real-world scenarios. arXiv preprint arXiv:2602.01675, 2026
32. Mohammadi M, Li Y, Lo J, et al. Evaluation and benchmarking of llm agents: A survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025. 6129–6139
33. Wei J, Sun Z, Papay S, et al. Browsecomp: A simple yet challenging benchmark for browsing agents. arXiv preprint arXiv:2504.12516, 2025
34. Rein D, Hou B L, Stickland A C, et al. Gpqa: A graduate-level google-proof q&a benchmark. In: Proceedings of First Conference on Language Modeling, 2024
35. Yang E, Wang D. Benchmark illusion: Disagreement among llms and its scientific consequences. arXiv preprint arXiv:2602.11898, 2026
36. Mousavi S M, Cecchinato E, Hornikova L, et al. Garbage in, reasoning out? why benchmark scores are unreliable and what to do about it. arXiv preprint arXiv:2506.23864, 2025
37. Shojaee P, Mirzadeh I, Alizadeh K, et al. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2506.06941, 2025
38. Elkins K, Chun J. Syntactic framing fragility: An audit of robustness in llm ethical decisions. arXiv preprint arXiv:2601.09724, 2025
39. Jacovi A, Bitton Y, Bohnet B, et al. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 4615–4634
40. Fu Y, Ou L, Chen M, et al. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. arXiv preprint arXiv:2305.17306, 2023
41. Zheng C, Zhang Z, Zhang B, et al. Processbench: Identifying process errors in mathematical reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 1009–1024
42. Tsoukalas G, Lee J, Jennings J, et al. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. Advances in Neural Information Processing Systems, 2024, 37: 11545–11569
43. Mirzadeh I, Alizadeh K, Shahrokhi H, et al. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229, 2024
44. Zhang X, Chen Y, Hu S, et al. ∞bench: Extending long context evaluation beyond 100k tokens. arXiv preprint arXiv:2402.13718, 2024
45. Hsieh C P, Sun S, Kriman S, et al. Ruler: What's the real context size of your long-context language models? arXiv preprint arXiv:2404.06654, 2024
46. Guo Z, Cheng S, Wang H, et al. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 11143–11156

 doi:10.20944/preprints202605.0806.v1

47. Yao S, Shinn N, Razavi P, et al. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045, 2024
48. Song M, Su Z, Qu X, et al. Prmbench: A fine-grained and challenging benchmark for process-level reward models. arXiv preprint arXiv:2501.03124, 2025
49. Li X, Yu H, Zhang X, et al. Socratic-prmbench: Benchmarking process reward models with systematic reasoning patterns. arXiv preprint arXiv:2505.23474, 2025
50. Parmar M, Patel N, Varshney N, et al. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. arXiv preprint arXiv:2404.15522, 2024
51. Wan Y, Wang W, Yang Y, et al. Logicasker: Evaluating and improving the logical reasoning ability of large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 2124–2155
52. Fujisawa I, Nobe S, Seto H, et al. Procbench: Benchmark for multi-step reasoning and following procedure. arXiv preprint arXiv:2410.03117, 2024
53. Lu J, Holleis T, Zhang Y, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 1160–1183
54. Guo T, Nan B, Liang Z, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. Advances in neural information processing systems, 2023, 36: 59662–59688
55. Zong Y, Qiu X. Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. arXiv preprint arXiv:2402.15745, 2024
56. He J, Hu N, Long W, et al. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. arXiv preprint arXiv:2412.17032, 2024
57. Mehri S, Kargupta P, August T, et al. Learning user preferences through interaction for long-term collaboration. arXiv preprint arXiv:2601.02702, 2026
58. Lin B Y, Bras R L, Richardson K, et al. Zebralogic: On the scaling limits of llms for logical reasoning. arXiv preprint arXiv:2502.01100, 2025
59. Phan L, Gatti A, Han Z, et al. Humanity's last exam. arXiv preprint arXiv:2501.14249, 2025
60. Qian Q, Huang C, Xu J, et al. Benchmark^2: Systematic evaluation of llm benchmarks. arXiv preprint arXiv:2601.03986, 2026
61. Wang H, Liu H, Liu X, et al. Fostering video reasoning via next-event prediction. arXiv preprint arXiv:2505.22457, 2025
62. Chen Q, Luan C, Wu J, et al. Omibench: Benchmarking olympiad-level multi-image reasoning in large vision-language model. arXiv preprint arXiv:2604.20806, 2026
63. Gema A P, Leang J O J, Hong G, et al. Are we done with mmlu? arXiv preprint arXiv:2406.04127, 2024
64. Lin X V, Mihaylov T, Artetxe M, et al. Few-shot learning with multilingual generative language models. In: Proceedings of the 2022 conference on empirical methods in natural language processing, 2022. 9019–9052
65. Basu K, Abdelaziz I, Chaudhury S, et al. Api-blend: A comprehensive corpora for training and benchmarking api llms. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 12859–12870
66. Mishra S, Finlayson M, Lu P, et al. Lila: A unified benchmark for mathematical reasoning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. 5807–5832
67. Zhong W, Cui R, Guo Y, et al. Agieval: A human-centric benchmark for evaluating foundation models. In: Proceedings of Findings of the association for computational linguistics, 2024. 2299–2314
68. Wang Y, Ma X, Zhang G, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems, 2024, 37: 95266–95290
69. Kazemi M, Fatemi B, Bansal H, et al. Big-bench extra hard. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 26473–26501
70. Srivastava A, Rastogi A, Rao A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023
71. Bai Y, Tu S, Zhang J, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. 2025, pages 3639–3664
72. Mialon G, Fourrier C, Swift C, et al. Gaia: a benchmark for general ai assistants. arXiv preprint arXiv:2311.12983, 2023
73. Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020

 doi:10.20944/preprints202605.0806.v1

74. Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021
75. Hendrycks D, Basart S, Kadavath S, et al. Measuring coding challenge competence with apps. arXiv preprint arXiv:2105.09938, 2021
76. Guan J, Chen Q, Qin L, et al. Beware of reasoning overconfidence: Pitfalls in the reasoning process for multi-solution tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2026. 30843–30851
77. Yue X, Ni Y, Zhang K, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502, 2023
78. Liu X, Yu H, Zhang H, et al. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688, 2023
79. Bai Y, Lv X, Zhang J, et al. Longbench: A bilingual, multitask benchmark for long context understanding. 2024, pages 3119–3137
80. Jimenez C E, Yang J, Wettig A, et al. Swe-bench: Can language models resolve real-world github issues? arXiv preprint arXiv:2310.06770, 2023
81. Zhong W, Wang S, Tang D, et al. Ar-lsat: Investigating analytical reasoning of text. arXiv preprint arXiv:2104.06598, 2021
82. Welleck S, Liu J, Bras R L, et al. Naturalproofs: Mathematical theorem proving in natural language. arXiv preprint arXiv:2104.01112, 2021
83. Dalvi B, Jansen P, Tafjord O, et al. Explaining answers with entailment trees. arXiv preprint arXiv:2104.08661, 2021
84. Saparov A, He H. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, 2023
85. Tafjord O, Dalvi B, Clark P. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 3621–3634
86. Chen G, Xu W, Zhang H, et al. Finereason: Evaluating and improving llms' deliberate reasoning through reflective puzzle solving. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 6685–6715
87. Waugh J. Pencil puzzle bench: A benchmark for multi-step verifiable reasoning. arXiv preprint arXiv:2603.02119, 2026
88. Luo M, Kumbhar S, Parmar M, et al. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. arXiv preprint arXiv:2310.00836, 2023
89. Patel N, Kulkarni M, Parmar M, et al. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. arXiv preprint arXiv:2406.17169, 2024
90. Gull A, Usman Safder M, Elbadry R, et al. Engchain: A symbolic benchmark for verifiable multi-step reasoning in engineering. arXiv e-prints, 2025, pages arXiv–2511
91. Gao X, Gao Q, Gong R, et al. Dialfred: Dialogue-enabled agents for embodied instruction following. IEEE Robotics and Automation Letters, 2022, 7: 10049–10056
92. Wang R, Jansen P, Côté M A, et al. Scienceworld: Is your agent smarter than a 5th grader? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. 11279–11298
93. Fan L, Wang G, Jiang Y, et al. Minedojo: Building open-ended embodied agents with internet-scale knowledge. Advances in Neural Information Processing Systems, 2022, 35: 18343–18362
94. Ye X, Chen Q, Dillig I, et al. Satlm: Satisfiability-aided language models using declarative prompting. Advances in Neural Information Processing Systems, 2023, 36: 45548–45580
95. Gui J, Liu Y, Cheng J, et al. Logicgame: Benchmarking rule-based reasoning abilities of large language models. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 1474–1491
96. Suzgun M, et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them, 2022
97. Pan A, Williams M A. Context is not comprehension. arXiv preprint arXiv:2506.04907, 2025
98. Qi C, Ma R, Li B, et al. Large language models meet symbolic provers for logical reasoning evaluation. 2025
99. Zhu Q, Huang F, Peng R, et al. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. arXiv preprint arXiv:2502.16906, 2025
100. Liu J, Fan Y, Jiang Z, et al. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. arXiv preprint arXiv:2505.19641, 2025
101. Clark P, Tafjord O, Richardson K. Transformers as soft reasoners over language. arXiv preprint arXiv:2002.05867, 2020

 doi:10.20944/preprints202605.0806.v1

102. Zheng K, Han J M, Polu S. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. arXiv preprint arXiv:2109.00110, 2021
103. Tian J, Li Y, Chen W, et al. Diagnosing the first-order logical reasoning ability through logicnli. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021. 3738–3747
104. Jin Z, Lalwani A, Vaidhya T, et al. Logical fallacy detection. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 7180–7198
105. Ontanon S, Ainslie J, Cvicek V, et al. Logicinference: A new dataset for teaching logical inference to seq2seq models. arXiv preprint arXiv:2203.15099, 2022
106. Lalwani A, Kim T, Chopra L, et al. Autoformalizing natural language to first-order logic: A case study in logical fallacy detection. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 2025. 132–147
107. Ren Z, Shao Z, Song J, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. arXiv preprint arXiv:2504.21801, 2025
108. Han S, Yu A, Shen R, et al. P-folio: Evaluating and improving logical reasoning with abundant human-written reasoning chains. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 16553–16565
109. Sun Y, Saparov A. Language models do not follow occam's razor: A benchmark for inductive and abductive reasoning. arXiv preprint arXiv:2509.03345, 2025
110. Thomas N. Chaosbench-logic: A benchmark for logical and symbolic reasoning on chaotic dynamical systems. arXiv preprint arXiv:2601.01982, 2026
111. Peng Z, Yao Y, Ma K, et al. Criticlean: Critic-guided reinforcement learning for mathematical formalization. arXiv preprint arXiv:2507.06181, 2025
112. Lu P, Gong R, Jiang S, et al. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. arXiv preprint arXiv:2105.04165, 2021
113. Azerbayev Z, Piotrowski B, Schoelkopf H, et al. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. arXiv preprint arXiv:2302.12433, 2023
114. Wang P, Li L, Shao Z, et al. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 9426–9439
115. Lee J, Hwang W. Symba: Symbolic backward chaining for structured natural language reasoning. In: 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), 2025. 2468–2484
116. Song P, Yang K, Anandkumar A. Lean copilot: Large language models as copilots for theorem proving in lean. arXiv preprint arXiv:2404.12534, 2024
117. Lu P, Sheng J, Lyu L, et al. Solving inequality proofs with large language models. arXiv preprint arXiv:2506.07927, 2025
118. Chakraborty M, Pirkelbauer P, Yi Q. Formalspeccpp: A dataset of c++ formal specifications created using llms. In: Proceedings of 2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR). IEEE, 2025. 758–762
119. Yu Z, Peng R, Ding K, et al. Formalmath: Benchmarking formal mathematical reasoning of large language models. arXiv preprint arXiv:2505.02735, 2025
120. Miranda B, Zhou Z, Nie A, et al. Veribench: End-to-end formal verification benchmark for ai code generation in lean 4. In: Proceedings of 2nd AI for Math Workshop@ ICML 2025, 2025
121. Schmitt J, Bérczi G, Dekoninck J, et al. Improofbench: Benchmarking ai on research-level mathematical proof generation. arXiv preprint arXiv:2509.26076, 2025
122. Barkallah S, Daruru S, Miranda B, et al. Veribench-ftp: A formal theorem proving benchmark in lean 4 for code verification. In: Proceedings of The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025
123. Borroto M, Kareem I, Ricca F. Towards automatic composition of asp programs from natural language specifications. arXiv preprint arXiv:2403.04541, 2024
124. Zeng L, Che F, Huang X, et al. Veriequivbench: An equivalence score for ground-truth-free evaluation of formally verifiable code. arXiv preprint arXiv:2510.06296, 2025
125. Cheng E Y, Weber L, Jin T, et al. Sharing state between prompts and programs. arXiv preprint arXiv:2512.14805, 2025

 doi:10.20944/preprints202605.0806.v1

126. Biyani P, Kirtania S, Bajpai Y, et al. Indimathbench: Autoformalizing mathematical reasoning problems with a human touch. arXiv preprint arXiv:2512.00997, 2025
127. Pandit S, Xu A, Nguyen X P, et al. Hard2verify: A step-level verification benchmark for open-ended frontier math. arXiv preprint arXiv:2510.13744, 2025
128. Zhou X, Lei Y, Zhou X, et al. Spark-prover-x1: Formal theorem proving through diverse data training. arXiv preprint arXiv:2511.13043, 2025
129. Zhang X, Zhu N, He Y, et al. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. arXiv preprint arXiv:2310.18021, 2023
130. Ambati M. Proofnet++: A neuro-symbolic system for formal proof verification with self-correction. arXiv preprint arXiv:2505.24230, 2025
131. Xu Q, Luan X, Wang R, et al. Neural theorem proving for verification conditions: A real-world benchmark. arXiv preprint arXiv:2601.18944, 2026
132. Balunović M, Dekoninck J, Petrov I, et al. Matharena: Evaluating llms on uncontaminated math competitions. arXiv preprint arXiv:2505.23281, 2025
133. Song C, Wang Z, Pu F, et al. Leangeo: Formalizing competitional geometry problems in lean. arXiv preprint arXiv:2508.14644, 2025
134. Yu W, Jiang Z, Dong Y, et al. Reclor: A reading comprehension dataset requiring logical reasoning. arXiv preprint arXiv:2002.04326, 2020
135. Moskvichev A, Odouard V V, Mitchell M. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. arXiv preprint arXiv:2305.07141, 2023
136. Li B, Donatelli L, Koller A, et al. Slog: A structural generalization benchmark for semantic parsing. In: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023. 3213–3232
137. Kamali D, Barezi E J, Kordjamshidi P. Nesycoco: A neuro-symbolic concept composer for compositional generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 4184–4193
138. Unsal M, Akkus A. Easyarc: Evaluating vision language models on true visual reasoning. arXiv preprint arXiv:2506.11595, 2025
139. Schmidt D M, Schubert R, Cimiano P. Compost: A benchmark for analyzing the ability of llms to compositionally interpret questions in a qald setting. In: Proceedings of International Semantic Web Conference. Springer, 2025. 3–22
140. McPheat L, Kaur N, Blackwell R, et al. Decompsr: A dataset for decomposed analyses of compositional multihop spatial reasoning. arXiv preprint arXiv:2511.02627, 2025
141. Yu Z, Zhao Y, Cohan A, et al. Humaneval pro and mbpp pro: Evaluating large language models on self-invoking code generation task. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 13253–13279
142. Wei A, Suresh T, Cao J, et al. Codearc: Benchmarking reasoning capabilities of llm agents for inductive program synthesis. arXiv preprint arXiv:2503.23145, 2025
143. Fan W, Zheng T, Hu Y, et al. Legal rule induction: Towards generalizable principle discovery from analogous judicial precedents. arXiv preprint arXiv:2505.14104, 2025
144. Patel A, Bhattamishra S, Goyal N. Are nlp models really able to solve simple math word problems? arXiv preprint arXiv:2103.07191, 2021
145. Gao L, Madaan A, Zhou S, et al. Pal: Program-aided language models. In: Proceedings of International conference on machine learning. PMLR, 2023. 10764–10799
146. Zhang H, Da J, Lee D, et al. A careful examination of large language model performance on grade school arithmetic. arXiv preprint arXiv:2405.00332, 2024
147. Li Q, Cui L, Zhao X, et al. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. arXiv preprint arXiv:2402.19255, 2024
148. Shalyt M, Elimelech R, Kaminer I. Asymob: Algebraic symbolic mathematical operations benchmark. arXiv preprint arXiv:2505.23851, 2025
149. Miao S Y, Liang C C, Su K Y. A diverse corpus for evaluating and developing english math word problem solvers. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, 2020. 975–984
150. Chen J, Tang J, Qin J, et al. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 513–523

 doi:10.20944/preprints202605.0806.v1

151. Zhang M L, Li Z Z, Yin F, et al. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. arXiv preprint arXiv:2407.07327, 2024
152. Chen J, Li T, Qin J, et al. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In: Proceedings of the 2022 conference on empirical methods in natural language processing, 2022. 3313–3323
153. Cao J, Xiao J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In: Proceedings of the 29th international conference on computational linguistics, 2022. 1511–1520
154. Xu X, Zhang J, Chen T, et al. Ugmathbench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models. arXiv preprint arXiv:2501.13766, 2025
155. O'Brien D, Haddow B, Allaway E, et al. Mathemagic: Generating dynamic mathematics benchmarks robust to memorization. arXiv preprint arXiv:2510.05962, 2025
156. He C, Luo R, Bai Y, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 3828–3850
157. Gao B, Song F, Yang Z, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. arXiv preprint arXiv:2410.07985, 2024
158. Sun H, Min Y, Chen Z, et al. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. arXiv preprint arXiv:2503.21380, 2025
159. Gulati A, Miranda B, Chen E, et al. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning in llms. 2025
160. Wei H, Xu Z, Yang B, et al. Skylenage technical report: Mathematical reasoning and contest-innovation benchmarks for multi-level math evaluation. arXiv preprint arXiv:2510.01241, 2025
161. Glazer E, Erdil E, Besiroglu T, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. arXiv preprint arXiv:2411.04872, 2024
162. Huang K, Guo J, Li Z, et al. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. arXiv preprint arXiv:2502.06453, 2025
163. An S, Cai X, Cao X, et al. Amo-bench: Large language models still struggle in high school math competitions. arXiv preprint arXiv:2510.26768, 2025
164. Luong M T, Hwang D, Nguyen H H, et al. Towards robust mathematical reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. 35406–35430
165. Duan B, Liang X, Lu S, et al. Gold-medal-level olympiad geometry solving with efficient heuristic auxiliary constructions. arXiv preprint arXiv:2512.00097, 2025
166. Mathematical Association of America. American Invitational Mathematics Examination (AIME). https://maa.org/maa-invitational-competitions/, 2024. Accessed: 2026-05-05
167. Lu P, Bansal H, Xia T, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023
168. Zou J, Wang Q, Thakur P, et al. Stem-pom: Evaluating language models math-symbol reasoning in document parsing. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 8183–8199
169. Bajpai A, Bhandari A, Nambi A, et al. Spatialmath: Spatial comprehension-infused symbolic reasoning for mathematical problem-solving. arXiv preprint arXiv:2601.17489, 2026
170. Wang X, Hu Z, Lu P, et al. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635, 2023
171. Chen W, Yin M, Ku M, et al. Theoremqa: A theorem-driven question answering dataset. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 7889–7901
172. Geva M, Khashabi D, Segal E, et al. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 2021, 9: 346–361
173. Talmor A, Yoran O, Catav A, et al. Multimodalqa: Complex question answering over text, tables and images. arXiv preprint arXiv:2104.06039, 2021
174. Trivedi H, Balasubramanian N, Khot T, et al. Musique: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 2022
175. Amouyal S J, Wolfson T, Rubin O, et al. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs. arXiv preprint arXiv:2205.12665, 2022
176. Ho M, Sharma A, Chang J, et al. Wikiwhy: Answering and explaining cause-and-effect questions. arXiv preprint arXiv:2210.12152, 2022

 doi:10.20944/preprints202605.0806.v1

177. Schnitzler J, Ho X, Huang J, et al. Morehopqa: More than multi-hop reasoning. arXiv preprint arXiv:2406.13397, 2024
178. Shen W, Wang M, Wang Y, et al. Are we on the right way for assessing document retrieval-augmented generation? arXiv preprint arXiv:2508.03644, 2025
179. Ho X, Nguyen A K D, Sugawara S, et al. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In: Proceedings of the 28th International Conference on Computational Linguistics, 2020. 6609–6625
180. Yang Z, Qi P, Zhang S, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. 2018, pages 2369–2380
181. Vu T, Iyyer M, Wang X, et al. Freshllms: Refreshing large language models with search engine augmentation. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 13697–13720
182. Press O, Zhang M, Min S, et al. Measuring and narrowing the compositionality gap in language models. In: Proceedings of Findings of the Association for Computational Linguistics, 2023. 5687–5711
183. Park S, Kim J, Han W S. Sparta: Scalable and principled benchmark of tree-structured multi-hop qa over text and tables. 2026
184. Zhu A, Hwang A, Dugan L, et al. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2024. 18–37
185. Tang Y, Yang Y. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. arXiv preprint arXiv:2401.15391, 2024
186. Arora S, Lewis P, Fan A, et al. Reasoning over public and private data in retrieval-based systems. Transactions of the Association for Computational Linguistics, 2023, 11: 902–921
187. Patil S G, Zhang T, Wang X, et al. Gorilla: Large language model connected with massive apis. arXiv preprint arXiv:2305.15334, 2023. URL https://arxiv.org/abs/2305.15334
188. Arun A, Dimino F, Agarwal T P, et al. Finreflectkg: Agentic construction and evaluation of financial knowledge graphs. In: Proceedings of the 6th ACM International Conference on AI in Finance, 2025. 283–290
189. Abdallah A, Ali M, Abdul-Mageed M, et al. Tempo: A realistic multi-domain benchmark for temporal reasoning-intensive retrieval. arXiv preprint arXiv:2601.09523, 2026
190. Abdallah A, Mounis M D, Abdalla M, et al. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. arXiv preprint arXiv:2601.09562, 2026
191. Masry A, Islam M S, Ahmed M, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. arXiv preprint arXiv:2504.05506, 2025
192. Shaham U, Segal E, Ivgi M, et al. Scrolls: Standardized comparison over long language sequences. arXiv preprint arXiv:2201.03533, 2022
193. Hudson G, Al Moubayed N. Muld: The multitask long document benchmark. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022. 3675–3685
194. Köksal A, Schick T, Korhonen A, et al. Longform: Effective instruction tuning with reverse instructions. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 7056–7078
195. Wu H, Zhan M, Tan H, et al. Vcsum: A versatile chinese meeting summarization dataset. arXiv preprint arXiv:2305.05280, 2023
196. Islam P, Kannappan A, Kiela D, et al. Financebench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944, 2023
197. Gu Z, Zhang L, Zhu X, et al. Detectbench: Can large language model detect and piece together implicit evidence? In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 199–222
198. Bertsch A, Pratapa A, Mitamura T, et al. Oolong: Evaluating long context reasoning and aggregation capabilities. arXiv preprint arXiv:2511.02817, 2025
199. Cohen V, Mooney R. Met-bench: Multimodal entity tracking for evaluating the limitations of vision-language and reasoning models. arXiv preprint arXiv:2502.10886, 2025
200. Li M, Zhang S, Zhang T, et al. Needlebench: Can llms do retrieval and reasoning in information-dense context? arXiv preprint arXiv:2407.11963, 2024
201. Li J, Wang M, Zheng Z, et al. Loogle: Can long-context language models understand long contexts? In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 16304–16333

 doi:10.20944/preprints202605.0806.v1

202. Wang C, Duan H, Zhang S, et al. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. arXiv preprint arXiv:2404.06480, 2024
203. Kuratov Y, Bulatov A, Anokhin P, et al. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. Advances in Neural Information Processing Systems, 2024, 37: 106519–106554
204. Liu X, Dong P, Hu X, et al. Longgenbench: Long-context generation benchmark. arXiv preprint arXiv:2410.04199, 2024
205. Yen H, Gao T, Hou M, et al. Helmet: How to evaluate long-context language models effectively and thoroughly. arXiv preprint arXiv:2410.02694, 2024
206. Zhuang T, Kuang C, Li X, et al. Docpuzzle: A process-aware benchmark for evaluating realistic long-context reasoning capabilities. arXiv preprint arXiv:2502.17807, 2025
207. Chen P, Jin H, Lee C C, et al. Longleader: A comprehensive leaderboard for large language models in long-context scenarios. In: 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), 2025. 8734–8750
208. Yang V, Jin H, Zhong S, et al. 100-longbench: Are de facto long-context benchmarks literally evaluating long-context ability? In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 17560–17576
209. Deng X, Gu Y, Zheng B, et al. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 2023, 36: 28091–28114
210. Chen Y, Ge Y, Ge Y, et al. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. International Journal of Computer Vision, 2026, 134: 118
211. Chang M, Chhablani G, Clegg A, et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. arXiv preprint arXiv:2411.00081, 2024
212. Lu Y, Wang J, Guo L, et al. R-horizon: How far can your large reasoning model really go in breadth and depth? arXiv preprint arXiv:2510.08189, 2025
213. Deng X, Da J, Pan E, et al. Swe-bench pro: Can ai agents solve long-horizon software engineering tasks? arXiv preprint arXiv:2509.16941, 2025
214. Monti S, Nicolini C, Pellegrini G, et al. Sokobench: Evaluating long-horizon planning and reasoning in large language models. arXiv preprint arXiv:2601.20856, 2026
215. Xu F, Yan H, Sun Q, et al. Odysseyarena: Benchmarking large language models for long-horizon, active and inductive interactions. arXiv preprint arXiv:2602.05843, 2026
216. Zhang Y, Jiang S, Li R, et al. Deepplanning: Benchmarking long-horizon agentic planning with verifiable constraints. arXiv preprint arXiv:2601.18137, 2026
217. Ziomek J, Bankes W, Wolf L, et al. Llm-wikirace: Benchmarking long-term planning and reasoning over real-world knowledge graphs. arXiv preprint arXiv:2602.16902, 2026
218. Hu X, Xia J, Xu S, et al. Ecogym: Evaluating llms for long-horizon plan-and-execute in interactive economies. arXiv preprint arXiv:2602.09514, 2026
219. Jubair S, Omayrah A, Alshammari A, et al. Lc-eval: A bilingual multi-task evaluation benchmark for long-context understanding. arXiv preprint arXiv:2510.16783, 2025
220. Huybrechts G, Ronanki S, Jayanthi S M, et al. Document haystack: A long context multimodal image/document understanding vision llm benchmark. arXiv preprint arXiv:2507.15882, 2025
221. Deng C, Yuan J, Bu P, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 1135–1159
222. Zhou S, Xu F F, Zhu H, et al. Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854, 2023
223. He H, Yao W, Ma K, et al. Webvoyager: Building an end-to-end web agent with large multimodal models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 6864–6890
224. Pan Y, Kong D, Zhou S, et al. Webcanvas: Benchmarking web agents in online environments. arXiv preprint arXiv:2406.12373, 2024
225. Fang S, Wang Y, Liu X, et al. Agentlongbench: A controllable long benchmark for long-contexts agents via environment rollouts. arXiv preprint arXiv:2601.20730, 2026
226. Miyai A, Zhao Z, Egashira K, et al. Webchorearena: Evaluating web browsing agents on realistic tedious web tasks. arXiv preprint arXiv:2506.01952, 2025

 doi:10.20944/preprints202605.0806.v1

227. Xi Y, Lin J, Zhu M, et al. Infodeepseek: Benchmarking agentic information seeking for retrieval-augmented generation. arXiv preprint arXiv:2505.15872, 2025
228. Gou B, Huang Z, Ning Y, et al. Mind2web 2: Evaluating agentic search with agent-as-a-judge. arXiv preprint arXiv:2506.21506, 2025
229. Yao S, Chen H, Yang J, et al. Webshop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems, 2022, 35: 20744–20757
230. Furuta H, Lee K H, Nachum O, et al. Multimodal web navigation with instruction-finetuned foundation models. arXiv preprint arXiv:2305.11854, 2023
231. Zhou X, Zhu H, Mathur L, et al. Sotopia: Interactive evaluation for social intelligence in language agents. arXiv preprint arXiv:2310.11667, 2023
232. Valmeekam K, Marquez M, Olmo A, et al. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. Advances in Neural Information Processing Systems, 2023, 36: 38975–38987
233. Paglieri D, Cupiał B, Coward S, et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. arXiv preprint arXiv:2411.13543, 2024
234. Shi W, Xu R, Zhuang Y, et al. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 22315–22339
235. Kapoor R, Butala Y P, Russak M, et al. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In: Proceedings of European Conference on Computer Vision. Springer, 2024. 161–178
236. Schmidgall S, Ziaei R, Harris C, et al. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. arXiv preprint arXiv:2405.07960, 2024
237. Trivedi H, Khot T, Hartmann M, et al. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 16022–16076
238. Jiang Y, Black K C, Geng G, et al. Medagentbench: a virtual ehr environment to benchmark medical llm agents. Nejm Ai, 2025, 2: AIdbp2500144
239. Rudakov E, Shock J, Cowley B U. Graph-based exploration for arc-agi-3 interactive reasoning tasks. arXiv preprint arXiv:2512.24156, 2025
240. Luo H, Zhang H, Zhang X, et al. Ultrahorizon: Benchmarking agent capabilities in ultra long-horizon scenarios. arXiv preprint arXiv:2509.21766, 2025
241. Moteki A, Masui S, Yang F, et al. Fieldworkarena: Agentic ai benchmark for real field work tasks. arXiv preprint arXiv:2505.19662, 2025
242. Starace G, Jaffe O, Sherburn D, et al. Paperbench: Evaluating ai's ability to replicate ai research. arXiv preprint arXiv:2504.01848, 2025
243. Kapoor S, Stroebl B, Kirgis P, et al. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. arXiv preprint arXiv:2510.11977, 2025
244. Xie T, Zhang D, Chen J, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems, 2024, 37: 52040–52094
245. Zeng Z, Liu J, Chen S, et al. Futurex: An advanced live benchmark for llm agents in future prediction. arXiv preprint arXiv:2508.11987, 2025
246. Xu T, Chen L, Wu D J, et al. Crab: Cross-environment agent benchmark for multimodal language model agents. arXiv preprint arXiv:2407.01511, 2024
247. Chai Y, Tang S, Xiao H, et al. A3: Android agent arena for mobile gui agents with essential-state procedural evaluation. arXiv preprint arXiv:2501.01149, 2025
248. Bonatti R, Zhao D, Bonacci F, et al. Windows agent arena: Evaluating multi-modal os agents at scale. arXiv preprint arXiv:2409.08264, 2024
249. Li M, Zhao Y, Yu B, et al. Api-bank: A comprehensive benchmark for tool-augmented llms. In: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023. 3102–3116
250. He W, Sun Y, Hao H, et al. Vitabench: Benchmarking llm agents with versatile interactive tasks in real-world applications. arXiv preprint arXiv:2509.26490, 2025
251. Tang Q, Deng Z, Lin H, et al. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. arXiv preprint arXiv:2306.05301, 2023

 doi:10.20944/preprints202605.0806.v1

252. Wu M, Zhu T, Han H, et al. Seal-tools: Self-instruct tool learning dataset for agent tuning and detailed benchmark. In: Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 2024. 372–384
253. Chen C, Hao X, Liu W, et al. Acebench: Who wins the match point in tool usage? arXiv preprint arXiv:2501.12851, 2025
254. Bandi C, Hertzberg B, Boo G, et al. Mcp-atlas: A large-scale benchmark for tool-use competency with real mcp servers. arXiv preprint arXiv:2602.00933, 2026
255. Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 2023, 36: 68539–68551
256. Qin Y, Liang S, Ye Y, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv preprint arXiv:2307.16789, 2023
257. Patil S G, Mao H, Yan F, et al. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In: Proceedings of Forty-second International Conference on Machine Learning, 2025
258. Lu S, Guo D, Ren S, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. arXiv preprint arXiv:2102.04664, 2021
259. Puri R, Kung D S, Janssen G, et al. Codenet: A large-scale ai for code dataset for learning a diversity of coding tasks. arXiv preprint arXiv:2105.12655, 2021
260. Zhang Z, Liu R, Liu A, et al. Code2bench: Scaling source and rigor for dynamic benchmark construction. In: Proceedings of The Fourteenth International Conference on Learning Representations, 2026
261. Austin J, Odena A, Nye M, et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021
262. Lai Y, Li C, Wang Y, et al. Ds-1000: A natural and reliable benchmark for data science code generation. In: Proceedings of International Conference on Machine Learning. PMLR, 2023. 18319–18345
263. Liu J, Xia C S, Wang Y, et al. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. Advances in neural information processing systems, 2023, 36: 21558–21572
264. Miao C, Zou H P, Li Y, et al. Recode-h: A benchmark for research code development with interactive human feedback. arXiv preprint arXiv:2510.06186, 2025
265. Wang Z, Liu S, Sun Y, et al. Codecontests+: High-quality test case generation for competitive programming. arXiv preprint arXiv:2506.05817, 2025
266. Yu H, Shen B, Ran D, et al. Codereval: A benchmark of pragmatic code generation with generative pre-trained models. arXiv preprint arXiv:2302.00288, 2023
267. Li J, Su Y, Lyu M R. From laboratory to real-world applications: Benchmarking agentic code reasoning at the repository level. arXiv preprint arXiv:2601.03731, 2026
268. Li J, Li G, Zhao Y, et al. Deveval: A manually-annotated code generation benchmark aligned with real-world code repositories. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 3603–3614
269. Fu L, Guan H, Zhang B, et al. Corecodebench: A configurable multi-scenario repository-level benchmark. arXiv preprint arXiv:2507.05281, 2025
270. Liu T, Xu C, McAuley J. Repobench: Benchmarking repository-level code auto-completion systems. arXiv preprint arXiv:2306.03091, 2023
271. Jain N, Han K, Gu A, et al. Livecodebench: Holistic and contamination free evaluation of large language models for code. arXiv preprint arXiv:2403.07974, 2024
272. Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021
273. Yang J, Prabhakar A, Narasimhan K, et al. Intercode: Standardizing and benchmarking interactive coding with execution feedback. Advances in Neural Information Processing Systems, 2023, 36: 23826–23854
274. Wang L, Ramalho L, Celestino A, et al. Swe-bench++: A framework for the scalable generation of software engineering benchmarks from open-source repositories. arXiv preprint arXiv:2512.17419, 2025
275. Yang J, Zhang J, Yang J, et al. Execrepobench: Multi-level executable code completion evaluation. arXiv preprint arXiv:2412.11990, 2024
276. Cheng Y, Chen J, Chen J, et al. Fullstack bench: Evaluating llms as full stack coders. arXiv preprint arXiv:2412.00535, 2024

 doi:10.20944/preprints202605.0806.v1

277. Zheng Z, Cheng Z, Shen Z, et al. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? arXiv preprint arXiv:2506.11928, 2025
278. Duan G, Liu M, Wang Y, et al. A hierarchical and evolvable benchmark for fine-grained code instruction following with multi-turn feedback. arXiv preprint arXiv:2507.00699, 2025
279. Ouyang S, Huang D, Guo J, et al. Dscodebench: A realistic benchmark for data science code generation. arXiv preprint arXiv:2505.15621, 2025
280. Wu C, Ge Y, Guo Q, et al. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. arXiv preprint arXiv:2405.07990, 2024
281. Zhang Y, Pan Y, Wang Y, et al. Pybench: Evaluating llm agent on various real-world coding tasks. arXiv preprint arXiv:2407.16732, 2024
282. Zhuo T Y, Vu M C, Chim J, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. arXiv preprint arXiv:2406.15877, 2024
283. Grötschla F, Müller L, Tönshoff J, et al. Agentsnet: Coordination and collaborative reasoning in multi-agent llms. arXiv preprint arXiv:2507.08616, 2025
284. Wang D, Cheng M, Yu S, et al. Paperarena: An evaluation benchmark for tool-augmented agentic reasoning on scientific literature. arXiv preprint arXiv:2510.10909, 2025
285. Li A, Zhang J, Li L, et al. M3mad-bench: Are multi-agent debates really effective across domains and modalities? arXiv preprint arXiv:2601.02854, 2026
286. Xu F F, Song Y, Li B, et al. Theagentcompany: Benchmarking llm agents on consequential real world tasks. arXiv preprint arXiv:2412.14161, 2024
287. Geng L, Chang E Y. Realm-bench: A benchmark for evaluating multi-agent systems on real-world, dynamic planning and scheduling tasks. arXiv preprint arXiv:2502.18836, 2025
288. Zhu K, Du H, Hong Z, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 8580–8622
289. Hyun J, Waytowich N R, Chen B. Crew-wildfire: Benchmarking agentic multi-agent collaborations at scale. arXiv preprint arXiv:2507.05178, 2025
290. Yang K, Swope A M, Gu A, et al. Leandojo: Theorem proving with retrieval-augmented language models, 2023
291. Chollet F. On the measure of intelligence. arXiv preprint arXiv:1911.01547, 2019
292. Patel B, Chakraborty S, Suttle W A, et al. Aime: Ai system optimization via multiple llm evaluators. arXiv preprint arXiv:2410.03131, 2024
293. Xu X, Xu Q, Xiao T, et al. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. arXiv preprint arXiv:2502.00334, 2025
294. Feng K, Zhao Y, Liu Y, et al. Physics: Benchmarking foundation models on university-level physics problem solving. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 11717–11743
295. Zhang Y, Ma Y, Gu Y, et al. Abench-physics: Benchmarking physical reasoning in llms via high-difficulty and dynamic physics problems. arXiv preprint arXiv:2507.04766, 2025
296. Huang Z, Wang Z, Xia S, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. Advances in Neural Information Processing Systems, 2024, 37: 19209–19253
297. Ling Z, Liu K, Yan K, et al. Longreason: A synthetic long-context reasoning benchmark via context expansion. arXiv preprint arXiv:2501.15089, 2025
298. Wang M, Chen L, Cheng F, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. 2024, pages 5627–5646
299. Zhou K, Tang Z, Ming L, et al. Mmlongcite: A benchmark for evaluating fidelity of long-context vision-language models. arXiv preprint arXiv:2510.13276, 2025
300. Shridhar M, Yuan X, Côté M A, et al. Alfworld: Aligning text and embodied environments for interactive learning. arXiv preprint arXiv:2010.03768, 2020
301. Huang Y, Shi J, Li Y, et al. Metatool benchmark for large language models: Deciding whether to use tools and which to use. arXiv preprint arXiv:2310.03128, 2023
302. Gu A, Rozière B, Leather H, et al. Cruxeval: A benchmark for code reasoning, understanding and execution. arXiv preprint arXiv:2401.03065, 2024
303. Prathifkumar T, Mathews N S, Nagappan M. Does swe-bench-verified test agent ability or model memory? arXiv preprint arXiv:2512.10218, 2025
304. Zhang L, He S, Zhang C, et al. Swe-bench goes live! arXiv preprint arXiv:2505.23419, 2025

 doi:10.20944/preprints202605.0806.v1

305. Dong Y, Zhu X, Pan Z, et al. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. 2024, pages 16290–16314
306. Sun H, Zhang S, Niu L, et al. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. 2025, pages 4922–4951
307. Ossowski T, Maqbool D, Chen J, et al. Comma: A communicative multimodal multi-agent benchmark. arXiv preprint arXiv:2410.07553, 2024
308. Yang H, Chen S, Nourzad N, et al. Emcoop: A framework and benchmark for embodied cooperation among llm agents. arXiv preprint arXiv:2603.00349, 2026
309. Zhang Y, Liu F, Shan Y, et al. Silo-bench: A scalable environment for evaluating distributed coordination in multi-agent llm systems. arXiv preprint arXiv:2603.01045, 2026
310. White C, Dooley S, Roberts M, et al. Livebench: A challenging, contamination-limited llm benchmark. arXiv preprint arXiv:2406.19314, 2024
311. Patel L, Arabzadeh N, Gupta H, et al. Deepscholar-bench: A live benchmark and automated evaluation for generative research synthesis. arXiv preprint arXiv:2508.20033, 2025
312. Kabir M, Ahmed T, Rahman M M, et al. Xcr-bench: A multi-task benchmark for evaluating cultural reasoning in llms. arXiv preprint arXiv:2601.14063, 2026
313. Lù X H, Kasner Z, Reddy S. Weblinx: Real-world website navigation with multi-turn dialogue. arXiv preprint arXiv:2402.05930, 2024
314. Wang Z, Chang Q, Patel H, et al. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. arXiv preprint arXiv:2508.20453, 2025
315. Xu Z, Zhou P, Ai J, et al. Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification. arXiv preprint arXiv:2503.12505, 2025
316. Xuan W, Yang R, Qi H, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. 1513–1532
317. Du X, Yao Y, Ma K, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. arXiv preprint arXiv:2502.14739, 2025
318. Lin B Y, Wu Z, Yang Y, et al. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. arXiv preprint arXiv:2101.00376, 2021
319. Pal A, Umapathi L K, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of Conference on health, inference, and learning. PMLR, 2022. 248–260
320. Zhang Y F, Zhang H, Tian H, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257, 2024
321. Wei A, Wu Y, Wan Y, et al. Satbench: Benchmarking llms' logical reasoning via automated puzzle generation from sat formulas. arXiv preprint arXiv:2505.14615, 2025
322. Tong H, Yue Z, Zhao F, et al. Cogtom: A comprehensive theory of mind benchmark inspired by human cognition for large language models. arXiv preprint arXiv:2601.15628, 2026
323. Liang Z, Guo K, Liu G, et al. Scemqa: A scientific college entrance level multimodal question answering benchmark. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2024. 109–119
324. Dubois Y, Galambosi B, Liang P, et al. Length-controlled alpacaeval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475, 2024
325. Chernyshev K, Polshkov V, Artemova E, et al. U-math: A university-level benchmark for evaluating mathematical skills in llms. arXiv preprint arXiv:2412.03205, 2024
326. Lee D, Park A, Lee H, et al. Typed-rag: Type-aware decomposition of non-factoid questions for retrieval-augmented generation. In: Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), 2025. 129–152
327. Ospanov A, Farnia F, Yousefzadeh R. minif2f-lean revisited: Reviewing limitations and charting a path forward. arXiv preprint arXiv:2511.03108, 2025
328. Merrill M A, Shaw A G, Carlini N, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. arXiv preprint arXiv:2601.11868, 2026
329. An C, Gong S, Zhong M, et al. L-eval: Instituting standardized evaluation for long context language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 14388–14411

 doi:10.20944/preprints202605.0806.v1

330. Ma Y, Zang Y, Chen L, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. Advances in Neural Information Processing Systems, 2024, 37: 95963–96010
331. Liu Z, Ping W, Roy R, et al. Chatqa: Surpassing gpt-4 on conversational qa and rag. Advances in Neural Information Processing Systems, 2024, 37: 15416–15459
332. Friel R, Belyi M, Sanyal A. Ragbench: Explainable benchmark for retrieval-augmented generation systems. arXiv preprint arXiv:2407.11005, 2024
333. Gong Z, Huang Y, Mai C. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering with hierarchical index and multi-granularity retrieval. arXiv e-prints, 2025, pages arXiv–2508
334. Chiang W L, Zheng L, Sheng Y, et al. Chatbot arena: An open platform for evaluating llms by human preference. arXiv preprint arXiv:2403.04132, 2024
335. Li X, Yao X, Qi G, et al. Findeepforecast: A live multi-agent system for benchmarking deep research agents in financial forecasting. arXiv preprint arXiv:2601.05039, 2026
336. Zhong L, Du Z, Zhang X, et al. Complexfuncbench: exploring multi-step and constrained function calling under long-context scenario. arXiv preprint arXiv:2501.10132, 2025
337. Lu Y, Liu S, Dong L. Orchdag: Complex tool orchestration in multi-turn interactions with plan dags. arXiv preprint arXiv:2510.24663, 2025
338. Xu Z, Soria A M, Tan S, et al. Toucan: Synthesizing 1.5 m tool-agentic data from real-world mcp environments. arXiv preprint arXiv:2510.01179, 2025
339. Zhou Y, Zhao M, Wang Z, et al. M^3-bench: Multi-modal, multi-hop, multi-threaded tool-using mllm agent benchmark. arXiv preprint arXiv:2511.17729, 2025
340. Rawles C, Clinckemaillie S, Chang Y, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. arXiv preprint arXiv:2405.14573, 2024
341. Xu W, Xiong J, Zhao C, et al. Swingarena: Competitive programming arena for long-context github issue solving. arXiv preprint arXiv:2505.23932, 2025
342. Qian C, Liu Z, Prabhakar A, et al. Userbench: An interactive gym environment for user-centric agents. arXiv preprint arXiv:2507.22034, 2025
343. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110, 2022
344. Guo Z, Chen T, Meng W, et al. Dynamic thinking-token selection for efficient reasoning in large reasoning models. arXiv preprint arXiv:2601.18383, 2026
345. Wang B, Xu C, Wang S, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. arXiv preprint arXiv:2111.02840, 2021
346. Yuan L, Chen Y, Cui G, et al. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. Advances in Neural Information Processing Systems, 2023, 36: 58478–58507
347. Stuhlmann L, Fadel Argerich M, Fürst J. Bench360: Benchmarking local llm inference from 360°. arXiv preprint arXiv:2511.16682, 2025
348. Kaiser D, Frigessi A, Ramezani-Kebrya A, et al. Decomposing reasoning efficiency in large language models. arXiv preprint arXiv:2602.09805, 2026
349. Mozannar H, Chen V, Alsobay M, et al. The realhumaneval: Evaluating large language models' abilities to support programmers. arXiv preprint arXiv:2404.02806, 2024
350. Qian Y, Wan C, Jia C, et al. Prism-bench: A benchmark of puzzle-based visual tasks with cot error detection. arXiv preprint arXiv:2510.23594, 2025
351. Chang M, Zhang J, Zhu Z, et al. Agentboard: An analytical evaluation board of multi-turn llm agents. Advances in neural information processing systems, 2024, 37: 74325–74362
352. Huang Y, Bai Y, Zhu Z, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. Advances in neural information processing systems, 2023, 36: 62991–63010
353. Li Y, Choi D, Chung J, et al. Competition-level code generation with alphacode. Science, 2022, 378: 1092–1097
354. Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step. In: Proceedings of The twelfth international conference on learning representations, 2024
355. Liu H, Liu J, Cui L, et al. Logiqa 2.0: An improved dataset for logical reasoning in natural language understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023
356. Vashurin R, Fadeeva E, Vazhentsev A, et al. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. In: Proceedings of Transactions of the Association for Computational Linguistics, 2025

 doi:10.20944/preprints202605.0806.v1

357. Wang X, Zhang Z, Chen G, et al. Ubench: Benchmarking uncertainty in large language models with multiple choice questions. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 8076–8107
358. Yang R, Zhang C, Zhang Z, et al. Uncle: Uncertainty expressions in long-form generation. arXiv e-prints, 2025, pages arXiv–2505
359. Müller P, Popovič N, Färber M, et al. Benchmarking uncertainty calibration in large language models for long-form question answering. arXiv preprint arXiv:2602.00279, 2026
360. Min S, Krishna K, Lyu X, et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 12076–12100
361. Li X, Cao Y, Pan L, et al. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 493–516
362. Aly R, Guo Z, Schlichtkrull M, et al. Feverous: Fact extraction and verification over unstructured and structured information. arxiv 2021. arXiv preprint arXiv:2106.05707
363. Ruan Y, Dong H, Wang A, et al. Identifying the risks of lm agents with an lm-emulated sandbox. arXiv preprint arXiv:2309.15817, 2023
364. Lin S, Hilton J, Evans O. Truthfulqa: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022
365. Li J, Cheng X, Zhao W X, et al. Halueval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023. 6449–6464
366. Zhang Z, Lei L, Wu L, et al. Safetybench: Evaluating the safety of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 15537–15553
367. Vidgen B, Agrawal A, Ahmed A M, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. arXiv preprint arXiv:2404.12241, 2024
368. Mazeika M, Phan L, Yin X, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249, 2024
369. Chao P, Debenedetti E, Robey A, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. Advances in Neural Information Processing Systems, 2024, 37: 55005–55029
370. Kim S, Yun S, Lee H, et al. Propile: Probing privacy leakage in large language models. Advances in Neural Information Processing Systems, 2023, 36: 20750–20762
371. Wang J, Yang T, Xie R, et al. Raccoon: Prompt extraction benchmark of llm-integrated applications. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 13349–13365
372. Mukhopadhyay S, Reddy S, Muthukumar S, et al. Privacybench: A conversational benchmark for evaluating privacy in personalized ai. arXiv preprint arXiv:2512.24848, 2025
373. Kiela D, Bartolo M, Nie Y, et al. Dynabench: Rethinking benchmarking in nlp. In: Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, 2021. 4110–4124
374. Goel K, Rajani N F, Vig J, et al. Robustness gym: Unifying the nlp evaluation landscape. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, 2021. 42–55
375. Zhu K, Wang J, Zhou J, et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In: Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis, 2023. 57–68
376. Fursin G, Altunay D. Framing ai system benchmarking as a learning task: Flexbench and the open mlperf dataset. arXiv preprint arXiv:2509.11413, 2025
377. Mehditabar M, Rajput S, Mastropaolo A, et al. Smart but costly? benchmarking llms on functional accuracy and energy efficiency. arXiv preprint arXiv:2511.07698, 2025
378. Shi Z, Wang Y, Yan L, et al. Retrieval models aren't tool-savvy: Benchmarking tool retrieval for large language models. arXiv preprint arXiv:2503.01763, 2025
379. Li Y, Yuan P, Feng S, et al. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In: Proceedings of The Twelfth International Conference on Learning Representations, 2024. arXiv:2401.10480
380. Wang J, Jain S, Zhang D, et al. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In: Proceedings of Al-Onaizan Y, Bansal M, Chen Y N, editors, the 2024 Conference on

 doi:10.20944/preprints202605.0806.v1

Empirical Methods in Natural Language Processing, Miami, Florida, USA: Association for Computational Linguistics, 2024. 19916–19939
381. Kaiser D, Frigessi A, Ramezani-Kebrya A, et al. Cogniload: A synthetic natural language reasoning benchmark with tunable length, intrinsic difficulty, and distractor density. arXiv preprint arXiv:2509.18458, 2025
382. Tschand A, Rajan A T R, Idgunji S, et al. Mlperf power: Benchmarking the energy efficiency of machine learning systems from microwatts to megawatts for sustainable ai. arXiv preprint arXiv:2410.12032, 2024
383. Gao T, Yen H, Yu J, et al. Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 6465–6488
384. Shen X, Wang S, Tan Z, et al. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. arXiv preprint arXiv:2510.04040, 2025
385. Gupta P, Wu C S, Liu W, et al. Dialfact: A benchmark for fact-checking in dialogue. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022. 3785–3801
386. Park J, Min S, Kang J, et al. Faviq: Fact verification from information-seeking questions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022. 5154–5166
387. Chakraborty M, Pahwa K, Rani A, et al. Factify3m: A benchmark for multimodal fact verification with explainability through 5w question-answering. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 15282–15322
388. Uddin M N, Saeidi A, Handa D, et al. Unseentimeqa: Time-sensitive question-answering beyond llms' memorization. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 1873–1913
389. Wang P, Tao R, Chen Q, et al. X-webagentbench: A multilingual interactive web benchmark for evaluating global agentic system. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2025, 2025. 19320–19335
390. Zhang Y, Liu X, Zhou R, et al. Cchall: A novel benchmark for joint cross-lingual and cross-modal hallucinations detection in large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 30728–30749
391. Wang Y, Li H, Han X, et al. Do-not-answer: A dataset for evaluating safeguards in llms. arXiv preprint arXiv:2308.13387, 2023
392. Röttger P, Kirk H, Vidgen B, et al. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024. 5377–5400
393. Yuan T, He Z, Dong L, et al. R-judge: Benchmarking safety risk awareness for llm agents. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 1467–1490
394. Shahriar S, Dara R. Priv-iq: A benchmark and comparative evaluation of large multimodal models on privacy competencies. AI, 2025, 6: 29
395. Debenedetti E, Rando J, Paleka D, et al. Dataset and lessons learned from the 2024 satml llm capture-the-flag competition. Advances in Neural Information Processing Systems, 2024, 37: 36914–36937
396. Zhan Q, Liang Z, Ying Z, et al. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. arXiv preprint arXiv:2403.02691, 2024
397. Zhang W, Aljunied M, Gao C, et al. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. Advances in Neural Information Processing Systems, 2023, 36: 5484–5505
398. Souly A, Lu Q, Bowen D, et al. A strongreject for empty jailbreaks. Advances in Neural Information Processing Systems, 2024, 37: 125416–125440
399. Wang R, Yu H, Zhang W, et al. Sotopia-π: Interactive learning of socially intelligent language agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 12912–12940
400. German E, Antebi S, Samira D, et al. Tab-mia: A benchmark dataset for membership inference attacks on tabular data in llms. arXiv preprint arXiv:2507.17259, 2025
401. Wang Y, Zhang P, Tang J, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. arXiv preprint arXiv:2504.18428, 2025
402. Puerto H, Gubri M, Yun S, et al. Scaling up membership inference: When and how attacks succeed on large language models. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 4165–4182

 doi:10.20944/preprints202605.0806.v1

403. Yang K, Deng J, Chen D. Generating natural language proofs with verifier-guided search. arXiv preprint arXiv:2205.12443, 2022
404. Queiroz J. Adversarial versification in portuguese as a jailbreak operator in llms. arXiv preprint arXiv:2512.15353, 2025
405. Li H, Zhang Y, Koto F, et al. Cmmlu: Measuring massive multitask language understanding in chinese. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 11260–11285
406. Abhyankar R, Qi Q, Zhang Y. Osworld-human: Benchmarking the efficiency of computer-use agents. arXiv preprint arXiv:2506.16042, 2025
407. Singh S, Romanou A, Fourrier C, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 18761–18799
408. Myung J, Lee N, Zhou Y, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. Advances in Neural Information Processing Systems, 2024, 37: 78104–78146
409. Hupkes D, Bogoychev N. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. arXiv preprint arXiv:2504.10356, 2025
410. Liu C, Zhang W, Ying J, et al. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 6119–6136
411. Chiu Y Y, Jiang L, Lin B Y, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. arXiv preprint arXiv:2410.02677, 2024
412. Gao Z, Xu Y, Thebault-Spieker J. Localbench: Benchmarking llms on county-level local knowledge and reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2026. 38487–38495
413. Frohberg J, Binder F. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022. 2126–2140
414. Chalkidis I, Jana A, Hartung D, et al. Lexglue: A benchmark dataset for legal language understanding in english. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022. 4310–4330
415. Liu C, Jin R, Ren Y, et al. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. arXiv preprint arXiv:2305.10263, 2023
416. Altakrori M H, Habash N, Freihat A A, et al. Dialectalarabicmmlu: Benchmarking dialectal capabilities in arabic and multilingual language models. arXiv preprint arXiv:2510.27543, 2025
417. Zhang X, Li C, Zong Y, et al. Evaluating the performance of large language models on gaokao benchmark. arXiv preprint arXiv:2305.12474, 2023
418. Guha N, Nyarko J, Ho D, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in neural information processing systems, 2023, 36: 44123–44279
419. Wang X, Yeo J, Lim J H, et al. Kulture bench: A benchmark for assessing language model in korean cultural context. arXiv preprint arXiv:2412.07251, 2024
420. Wang Z. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In: Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), 2024. 143–151
421. Romanou A, Foroutan N, Sotnikova A, et al. Include: Evaluating multilingual language understanding with regional knowledge. arXiv preprint arXiv:2411.19799, 2024
422. Cao C, Zhu Z, Zhu J, et al. Measuring hong kong massive multi-task language understanding. arXiv preprint arXiv:2505.02177, 2025
423. Chen Y, Singh V K, Ma J, et al. Counterbench: A benchmark for counterfactuals reasoning in large language models. arXiv preprint arXiv:2502.11008, 2025
424. Pramodya A, Nelki N, Shalinda H, et al. Sinhalammlu: A comprehensive benchmark for evaluating multitask language understanding in sinhala. arXiv preprint arXiv:2509.03162, 2025
425. Fabbri A R, Mares D, Flores J, et al. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. arXiv preprint arXiv:2507.17476, 2025
426. Nacar O, Sibaee S T, Ahmed S, et al. Towards inclusive arabic llms: A culturally aligned benchmark in arabic large language model evaluation. In: Proceedings of the First Workshop on Language Models for Low-Resource Languages, 2025. 387–401

 doi:10.20944/preprints202605.0806.v1

427. Ponti E M, Glavaš G, Majewska O, et al. Xcopa: A multilingual dataset for causal commonsense reasoning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 2362–2376
428. Shi F, Suzgun M, Freitag M, et al. Language models are multilingual chain-of-thought reasoners. arXiv preprint arXiv:2210.03057, 2022
429. Gusev I, Tikhonov A. Headlinecause: A dataset of news headlines for detecting causalities. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022. 6153–6161
430. Lin B Y, Lee S, Qiao X, et al. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. arXiv preprint arXiv:2106.06937, 2021
431. Lai V D, Veyseh A P B, Van Nguyen M, et al. Meci: A multilingual dataset for event causality identification. In: Proceedings of the 29th international conference on computational linguistics, 2022. 2346–2356
432. Lai V D, Van Nguyen C, Ngo N T, et al. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. arXiv preprint arXiv:2307.16039, 2023
433. Bandarkar L, Liang D, Muller B, et al. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 749–775
434. Sakai Y, Kamigaito H, Watanabe T. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 14182–14214
435. Liu Y, Xu M, Wang S, et al. Omgeval: An open multilingual generative evaluation benchmark for large language models. arXiv preprint arXiv:2402.13524, 2024
436. Chang T A, Arnett C, Eldesokey A, et al. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. arXiv preprint arXiv:2510.24081, 2025
437. Edwards C, Han C, Lee G, et al. mclm: A function-infused and synthesis-friendly modular chemical language model. arXiv preprint arXiv:2505.12565, 2025
438. Luo W, Zhao W X, Sha J, et al. Mmath: A multilingual benchmark for mathematical reasoning. arXiv preprint arXiv:2505.19126, 2025
439. Sobhani M E, Sayeedi M F A, Mohiuddin T, et al. Mathmist: A parallel multilingual benchmark dataset for mathematical problem solving and reasoning. arXiv preprint arXiv:2510.14305, 2025
440. Gurgurov D, Ghussin Y A, Baeumel T, et al. Clas-bench: A cross-lingual alignment and steering benchmark. arXiv preprint arXiv:2601.08331, 2026
441. Blandón M A C, Talur J, Charron B, et al. Memerag: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 22577–22595
442. Kulkarni M, Mazzia V, Gaspers J, et al. Massive-agents: A benchmark for multilingual function-calling in 52 languages. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 20193–20215
443. Hofman O, Brokman J, Rachmil O, et al. Maps: A multilingual benchmark for global agent performance and security. arXiv preprint arXiv:2505.15935, 2025
444. Li J, Lu W, Fei H, et al. A survey on benchmarks of multimodal large language models. arXiv preprint arXiv:2408.08632, 2024
445. Masry A, Do X L, Tan J Q, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Proceedings of Findings of the association for computational linguistics, 2022. 2263–2279
446. Cheng Z, Chen Q, Zhang J, et al. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 23678–23686
447. Ji Y, Chen H, Chen Q, et al. Mpcc: A novel benchmark for multimodal planning with complex constraints in multimodal large language models. In: Proceedings of the 33rd ACM International Conference on Multimedia, 2025. 5188–5197
448. Huang J, Zhang J. A survey on evaluation of multimodal large language models. arXiv preprint arXiv:2408.15769, 2024
449. Mathew M, Karatzas D, Jawahar C. Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021. 2200–2209
450. Yan J, Ren R, Liu J, et al. Teleego: Benchmarking egocentric ai assistants in the wild. arXiv preprint arXiv:2510.23981, 2025

 doi:10.20944/preprints202605.0806.v1

451. Guan T, Liu F, Wu X, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024. 14375–14385
452. Wang W, Gao Z, Chen L, et al. Visualprm: An effective process reward model for multimodal reasoning. arXiv preprint arXiv:2503.10291, 2025
453. Verma A, Puttagunta S, Subramanian S, et al. Graft: Graph and table reasoning for textual alignment–a benchmark for structured instruction following and visual reasoning. arXiv preprint arXiv:2508.15690, 2025
454. Chen S, Chen Y, Li Z, et al. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. 10091–10109
455. Cassano F, Gouwar J, Nguyen D, et al. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. arXiv preprint arXiv:2208.08227, 2022
456. Athiwaratkun B, Gouda S K, Wang Z, et al. Multi-lingual evaluation of code generation models. arXiv preprint arXiv:2210.14868, 2022
457. Wang S, Li Z, Qian H, et al. Recode: Robustness evaluation of code generation models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023. 13818–13843
458. Huang J, Wang C, Zhang J, et al. Execution-based evaluation for data science code generation models. In: Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), 2022. 28–36
459. Hao Y, Li G, Liu Y, et al. Aixbench: A code generation benchmark dataset. arXiv preprint arXiv:2206.13179, 2022
460. Yin P, Li W D, Xiao K, et al. Natural language to code generation in interactive data science notebooks. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023. 126–173
461. Tang X, Qian B, Gao R, et al. Biocoder: a benchmark for bioinformatics code generation with large language models. Bioinformatics, 2024, 40: i266–i276
462. Li R, Fu J, Zhang B W, et al. Taco: Topics in algorithmic code generation dataset. arXiv preprint arXiv:2312.14852, 2023
463. Fu L, Chai H, Luo S, et al. Codeapex: A bilingual programming evaluation benchmark for large language models. arXiv preprint arXiv:2309.01940, 2023
464. Xu Y, Chen Y, Zhang X, et al. Cloudeval-yaml: A practical benchmark for cloud configuration generation. Machine Learning and Systems, 2024, 6: 173–195
465. Ahmed T, Hirzel M, Pan R, et al. Tdd-bench verified: Can llms generate tests for issues before they get resolved? arXiv preprint arXiv:2412.02883, 2024
466. Huang D, Qing Y, Shang W, et al. Effibench: Benchmarking the efficiency of automatically generated code. Advances in Neural Information Processing Systems, 2024, 37: 11506–11544
467. Dinella E, Chandra S, Maniatis P. Crqbench: A benchmark of code reasoning questions. arXiv preprint arXiv:2408.08453, 2024
468. Yang J, Jimenez C E, Zhang A L, et al. Swe-bench multimodal: Do ai systems generalize to visual software domains? arXiv preprint arXiv:2410.03859, 2024
469. Yang J, Lieret K, Yang J, et al. Codeclash: Benchmarking goal-oriented software engineering. arXiv preprint arXiv:2511.00839, 2025
470. Yang L, Jin R, Shi L, et al. Probench: Benchmarking large language models in competitive programming. arXiv preprint arXiv:2502.20868, 2025
471. Tang Y, Zhu K, Ruan B, et al. Devops-gym: Benchmarking ai agents in software devops cycle. arXiv preprint arXiv:2601.20882, 2026
472. Zan D, Huang Z, Liu W, et al. Multi-swe-bench: A multilingual benchmark for issue resolving. arXiv preprint arXiv:2504.02605, 2025
473. Zhang L, Wang J, He S, et al. Di-bench: Benchmarking large language models on dependency inference with testable repositories at scale. arXiv preprint arXiv:2501.13699, 2025
474. Xie D, Zheng M, Liu X, et al. Core: Benchmarking llms code reasoning capabilities through static analysis tasks. arXiv preprint arXiv:2507.05269, 2025
475. Roy M K, Chen S, Steenhoek B, et al. Codesense: a real-world benchmark and dataset for code semantic reasoning. arXiv preprint arXiv:2506.00750, 2025

 doi:10.20944/preprints202605.0806.v1

476. Li X, Chen W, Liu Y, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks. arXiv preprint arXiv:2602.12670, 2026
477. Wang Y, Zhang Z, Wang C, et al. Realsec-bench: A benchmark for evaluating secure code generation in real-world repositories. arXiv preprint arXiv:2601.22706, 2026
478. Lim S, Hahn J, Park H, et al. Contracteval: A benchmark for evaluating contract-satisfying assertions in code generation. arXiv preprint arXiv:2510.12047, 2025
479. Ye Z, Chen J, Shao Z, et al. Solcontracteval: A benchmark for evaluating contract-level solidity code generation. arXiv preprint arXiv:2509.23824, 2025
480. Alghanmi I, Anke L E, Schockaert S. Probing pre-trained language models for disease knowledge. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 3023–3033
481. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. arXiv e-prints, 2020, pages arXiv–2009
482. Gao Y, Dligach D, Miller T, et al. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. Journal of biomedical informatics, 2023, 138: 104286
483. Blinov P, Reshetnikova A, Nesterov A, et al. Rumedbench: a russian medical language understanding benchmark. In: Proceedings of International Conference on Artificial Intelligence in Medicine. Springer, 2022. 383–392
484. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature, 2023, 620: 172–180
485. Xiong J, Shen J, Yuan Y, et al. Trigo: Benchmarking formal mathematical proof reduction for generative language models. arXiv preprint arXiv:2310.10180, 2023
486. Kim Y, Wu J, Abdulle Y, et al. Medexqa: Medical question answering benchmark with multiple explanations. In: Proceedings of the 23rd Workshop on biomedical natural language processing, 2024. 167–181
487. Nguyen D, Ho M K, Ta H, et al. Localizing before answering: A benchmark for grounded medical visual question answering. In: Proceedings of Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25). International Joint Conferences on Artificial Intelligence Organization, 2025
488. Zhao H, Tang X, Yang Z, et al. Chemsafetybench: Benchmarking llm safety on chemistry domain. arXiv preprint arXiv:2411.16736, 2024
489. Zhang J, Gan J, Wang X, et al. Matscibench: Benchmarking the reasoning ability of large language models in materials science. arXiv preprint arXiv:2510.12171, 2025
490. Ye X, Li C, Chen S, et al. Mmscibench: Benchmarking language models on chinese multimodal scientific problems. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 14621–14663
491. Kweon S, Choi B, Chu G, et al. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations. arXiv preprint arXiv:2403.01469, 2024
492. Arora R K, Wei J, Hicks R S, et al. Healthbench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775, 2025
493. Adams L, Busch F, Han T, et al. Longhealth: A question answering benchmark with long clinical documents. Journal of Healthcare Informatics Research, 2025, 9: 280–296
494. Su E, Wu J, Tang C, et al. Sciif: Benchmarking scientific instruction following towards rigorous scientific intelligence. arXiv preprint arXiv:2601.04770, 2026
495. Banerjee O, Kim S E, Willauer A N, et al. Rexinthewild: A unified benchmark for medical photograph understanding. arXiv preprint arXiv:2603.19517, 2026
496. Zi X, Zhou X, Xiao J, et al. Shattering the shortcut: A topology-regularized benchmark for multi-hop medical reasoning in llms. arXiv preprint arXiv:2603.12458, 2026
497. Yoshitake M, Suzuki Y, Igarashi R, et al. Materialfigbench: benchmark dataset with figures for evaluating college-level materials science problem-solving abilities of multimodal large language models. arXiv preprint arXiv:2603.11414, 2026
498. Zhang D, Shen Z, Xie R, et al. Mobile-env: Building qualified evaluation benchmarks for llm-gui interaction. arXiv preprint arXiv:2305.08144, 2023
499. Koh J Y, Lo R, Jang L, et al. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 881–905
500. Xue T, Qi W, Shi T, et al. An illusion of progress? assessing the current state of web agents. arXiv preprint arXiv:2504.01382, 2025
501. Anupam S, Brown D, Li S, et al. Browserarena: Evaluating llm agents on real-world web navigation tasks. arXiv preprint arXiv:2510.02418, 2025

 doi:10.20944/preprints202605.0806.v1

502. Lyu Y, Zhang X, Yan L, et al. Deepshop: A benchmark for deep research shopping agents. arXiv preprint arXiv:2506.02839, 2025
503. Cao Y, Wang Y, Bu P, et al. Androidlens: Long-latency evaluation with nested sub-targets for android gui agents. arXiv preprint arXiv:2512.21302, 2025
504. Li S, Kallidromitis K, Gokul A, et al. Mobileworldbench: Towards semantic world modeling for mobile agents. arXiv preprint arXiv:2512.14014, 2025
505. Shi Y, Li J, Zhang L, et al. Androtmem: From interaction trajectories to anchored memory in long-horizon gui agents. arXiv preprint arXiv:2603.18429, 2026
506. Zhao K, Song J, Sha L, et al. Gui testing arena: A unified benchmark for advancing autonomous gui testing agent. arXiv preprint arXiv:2412.18426, 2024
507. Li Y, Liu Y, Lu H, et al. Gui-ceval: A hierarchical and comprehensive chinese benchmark for mobile gui agents. arXiv preprint arXiv:2603.15039, 2026
508. Sun J, Li M, Zhang Y, et al. Ambibench: Benchmarking mobile gui agents beyond one-shot instructions in the wild. arXiv preprint arXiv:2602.11750, 2026
509. Zhao H H, Yang K, Yu W, et al. Worldgui: An interactive benchmark for desktop gui automation from any starting point. arXiv preprint arXiv:2502.08047, 2025
510. Farn N, Shin R. Tooltalk: Evaluating tool-usage in a conversational setting. arXiv preprint arXiv:2311.10775, 2023
511. Zhuang Y, Yu Y, Wang K, et al. Toolqa: A dataset for llm question answering with external tools. arXiv preprint arXiv:2306.13304, 2023
512. Chen Z, Du W, Zhang W, et al. T-eval: Evaluating the tool utilization capability of large language models step by step. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024
513. Mou X, Liang J, Lin J, et al. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. arXiv preprint arXiv:2410.19346, 2024
514. Khanna M, Ramrakhya R, Chhablani G, et al. Goat-bench: A benchmark for multi-modal lifelong navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 16373–16383
515. Zhao W, Schmidt L, Zou J, et al. Zebraarena: A diagnostic simulation environment for studying reasoning-action coupling in tool-augmented llms. arXiv preprint arXiv:2603.18614, 2026
516. Jiang X, Chang D, McAuley J, et al. When benchmarks age: Temporal misalignment through large language model factuality evaluation. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), 2026. 500–512
517. Thakur N, Reimers N, Rücklé A, et al. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663, 2021
518. Zhang X, Thakur N, Ogundepo O, et al. Miracl: A multilingual retrieval dataset covering 18 diverse languages. Transactions of the Association for Computational Linguistics, 2023, 11: 1114–1131
519. Bordes F, Ross C, Kao J T, et al. Eval factsheets: A structured framework for documenting ai evaluations. arXiv preprint arXiv:2512.04062, 2025
520. Taghanaki S A, Khani A, Khasahmadi A. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. arXiv preprint arXiv:2409.02257, 2024
521. Liu J, Qian C, Su Z, et al. Costbench: Evaluating multi-turn cost-optimal planning and adaptation in dynamic environments for llm tool-use agents. arXiv preprint arXiv:2511.02734, 2025
522. Ahuja S, Gumma V, Sitaram S. Contamination report for multilingual benchmarks. arXiv preprint arXiv:2410.16186, 2024
523. Feuer B, Tseng C Y, Lathe A S, et al. When judgment becomes noise: How design failures in llm judge benchmarks silently undermine validity. arXiv preprint arXiv:2509.20293, 2025
524. Li T, Chiang W L, Frick E, et al. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In: Proceedings of the International Conference on Machine Learning, 2025
525. Zhu K, Zhao Q, Chen H, et al. Promptbench: A unified library for evaluation of large language models. arXiv preprint arXiv:2312.07910, 2023
526. Chezelles D, Le Sellier T, Shayegan S O, et al. The browsergym ecosystem for web agent research. arXiv preprint arXiv:2412.05467, 2024
527. Filali A E, Bedar I. Towards more standardized ai evaluation: From models to agents. arXiv preprint arXiv:2602.18029, 2026

 doi:10.20944/preprints202605.0806.v1

528. Reuel A, Hardy A, Smith C, et al. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. Advances in Neural Information Processing Systems, 2024, 37: 21763–21813
529. Eriksson M, Purificato E, Noroozian A, et al. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025. 850–864
530. Cheng Z, Wohnig S, Gupta R, et al. Benchmarking is broken–don't let ai be its own judge. arXiv preprint arXiv:2510.07575, 2025
531. He P, Dai Z, He B, et al. Traject-bench: A trajectory-aware benchmark for evaluating agentic tool use. arXiv preprint arXiv:2510.04550, 2025
532. Chen Y, Jiang J, Liu J, et al. Trace: Trajectory-aware comprehensive evaluation for deep research agents. arXiv preprint arXiv:2602.21230, 2026
533. Kim S, Wang J, Xie X, et al. Harnessing temporal databases for systematic evaluation of factual time-sensitive question-answering in llms. In: Proceedings of The Fourteenth International Conference on Learning Representations, 2026
534. Meem J, Rashid M, Dong Y, et al. Pat-questions: A self-updating benchmark for present-anchored temporal question-answering. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 13129–13148
535. Wang Z, Yu W, Ren X, et al. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. arXiv preprint arXiv:2505.10610, 2025
536. Wei J, Yang C, Song X, et al. Long-form factuality in large language models. Advances in Neural Information Processing Systems, 2024, 37: 80756–80827
537. Shridhar M, Thomason J, Gordon D, et al. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020. 10740–10749
538. Srivastava S, Li C, Lingelbach M, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: Proceedings of Conference on robot learning. PMLR, 2022. 477–490
539. Li C, Zhang R, Wong J, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Proceedings of Conference on Robot Learning. PMLR, 2023. 80–93
540. Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691, 2022
541. Driess D, Xia F, Sajjadi M S, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023
542. Zitkovich B, Yu T, Xu S, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Proceedings of Conference on Robot Learning. PMLR, 2023. 2165–2183
543. Kim T, Min C, Kim B, et al. Realfred: An embodied instruction following benchmark in photo-realistic environments. In: Proceedings of European Conference on Computer Vision. Springer, 2024. 346–364
544. Yang R, Chen H, Zhang J, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560, 2025
545. Majumdar A, Ajay A, Zhang X, et al. Openeqa: Embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024. 16488–16498
546. Cao J, Chan Y K, Ling Z, et al. How should we build a benchmark? revisiting 274 code-related benchmarks for llms. arXiv preprint arXiv:2501.10711, 2025
547. Jiang H, Zhang S, Yi X, et al. Position: Science of ai evaluation requires item-level benchmark data. arXiv preprint arXiv:2604.03244, 2026
548. Diddee H, Yauney G, Swayamdipta S, et al. Benchbrowser–collecting evidence for evaluating benchmark validity. arXiv preprint arXiv:2603.18019, 2026
549. Li G, Xie Y, Liu Y, et al. The world won't stay still: Programmable evolution for agent benchmarks. arXiv preprint arXiv:2603.05910, 2026
550. Joaquin A S, Gipiškis R, Staufer L, et al. Deprecating benchmarks: Criteria and framework. arXiv preprint arXiv:2507.06434, 2025
551. Huang J, Chang K C C. Towards reasoning in large language models: A survey. In: Proceedings of Findings of the association for computational linguistics, 2023. 1049–1065

 doi:10.20944/preprints202605.0806.v1

552. Sun J, Zheng C, Xie E, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. ACM Computing Surveys, 2025, 57: 1–43
553. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology, 2024, 15: 1–45
554. Xu C, Guan S, Greene D, et al. Benchmark data contamination of large language models: A survey. arXiv preprint arXiv:2406.04244, 2024
555. Yang L, Shirvaikar V, Clivio O, et al. A critical review of causal reasoning benchmarks for large language models. arXiv preprint arXiv:2407.08029, 2024