

# Toward Generalizable Evaluation in the LLM Era: A Survey Beyond Benchmarks

Yixin Cao<sup>1</sup>, Shibo Hong<sup>1</sup>, Xinze Li<sup>2</sup>, Jiahao Ying<sup>3</sup>, Yubo Ma<sup>2</sup>, Haiyuan Liang<sup>1</sup>, Yantao Liu<sup>1</sup>, Zijun Yao<sup>4</sup>, Xiaozhi Wang<sup>4</sup>, Dan Huang<sup>3</sup>, Wenxuan Zhang<sup>5</sup>, Lifu Huang<sup>6</sup>, Muhao Chen<sup>6</sup>, Lei Hou<sup>4</sup>, Qianru Sun<sup>3</sup>, Xingjun Ma<sup>1</sup>, Zuxuan Wu<sup>1</sup>, Min-Yen Kan<sup>7</sup>, David Lo<sup>3</sup>, Qi Zhang<sup>1</sup>, Heng Ji<sup>8</sup>, Jing Jiang<sup>9</sup>, Juanzi Li<sup>4</sup>, Aixin Sun<sup>2</sup>, Xuanjing Huang<sup>1</sup>, Tat-Seng Chua<sup>7</sup>, Yu-Gang Jiang<sup>1</sup>

<sup>1</sup>Fudan University, <sup>2</sup>Nanyang Technological University, <sup>3</sup>Singapore Management University, <sup>4</sup>Tsinghua University, <sup>5</sup>Singapore University of Technology and Design, <sup>6</sup>University of California Davis, <sup>7</sup>National University of Singapore, <sup>8</sup>University of Illinois Urbana-Champaign, <sup>9</sup>Australian National University

**Abstract**—Large Language Models (LLMs) are advancing at an amazing speed and have become indispensable across academia, industry, and daily applications. To keep pace with the status quo, this survey probes the core challenges that the rise of LLMs poses for evaluation. We identify and analyze two pivotal transitions: (i) from task-specific to capability-based evaluation, which reorganizes benchmarks around core competencies such as knowledge, reasoning, instruction following, multi-modal understanding, and safety; and (ii) from manual to automated evaluation, encompassing dynamic dataset curation and “LLM-as-a-judge” scoring.

Yet, even with these transitions, a crucial obstacle persists: the evaluation generalization issue. Bounded test sets cannot scale alongside models whose abilities grow seemingly without limit. We will dissect this issue, along with the core challenges of the above two transitions, from the perspectives of methods, datasets, evaluators, and metrics. Due to the fast evolving of this field, we will maintain a living GitHub repository (links are in each section) to crowd-source updates and corrections, and warmly invite contributors and collaborators.

**Index Terms**—Large language model, evaluation, benchmark, survey



## 1 INTRODUCTION

Large Language Models (LLMs) have achieved unprecedented success in both academia and industry, largely attributed to the rapid advancements in training and evaluation techniques. As the “quality-control system”, evaluation not only guides the trajectory of technological progress but also serves as an early-warning mechanism for potential risks. Recent reasoning LLMs like OpenAI o1 or DeepSeek-R1 further underscore this importance of evaluation — by integrating reasoning, evaluation, and subsequent re-reasoning (i.e., refinement or correction) into a single Chain-of-Thought (CoT), their inference quality got greatly improved. These advances have invigorated the evaluation community, producing an ever-expanding array of benchmarks and assessment studies. To keep pace with this rapid growth, our survey goes beyond mere cataloging or facet-specific reviews. Instead, we delve into the fundamental challenges by examining how the advent of LLMs has reshaped the evaluation landscape, a phenomenon we term the **evaluation generalization**.

Upon reviewing current research in this area, we identify two critical transitions. As shown in Figure 1, one transition in evaluation is *from task-specific to capability-based*. Traditional evaluation methods focused on specific tasks (e.g., text classification, information extraction). As LLMs unify various NLP tasks in the same form of natural language generation, the definition of each task and the boundaries

between them has become increasingly blurred. In this new paradigm, each instruction or prompt can be viewed as an individual task, shifting attention toward assessing the core capabilities needed to tackle real-world needs. In this survey, we identify five key capabilities: knowledge, reasoning, instruction following<sup>1</sup>, multi-modal understanding, and safety. In Section 2, we survey existing benchmarks and categorize them within this capability framework, further dividing them into more detailed sub-categories. In addition, we discuss comprehensive evaluations that assess the interplay between different capabilities and current live leaderboards. This shift from task-based to capability-based evaluation enables a comprehensive understanding of a model’s true potential, beyond its performance in predefined tasks.

Another transition in evaluation is *from manual to automated* methods, including data curation and judgment. On the data side, rapidly evolving model performance demands increasingly frequent benchmark updates and manual curation processes have become unsustainable, highlighted by the accuracy surge on GSM8K (Grade School Math 8K) from 74% to 95% within two years. Automated pipelines can address both the cost and efficiency challenges inherent in dataset creation. Another benefit of automation is its potential to mitigate data contamination, where test data are inadvertently exposed during pre-training or post-training, leading to overestimated performance. In response,

1. Conventional NLP tasks are considered part of instruction following.

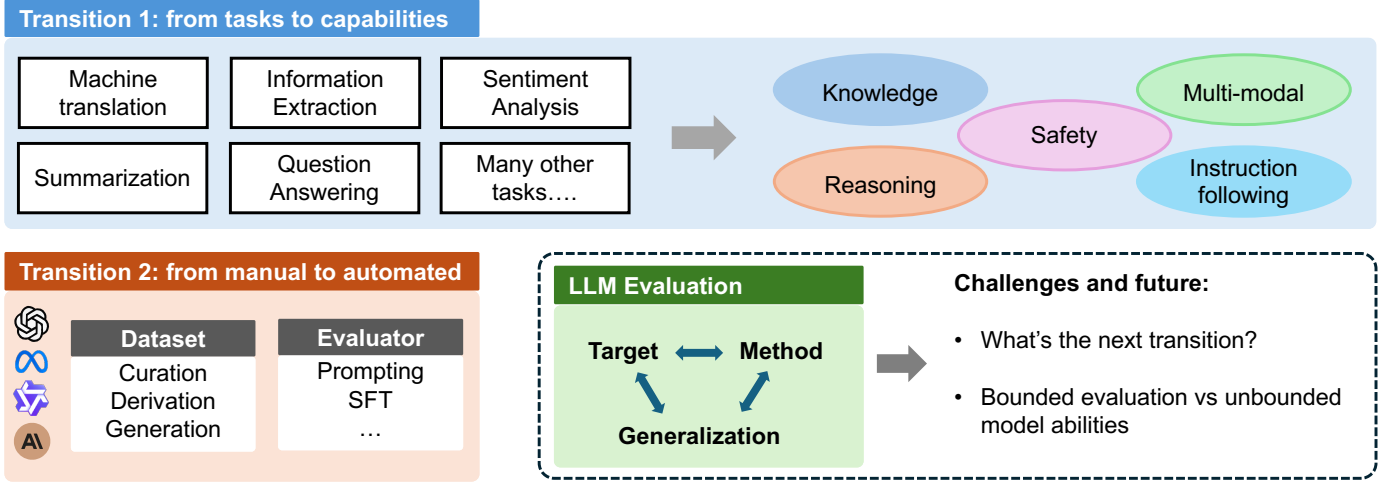


Fig. 1. Illustration of two transitions in the field of evaluation for LLMs.

automated approaches can be one of the solutions, which continually updates or refines test sets, known as dynamic benchmarks, ensuring that no test data are seen in advance. On the judgment side, as mentioned above, the shift to user prompts brings more open-ended responses, which pose further complexities: human judgment is expensive. Automated evaluators (i.e., “LLMs-as-a-judge”) not only show promise in providing reliable, efficient assessments but also can produce more detailed, fine-grained evaluations of human-like responses. In Section 3, we provide a comprehensive survey of these automated methods.

Although researchers have made significant progress along the two transitions outlined above, we argue that a fundamental contradiction persists between the training paradigm implied by scaling laws and the bounded evaluation practice. As model parameters, training FLOPs and data increase, the performance can be improved seemingly without bound. However, evaluation datasets cannot be expanded or diversified unbounded in practice considering the efficiency. That being said, current evaluation pipeline do not scale in tandem with model capabilities. The result is a growing mismatch between what models can do and what our tests can cover. This tension underlies many known challenges in LLM evaluation. Take data contamination as an example, because the limited testing dataset can cover only a subset of a model’s capabilities, different models may gain heterogeneous advantages during evaluation, leading to unfair comparisons. That is, if a model has encountered and memorized the test samples during training, its measured abilities will align perfectly with what the dataset evaluates, granting it an outsized edge that does not necessarily reflect stronger true capabilities.

We designate the above problem — how to leverage a bounded evaluation pipeline to assess an unbounded model capacity — as the evaluation generalization issue. In other words, existing evaluation tend to concentrate on capabilities that models already exhibit or that can be expressed by a fixed test set, inherently limiting the scope. Thus, the core challenge of evaluation in the era of LLMs is to develop generalizable evaluation methods capable of anticipating future or unexpressed abilities. In this survey, we

examine this challenge from different perspectives: datasets, evaluators, and metrics, and explore potential solutions. For example, some work focuses on predictive evaluation that carefully curates various tasks to estimate the performance of larger scale models based on that of smaller ones [1]. Or, Cao et. al. [2] propose to combine performance and a new interpretability-based metric, Model Utilization Index (MUI), for evaluating the potential of LLMs beyond the given datasets. The basic idea mirrors human assessment practices: when judging an individual’s overall ability, we consider both the result and the effort required (i.e., MUI) — less effort for equal performance denotes greater proficiency.

It is important to acknowledge that LLM evaluation is a rapidly evolving field. While we have endeavored to catalog the latest work on text-centric evaluations, many studies remain at the preprint stage. Consequently, our emphasis here is on forward-looking insights and research directions. Inevitably, some omissions or inaccuracies may occur. We plan to maintain a dedicated GitHub repository and invite the community to help us for refinement; major contributors will be gratefully acknowledged or invited as collaborators.

## 2 CORE CAPABILITIES AND DATASETS

As LLMs unify a wide range of tasks, the first type transition is from task-specific to capability-based assessment. In this section, we first discuss five core capabilities: knowledge, reasoning, instruction following, multimodal, and safety, with corresponding datasets, followed by their intersections and current live leaderboards. An illustration is shown in Figure 6. The Github page we will maintain and welcome any collaborators is <https://github.com/ALEX-nlp/Benchmark-of-core-capabilities/tree/main>.

### 2.1 Knowledge Evaluation

Knowledge evaluation determines the models’ ability to accurately recall, understand, and utilize factual information or human priors. Ensuring that LLMs possess a robust and reliable knowledge base is crucial for applications where precision and correctness are paramount. For example, it

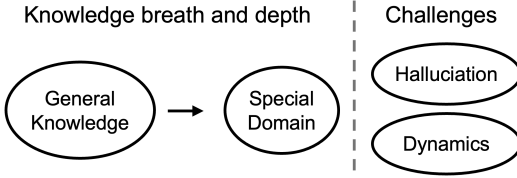


Fig. 2. The logic of reviewing knowledge evaluation.

focuses on knowledge-intensive questions such as *Who is the president of the United States?* or *What is the capital of France?*

In the early stage, benchmarks were primarily designed to assess the breadth of world knowledge in LLMs [3]. These benchmarks focused on general knowledge derived from sources such as Wikipedia, ConceptNet, and other knowledge bases, typically adopting a question-answering format—sometimes accompanied by supporting documents. Representative examples include TRIVIAQA [4], NATURALQUESTIONS [5], WEBQUESTIONS [6], and COMMONSENSEQA [7]. With the advent of LLM scaling laws, vast pre-training corpora have endowed these models with an extensive repository of general knowledge. Consequently, research attention has increasingly shifted from evaluating the breadth to probing the depth of domain-specific expertise in areas such as finance [8–10] and law [11–13].

Besides the breath and depth, knowledge evaluation also faces two major challenges. First, when it comes to unfamiliar or conflicting knowledge, LLMs may not admit their lack of understanding as humans do; instead, they may fabricate information, resulting in hallucinations. Worse still, as LLMs continue to evolve, they might even learn false information from the Internet. Consequently, LLMs could generate a large number of incorrect answers that are deceptive and potentially misleading to humans, which requires serious attention. To address this issue, TRUTHFULQA [14] collected a set of well-known false claims or misconceptions, while HALUEVAL [15] curated questions that have no answers or are impossible to answer, requiring the model to point out that the question is unanswerable instead of generating a fake answer.

The second challenge concerns the dynamic nature of knowledge. Early datasets emphasized the timeliness and chronological order of knowledge [16–21], whereas later datasets focused more on addressing the data contamination issue using the latest knowledge from News articles [22], Wikipedia [23, 24]. The motivation is to accurately assess model advancements, evaluation datasets need to be continuously updated to prevent false negatives caused by outdated information. Moreover, some scholars argue that consistently updating data can prevent performance overestimation due to data contamination, since as long as the evaluation data pertains to the latest knowledge, the model would not have been exposed to it, thereby eliminating data contamination issues. However, other scholars point out that the risk lies in the difficulty of completely distinguishing new from old knowledge based on a specific cutoff date (such as the model’s release date). For instance, even if a movie is released after this date and the model should not have seen it, necessary information might have already been exposed to the model through early promotions and related

activities [23]. We will detail dynamic datasets in Section 3.1.

## 2.2 Reasoning Evaluation

Reasoning is a core component of intelligence in applying logic by drawing valid conclusions from new or existing information. Its evaluation is the key to gauge the true cognitive abilities of a model, such as problem solving, decision-making and human-like thought process. However, reasoning cannot be fully evaluated from a single perspective. Instead, researchers have developed methods to assess reasoning across multiple dimensions. In the following, we highlight several key domains: mathematics, coding, commonsense, long-context understanding, logic, planning, and miscellaneous tasks.

### 2.2.1 Mathematics Evaluation

Mathematical reasoning represents one of the most rigorously scrutinized aspects of reasoning evaluation. Its structured and precise reasoning process facilitates straightforward assessment. Furthermore, mathematics, as a cornerstone of abstract thought, is indispensable in scientific research, engineering, and related fields. As illustrated in Figure 3, mathematics benchmarks have evolved alongside the advancements in LLM capabilities, progressing from primary school-level problems to challenges of Olympiad-level difficulty. In particular, before 2021, mathematics datasets mainly focused on primary school-level problems, reflecting the limited capabilities of language models at that time [25–27], where GSM8K [28] is still widely used. After 2021, research efforts shifted toward high school [29] or university-level problems [30]. Example datasets include MATH [31], which comprises 12,500 advanced high school math competition problems annotated with five difficulty levels. Since 2024, the rapid advancement of LLMs has prompted researchers to further escalate the difficulty of benchmarks to the Olympiad contest level, aiming to extend the boundaries of these models [32, 33]. Currently, the most difficult dataset is FrontierMath [34] crafted by expert mathematicians, covering major branches of modern mathematics — from computational number theory to abstract algebraic geometry — and often requiring hours or even days for specialists to solve. Even the most advanced reasoning LLMs like OpenAI o1 can only achieve around 3% accuracy.

### 2.2.2 Coding Evaluation

Coding is another widely used reasoning evaluation task. Compared with mathematics, its reasoning process (i.e., code snippets) is also highly rigorous yet holds significant practical applications. A variety of benchmarks have been introduced to evaluate LLMs’ capabilities in code understanding and generation. We provide a high-level categorization of these benchmarks by programming languages and the primary coding tasks evaluated in Figure 4. Observe that most benchmarks concentrate on Python and code generation task, given its wide adoption in both industry and academia [35–40]. To further evaluate cross-lingual capabilities, several benchmarks feature tasks in multiple programming languages [41–44]. Beyond language diversity, some benchmarks explicitly focus on different aspects of

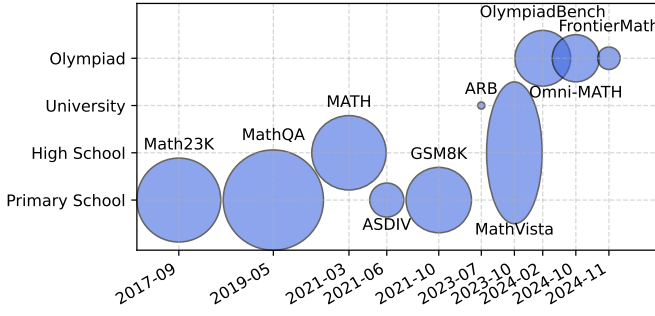


Fig. 3. Mathematics benchmarks in chronological order. The y-axis represents four difficulty levels: Primary School, High School, University, and Olympiad. The area represents the size of each dataset.

the software development life-cycle, including debugging, clone detection, defect detection, code completion, code-to-code translation, and requirement switching [45, 46].

Except for precise evaluation criteria, coding tasks also have strong practical values, which are becoming a high-visibility benchmark for LLM reasoning. Nevertheless, they remain especially vulnerable to data contamination — vast repositories of public code are ingested during pre-training. To mitigate this, Livecodebench [47] continuously ingests newly released problems from coding competitions on platforms such as LeetCode, AtCoder, and Codeforces. By annotating each problem with its official release date, Livecodebench ensures that test items were unavailable during a model’s pre-training period, effectively preventing contamination and overfitting and yielding a trustworthy, time-aware assessment of coding performance.

### 2.2.3 Logic Reasoning

Another group of reasoning task is logic reasoning, usually involving three types: deduction (drawing conclusions), induction (recognizing patterns), and abductive reasoning (forming explanations). Logic reasoning is similar with math or coding in its well-defined nature, yet focuses on domain-independent inference patterns. This means that a model is expected to follow a structured reasoning process given the information at hand without any prior knowledge. Therefore, when benchmarking logic reasoning, except for real-world scenario, many attempts build a confining setting to minimize the unfair advantage of accumulated knowledge or learned information.

Deductive reasoning proceeds from general premises or rules to a guaranteed specific conclusion. If all premises are true and the logical steps are valid, the conclusion must be true, as in classical syllogisms or formal proofs. For example, given premises “All birds can fly” and “Magpie is a bird”, a deductive model infers “Magpie can fly”. In the context of LLM evaluation, deductive benchmarks often involve determining whether a hypothesis holds true or false from provided premises [48] or producing a step-by-step proof [49]. We can see that such tasks require the model to carry out multi-step logical derivations without introducing outside knowledge, while there are also other studies that curate probing benchmarks across different domains

towards practical values, such as everyday situations [50] or exams [51, 52].

Inductive reasoning is essentially the inverse of deduction: it draws general conclusions or rules from specific observations or instances. Here, the inference is probabilistic rather than certain — the conclusion goes beyond the information provided. For example, given observations “Magpie is a bird” and “Magpie can fly”, an inductive model may hypothesize “All birds can fly”, which could later be proven wrong by a counterexample (“Ostrich cannot fly”). Clear, the more the observations, the higher probability the inferred hypothesis holds true. This type of reasoning is easily influenced by prior knowledge. Thus, ARC-AGI benchmark [53] only assume core knowledge priors (“cognitive building blocks that are either present at birth or acquired very early in human development with minimal explicit instruction”) and design problems in a formal setting: given a set of input-output examples specifying some behavior for recognition, which is further simplified to 1-D pixel pattern in images [54]. Similarly, syntax-guided synthesis (SyGuS) [55] setup the task based on string transformation, and CLUTTR [56] focuses on relational logic in narratives.

Abductive reasoning, also known as explanatory reasoning, involves generating the most plausible explanation for a given set of observations or facts. Clearly, certainty is not guaranteed. The proposed explanation is a guess that could be wrong, but unlike induction, the goal of abductive reasoning is not a general rule but rather a specific hypothesis that accounts for the data. For example, given the observation “The road is wet”, we may guess “it probably rained recently”. Such guessing heavily relies on the experience, so benchmarks for abductive reasoning often requires commonsense (which will be detailed next section) and an understanding of likely causal chains in everyday scenarios. For example,  $\alpha$ NLI [57] select the task of story completion and targets the more plausible connective explanation for how the characters got from the start to that end. To further challenge the deep abductive reasoning capabilities, True Detective benchmark [58] setup the questions in murder-mystery narratives to ask who the crime is or what explains the mystery. Considering the impacts of prior knowledge, there are also attempts like AbductionRules [59], which constructs synthetic logic puzzles for abduction. It presents a knowledge base of facts and rules (expressed in natural language) along with an “unexpected” observation, and the task is to hypothesize a missing fact or rule that would explain the observation.

Some recent surveys [60, 61] and comprehensive benchmarks [62] target all three types of reasoning. An interesting finding is that LLMs often perform the best for abductive reasoning and the worse for inductive reasoning. But it is still challenging how to design benchmarks that truly measure reasoning and not just language proficiency or shallow pattern matching. As mentioned above, one proposal is to abstract away rich semantic content in tasks, so that an LLM’s performance reflects its grasp of reasoning structure rather than any prior knowledge. For example, using arbitrary symbols or “neutral” facts prevents the model from relying on memorized world knowledge, forcing it to rely on pure logic [63]. Indeed, models achieved high scores by learning the formal patterns, but such content-abstracted



benchmarks have limits: they risk oversimplifying language understanding and may introduce unnatural regularities that models can exploit but that don’t translate to real-world reasoning [61]. On the other hand, benchmarks within some domains like science exams or detective stories ensure that models must deal with realistic language and background knowledge, but then it becomes harder to disentangle logical reasoning from domain knowledge. The field is grappling with this balance between symbolic abstraction and natural complexity when evaluating reasoning. Besides, as the True Detective [58] results indicate, scaling reasoning to long contexts or more complex problems is still an open problem. Future benchmarks will likely need to push beyond toy tasks and short paragraphs, testing whether LLMs can maintain logical coherence over extended reasoning chains or in interactive, multi-turn settings.

#### 2.2.4 Commonsense Reasoning

Commonsense reasoning refers to the fundamental level of practical knowledge and reasoning about everyday situations and events that is widely shared among people. Sometimes it adopts the same form of logic reasoning with commonsense knowledge. Still, it is essential not only for humans to navigate daily life and interact with one another but also for artificial intelligence (AI) systems to better understand human needs and actions. In terms of scenario, we can roughly categorize the evaluation into three domains: social, temporal, and physical commonsense. 1) Social commonsense involves understanding interpersonal interactions and human behavior. Representative datasets in this category include Naive Psychology [64], ROCStories [65], Social IQa, the Winograd Schema Challenge (WSC) [66], Choice of Plausible Alternatives (COPA) [67], VCR (visual commonsense reasoning) [68], and e-CARE [69] (explainable commonsense). 2) Temporal commonsense pertains to the sequencing of events, causality, and time-related inferences like duration, frequency, or ordering. Key datasets here include MCTaco [70], UDS-T [71], and MavenERE [72]. 3) Finally, physical commonsense encompasses fundamental knowledge about the physical world, including object properties and spatial relationships, such as Physical IQa [73], HellaSwag [74], Abductive NLI [57], SWAG [75], CommonsenseQA [7], and JHU Ordinal Commonsense [76].

Besides datasets, there are also many commonsense resources available for both training and evaluation purposes. Early projects were primarily developed by human experts including Cyc [77] and OpenCyc [78]. These systems encoded ontological relationships between objects using formal logic, categorizing entities into types such as entities, sets, functions, and truth functions, and contained a wealth of commonsense assertions. Concurrently, a team at the MIT Media Lab developed the Open Mind Common Sense project [79], later evolving into ConceptNet [80]. This project harvested online data and integrated diverse knowledge sources. The latest version, ConceptNet 5.5, employs automated extraction techniques and comprises over eight million nodes and more than twenty-one million links, incorporating multilingual resources as well as connections to other knowledge graphs. Among those automatically curated resources [81, 82], a notable one is ATOMIC [83], a crowd-sourced knowledge graph that features textual descrip-

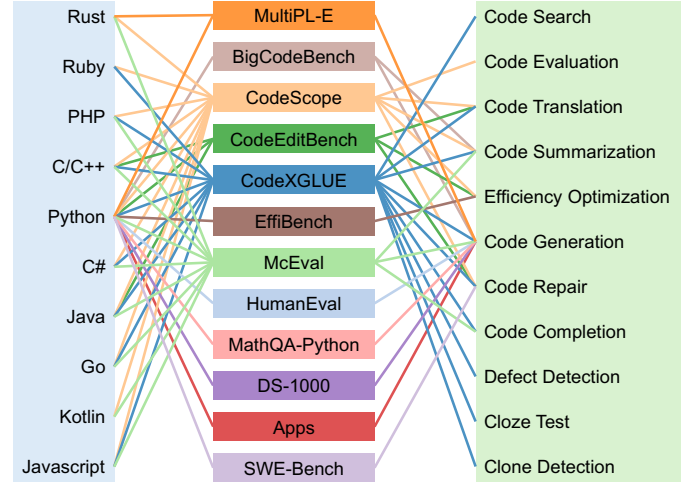


Fig. 4. Coding benchmarks, the programming languages (left), and coding tasks (right) that each benchmark use. We assign one color to each benchmark.

tions for around 300,000 event nodes and approximately 877,000 “if-event-then” triplets, capturing nine distinct types of causal relationships between everyday events. Building upon these foundations, the VisualComet project [84] extended the realm of commonsense reasoning into the multi-modal domain by proposing Visual Commonsense Graphs. Additionally, there are domain-specific or purpose-oriented commonsense reasoning resources, such as for sentiment analysis [85], causal reasoning [86], and e-commerce intention [87].

Overall, commonsense reasoning has a long research heritage, supported by extensive resources and benchmark datasets. More recently, the emergence of LLMs exhibiting super-human performance on certain commonsense tasks. However, this does not imply that the challenge has been completely resolved. In real-world applications, LLMs still lag behind human capabilities when it comes to complex commonsense reasoning, especially in multimodal tasks. Two core challenges underlie this gap: 1) commonsense knowledge is never explicitly stated in text, images or other modalities, which hampers acquirement or robust reasoning; 2) Unlike mathematics or coding, commonsense tasks do not possess clear formal structure or single “correct” answers. This ambiguity complicates both the construction of datasets and the accurate evaluation of open-ended responses. To conclude, commonsense task may be the key in extending reasoning evaluation from clear, well-formed tasks to more ambiguous, open-world problems in the near future.

#### 2.2.5 Long Context Reasoning

Text remains the primary medium for interacting with LLMs, and knowledge can be embedded in long texts in diverse forms. As a result, the ability to reason over extended contexts has become a critical capability. Long-context reasoning refers to increasing the input length that LLMs can process, while ensuring they can effectively understand, learn from, and reason over the information contained in these longer inputs.

We classify it as reasoning because, compared to methods like retrieval-augmented generation (RAG), which are more suited for extractive or localized tasks, long-context reasoning excels in scenarios where the model needs to perform global reasoning by leveraging all the input information. Also, this is highly significant in practical applications, as it serves both as an important means for integrating external knowledge, recording historical behaviors or interactions, and following complex instructions [88].

To propel this line of work, long-context LLMs have adopted techniques such as interpolation [89], extrapolation, fine-tuning, and architectural optimizations to rapidly extend their supported token windows. Correspondingly, evaluation benchmarks continuously raise both the maximum input sequence length and the complexity of tasks, to ensure that assessment keeps pace with ever-growing model capabilities. At first, L-Eval [90] and LongBench [91] are at a moderate scale — contexts from roughly 3K to 60K tokens, including tasks like single-document QA, multi-document QA, summarization, and code completion. Then, InfinityBench [92] steps up to ultra-long contexts (about 100K tokens) at domains such as novel and coding. Recently, LongBench v2 [93] is at the extreme frontier, which pushes to 2 million words of context across 503 questions in six categories: single- and multi-document QA, extended in-context learning, long-dialogue comprehension, large code-repository understanding, and structured-data reasoning. This benchmark emphasizes deep logical inference, cross-document linking, and structured-data extraction at unprecedented scale.

Although there are many efforts mentioned above, the definition of “context” is still not clear enough [94], which dictates both dataset structure and evaluation focus when designing long-context benchmarks. If context is in the form of a single coherent document, such as a novel or research paper, the benchmark must ensure tight question–passage alignment and test a model’s ability to integrate clues spread across multiple sections. These datasets probe deep reading comprehension and multi-paragraph reasoning but are costly to curate, subject to copyright constraints, and slow to refresh. Conversely, if context is provided as an artificial concatenation of shorter excerpts, e.g., simply grouping Wikipedia articles together, benchmarks can be scaled quickly. Yet the relevance between questions and information becomes uneven, and the task shifts toward retrieving salient facts amid noise, exposing failures such as position bias or the “lost-in-the-middle” effect.

This dual interpretation poses three core challenges for evaluation. First, benchmarks must strike a balance between reading and retrieval skills, ensuring that neither devolves into trivial keyword matching nor pure long-span memorization. Second, they require reliable metrics of question–context relevance; without such controls, high scores may reflect chance matches rather than genuine understanding. Third, as token windows expand, benchmarks must evolve dynamically: coherent long texts are hard to source continuously, while concatenated corpora risk overlapping with pre-training data, demanding fine-grained de-duplication and release-date tagging to prevent contamination. Addressing these challenges is essential for keeping long-context evaluation both realistic and forward-looking.

On the other hand, while long-context reasoning can be partially reflected in benchmark results, the specific reasoning capabilities assessed may be domain-specific, depending on the nature of the documents used in the benchmarks.

### 2.2.6 Planning

Planning is a special type of reasoning. Instead of inferring new knowledge from existing ones, it aims at decompose high-level objectives to fine-grained, relatively simple steps. Due to the task’s complexity, this process usually needs to combine various reasoning skills, setting a high bar for the model. Nevertheless, planning is the key for models dealing with dynamic and complex tasks then stepping from simple textual contexts, to virtual environments and to physical worlds. According to the model’s working environments, we examine three dimensions of planning benchmarks: (1) *task planning* for textual goal decomposition, (2) *agent planning* for autonomous decision-making in virtual, interactive environments, and (3) *embodied agent planning* that integrates physical interaction with spatial reasoning.

Textual task planning focuses on generating structured sequences of steps to achieve specified goals, often requiring hierarchical decomposition. Early work in Goal-Oriented Script Construction (GOSC) [95] established baselines using the WikiHow [96] benchmark for step sequence generation, later extended by Instructables [97] to incorporate hierarchical subgoals. Subsequent benchmarks like PlanBench [98] systematically evaluate validity and cost-optimality of generated plans, revealing significant gaps between LLM capabilities and human reasoning. TaskBench [99] introduces tool invocation graphs to assess execution consistency alongside planning precision. Real-world applications are explored through Natural Plan [100] for trip scheduling and meeting coordination, WorkBench [101] for digital workplace task management, and UltraTool [102] for end-to-end tool utilization in complex problem-solving scenarios.

Virtual agent planning evaluates autonomous systems’ capacity to navigate in simulated or digital environments. WebShop [103] targets e-commerce simulations requiring complex query interpretation and purchase optimization across 1.18 million real products. TravelPlanner [104] chooses requiring agents to balance budget, logistics, and commonsense constraints while coordinating multiple information sources for travel itineraries. SmartPlay [105] tests adaptive reasoning through six game environments requiring spatial and strategic planning. Theoretical foundations are strengthened by the tri-modal evaluation framework (autonomous/heuristic/human-in-the-loop) and the benchmark for action reasoning and plan reuse [106, 107]. Robotic integration is pioneered by SayCan [108], which grounds planning in physical affordances for real-world mobile manipulation.

Embodied agent planning bridges digital reasoning with physical world, demanding tight integration of perception, environment interaction, and actions. According to the key components, we discuss the evaluation from two perspectives of environment and planning types. In terms of environments, embodied planning benchmarks span from abstract symbolic worlds to high-fidelity 3D simulations. For symbolic or text-based virtual environments, which is different from the aforementioned task planning and virtual

agent planning, they leverage descriptive language or abstract state representations for rooms and objects instead of seeing pixels [109]. For 3D simulations, many benchmarks adopt first-person view based on simulators like Habitat [110] and iGibson [111] and an agent must interact with objects or navigate spaces. This type of evaluation emphasizes photorealism and physics, featuring realistic lighting, textures, and physical object dynamics. Examples include ALFRED [112] and BEHAVIOR [113]. By combining both, ALFWorld [114] is a hybrid platform that aligns a text-based world with the 3D tasks from ALFRED, allowing agents to practice in a simplified symbolic setting before transferring to a realistic simulator.

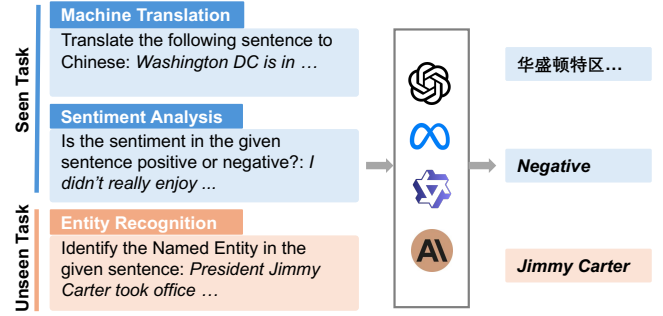
In terms of planning types, different benchmarks demand different levels of planning granularity. Low-level action planning requires sequences of fine-grained actions (navigation steps, motor primitives). For example, an agent needs to plan a path through a 3D scene, issuing low-level motions (forward, turn) to reach a target coordinate or object [110]. This is often framed as visual navigation and tests short-term planning and obstacle avoidance, albeit potentially over long distances. In contrast, high-level task planning involves deciding on a sequence of sub-tasks or goals to satisfy an overall objective, e.g., “clean the coffee cup and put it back”. Examples datasets include ALFRED [112] and BEHAVIOR [113]. Although high-level tasks have achieved promising results, some studies argue the potential overestimation and delve into single-step planning [115, 116]. Their analysis reveals both types of evaluation are important since notable drawbacks like numerical comprehension, heavy selective biases over directional concepts, or recurrent issues, still exist and may be critical when transferring to real world.

### 2.2.7 Miscellaneous Reasoning

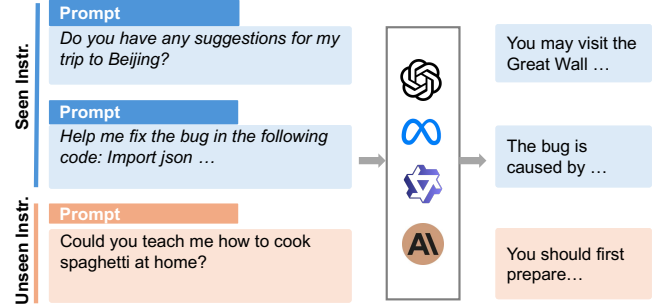
Apart from the aforementioned reasoning tasks, there exists a diverse range of reasoning skills that we collectively refer to as miscellaneous reasoning. Example symbolic reasoning tasks include coin flip reasoning and last letter concatenation [117]. The basic idea is to define a set of transformation rules, and the model is required to apply these rules systematically to infer the correct outcome given an initial state. Similarly, there are also some common IQ test puzzles and algorithmic problems, such as classic puzzles like the tiger-eats-sheep problem or the gold division problem. Clearly, these tasks are in-between logical reasoning or instruction-following.

Other examples include visual reasoning and designs tasks defined in [118]. These tasks include counting line intersections, determining the relationship between two circles, identifying the circled letter, counting overlapping shapes, counting nested squares, counting the rows and columns of a grid, and following single-colored paths. Though simple, most vision language models (VLMs) perform unsatisfactory.

Spatial reasoning also attracts many research attention. The work [119] defines 2D and 3D trajectory labeling and relationship identification. This work employs the CALVIN benchmark [120], which assesses LLMs in long-horizon, language-conditioned robotic manipulation tasks.



(a) Multi-task generalization.



(b) Prompt generalization.

Fig. 5. Illustration of instruction following’s paradigm shift: from tasks, where tasks are gathered and their descriptions are used as instructions, to user needs where user prompts are regarded as “fine-grained” tasks’ instructions.

## 2.3 Instruction Following

Instruction following aims to assess whether models can comprehend human inputs and provide appropriate responses. As model capabilities have advanced, this evaluation method has progressively evolved from conventional NLP tasks to diverse human needs (Figure 5).

In the early stages, models were limited to performing single tasks, with training focused on mapping inputs to outputs on specific datasets. To enhance generalization, multi-task learning was introduced, shifting model predictions from  $p(y|x)$  to  $p(y|x, \text{task description})$  [121]. Consequently, researchers began aggregating diverse tasks and crafting detailed task descriptions, resulting in multi-task fine-tuning datasets collectively known as instruction tuning. Held-out tasks not included in the training set were then used for evaluation. Early instruction following evaluations thus centered on traditional NLP problems, such as question answering, question generation, and text classification. Representative benchmarks in this phase included NATURAL INSTRUCTIONS [122], TO-EVAL [123], INSTRUCT-EVAL [124], FLAN [125] and SUPERNI-TEST [126].

As more tasks are included to enhance generalization, the goal of instruction tuning shifts to real-world human needs, from which modern NLP tasks are derived. Thus, task descriptions can be viewed as scientific definitions of those needs. Researchers then began collecting a diverse array of real-world user prompts, moving beyond the confines of specific tasks, such as SHAREGPT [127], FREEDOLLY [128], OPENASSISTANT [129]. The release of ChatGPT further accelerated this trend, as LLMs were

integrated into online service environments that allow users to issue a wide variety of instructions framed as open-world “tasks”. A representative example is CHATBOT ARENA [130], renowned for its Elo rating score system. In this platform, users provide instructions to chatbots, and two randomly selected LLMs generate responses for direct comparison. Human annotators then select the better response in real time, with each outcome contributing to an evolving Elo rating that dynamically quantifies the relative performance of the models—thereby more closely mirroring real-world usage scenarios.

This shift from task-specific evaluations to user-driven needs ensures that models can handle a wide variety of instructions and respond effectively to the multifaceted demands encountered in practical applications. However, it also introduces significant evaluation challenges: responses become more open-ended and non-structured, making it difficult to achieve reliable and consistent scores through human evaluation, which is both resource-intensive and time-consuming. Consequently, researchers have begun exploring automated evaluation methods to improve efficiency and scalability. For example, ALPACAEVAL [131] and VICUNAEVAL [132] have experimented with using LLMs to score the quality of responses. These studies found that LLMs can not only generate relatively consistent scores but also provide detailed explanations for those scores. Furthermore, benchmarks such as ARENA HARD AUTO [133] have been developed. In these benchmarks, user instructions collected from online environments are evaluated in a pairwise manner — similar to CHATBOT ARENA — where two LLMs generate responses to the same instruction. The key difference is that evaluations are conducted by a powerful LLM, such as GPT-4, rather than human annotators, thereby improving scalability and efficiency [130].

Nevertheless, this automated approach has faced criticism for introducing potential biases inherent in the judging LLMs. Such biases, including preferences for verbosity or specific response styles, may lead to unfair evaluation outcomes [134–137]. To mitigate these issues, style-control variants such as STYLECONTROL ARENA [134] and LENGTH-CONTROLLED ALPACAEVAL [136] have been introduced. These benchmarks seek to disentangle stylistic factors from the substantive content of responses, enabling fairer comparisons between LLM outputs. A more detailed discussion of automated evaluation methods will be presented in the Section 3.

Although pairwise comparison benchmarks are valuable for assessing relative performance, they do not provide fine-grained scores for specific capabilities. To address this limitation, a new class of benchmarks has been developed to evaluate instruction-following ability in an absolute and fine-grained manner [138–140]. For instance, IFEVAL [138] introduced 25 rule-verifiable constraints (e.g., Output your response in all uppercase letters, within 10 words, without using the word “I”, etc.), requiring LLMs to generate responses that satisfy these constraints. FOLLOWBENCH [139] extended this idea by expanding the constraints from rule-verifiable to model-verifiable, wherein another strong LLM is tasked with verifying whether the generated responses meet the constraints. WILDBENCH [140] further advanced this approach by providing a human-annotated checklist for

each instruction, with a judge LLM verifying whether the generated response satisfies the checklist — thus adding an additional layer of human oversight.

## 2.4 Multi-modal Evaluation

Multimodal evaluation measures the ability of LLMs to process different data modalities beyond text, such as audio or tabular data. This capability enhances the versatility of AI models in real-world applications. Several survey papers [141–143] have focused on evaluations in this domain; therefore, we select visual information as a complementary aspect to the text-based capability assessments discussed above. Below are some representative tasks.

### 2.4.1 Visual Question Answering

Visual Question Answering (VQA) is to answer questions based on both textual and visual information. We start with basic visual perception tasks, like RealWorldQA [144] that evaluate real-world spatial understanding including counting, identifying, and locating objects in images. These tasks are easy for humans but still challenging for models. To further assess cognitive abilities, MME [141] designs reasoning, coding, and planning tasks. MMT-BENCH [145] dives deeper by decomposing visual abilities into 32 meta-abilities (e.g., counting, locating, and identifying) and constructing a comprehensive benchmarks. Except for perception and cognition, MMMU [146] and MMMU-Pro [146] curated massive multi-discipline tasks demanding college-level subject knowledge. While, recent works have shown that many MMMU samples could be answered without visual information, this raises concerns about uni-modal bias [147]. To address this issue, MMSTAR [148], a vision-indispensable multi-modal benchmark, was proposed. Each sample in MMSTAR is verified by human to ensure the visual content is essential to answer the question. Furthermore, hallucination and long-tail issue are also considered in MMBENCH [149] and HallusionBench [150], respectively.

### 2.4.2 Visual Document Comprehension

Visual Document Comprehension regards document, including text, tables, and diagrams, etc., as visual inputs (e.g., images) for the following tasks. Compared with textual document understanding, this type of methods enjoy an efficient end-to-end manner and can achieve maximum retention of information. For example, there is no additional step to parse text from images and the layout is maintained. Therefore, benchmarks are curated to evaluate text understanding from document screenshot [151] or taken in the wild [152, 153], infographic text comprehension [151], multi-modal document understanding [154], and scientific diagram comprehension [155].

### 2.4.3 Multi-image Understanding

While earlier MLLMs are mostly trained to align single images with natural language components, one emergent capability that recent efforts seek to extend is multi-image understanding, or more broadly, interleaved processing of (multiple) images and texts. In this context, earlier benchmarks including multi-image examples typically focus on



specific scopes of reasoning and do not provide a comprehensive assessment [145, 156, 156–158]. Recent efforts assess MLLMs in multi-image scenarios. For example, MANTIS-Eval [159] is a human-annotated benchmark comprising 207 examples for multi-image reasoning, such as size perceptions and weight comparisons, while DEMON [160] evaluates whether MLLMs can follow zero-shot demonstrative instructions. A milestone benchmark for this challenge is MuirBench [161]. This comprehensive benchmark contains 11,264 images and 2,600 multiple-choice questions, evaluates on a range of 12 multi-image understanding abilities (e.g. geographic understanding, diagram understanding, visual retrieval, etc.) and 10 diverse multi-image relations (e.g. narrative, complementary, etc.).

Similarly, video understanding can be regarded as an extension of image understanding to a sequence of images, considering the temporal and spatial features among images. For example, MVBench [162] and PerceptionTest [163] evaluates general video comprehension. Clearly, along with the increasing video length, MLLMs is required to processing massive images within the context window like “visual long context reasoning” (visual version of Section 2.2.5). EgoSchema [164] and Video-MME [165] target the comprehension of long-term video up to one hour.

## 2.5 Safety

Along with the increasing capabilities of LLMs, their deployment raises serious safety concerns. Safety evaluation aims at assessing a model’s ability to avoid generating harmful, unethical, or biased outputs, ensuring its alignment with human values and societal norms. A recent survey [166] classified existing works into various attack and defense groups, including adversarial attacks/defenses, backdoor & poisoning attacks/defenses, jailbreak attacks/defenses, intellectual property protection, membership inference attacks, data extraction attacks, prompt injection attacks, etc. While, another survey [167] comprehensively introduce the open datasets and categorizes them into five main purposes: broad safety, narrowly defined safety, value integrity, bias, and other. Differently, our categorization is driven by analyzing evaluation trends and contains four directions: 1) content safety, 2) multi-dimensional trustworthiness, 3) adversarial robustness, and 4) agentic safety.

### 2.5.1 Content Safety

At the most fundamental level, content-safety benchmarks probe whether an LLM can identify, refuse, or filter toxic, hateful, violent, or otherwise disallowed text under non-adversarial conditions. Evaluations appear in three formats. The first one adopts single-sentence classification, e.g., ToxiGen [168] includes 274,000 machine-generated statements targeting 13 minority groups, each labeled as either toxic or benign. Second, recent studies, like RealToxicityPrompts [169], ToxicChat [170], BeaverTails [171], and DiaSafety [172], mimic the settings in real-world applications, which collects prompt-response pairs. Thus, evaluation can either treat it as text classification, the same as the first format above, or, third, feed the testing prompt into LLMs and leverage external tools to judge the newly generated response. Clearly, two major challenges lie in the design

of prompts and the performance of judge tools, which are the main focus of adversarial robustness as discussed later. In addition, there is a growing emphasis on multilingual content moderation or specific domains such as gender and sexuality [173]. To conclude, the key challenges for content safety benchmarks are two-fold: 1) the hate speech may be nuanced and contain no obvious slurs or profanity, which motivates ToxiGen [168] to design an adversarial classifier-in-the-loop generation process. 2) the hate speech should, as much as possible, originate from or closely resemble everyday life. Example datasets including RealToxicityPrompts [169] and DiaSafety [172] then collect data from real world like Reddit.

### 2.5.2 Multi-Dimensional Trustworthiness

LLM “safety” is not a single metric. Complementary to content toxicity or hatefulness as mentioned above, several recent benchmarks aim to evaluate LLMs holistically across multiple dimensions like bias [174]. DecodingTrust [175] assembles tests for eight different aspects including toxicity, stereotype bias, privacy, ethics, fairness, as well as adversarial and out-of-distribution robustness. HELM Safety<sup>2</sup> combines five benchmarks, covering six harm domains: violence, sexual content, harassment, self-harm, deception, and discrimination, and draws on specialized sub-benchmarks for each. The AegisSafety dataset [176] define a broad taxonomy of 13 critical risk and 9 sparse risk categories. SorryBench [177] spans 45 fine-grained safety categories targeting refusal behaviors. Meanwhile, it includes multilingual variations, which is also highlighted by XSafety [178] and S-Eval [179].

In terms of evaluation format, most benchmarks follows similar settings with those for content safety and adopt multi-choice questions, such as SafetyBench [180], DecodingTrust [175], SGBench [181]. To improve the difficulty levels, SALAD-Bench [182] introduces attack and defense methods to enhance the prompts categorized into 6 domains, 16 tasks, and 66 specific categories. While, CHiSafetyBench [183] designs a hierarchical benchmark across 5 risk areas and 31 categories to better organize the multiple safety dimensions. Except for structured tests, scenario-based tests are gaining popularity for practical values. The model is placed in a concrete situation and must take a stance or choose an action consistent with safety or ethics. For instance, the HHH benchmark [184] compares pairs of model outputs in different interaction scenarios and asks which response better aligns with ethical values: Helpfulness, Honesty, and Harmlessness. This format, using human preference judgments on model outputs, checks if the model can be simultaneously useful, truthful, and non-harmful. Another example is the ETHICS [185], which poses ethical dilemmas or scenario-based questions covering dimensions like justice, deontology, virtue ethics. For better understanding model’s safety, DoNotAnswer [186] provides an explanation for why a response should be refused, enabling evaluators to check not just if the model refuses, but whether it understood the risk.

A core challenge in multi-dimensional trust evaluation is coverage and scalability. While, curating tests for every

2. <https://crfm.stanford.edu/2024/11/08/helm-safety.html>

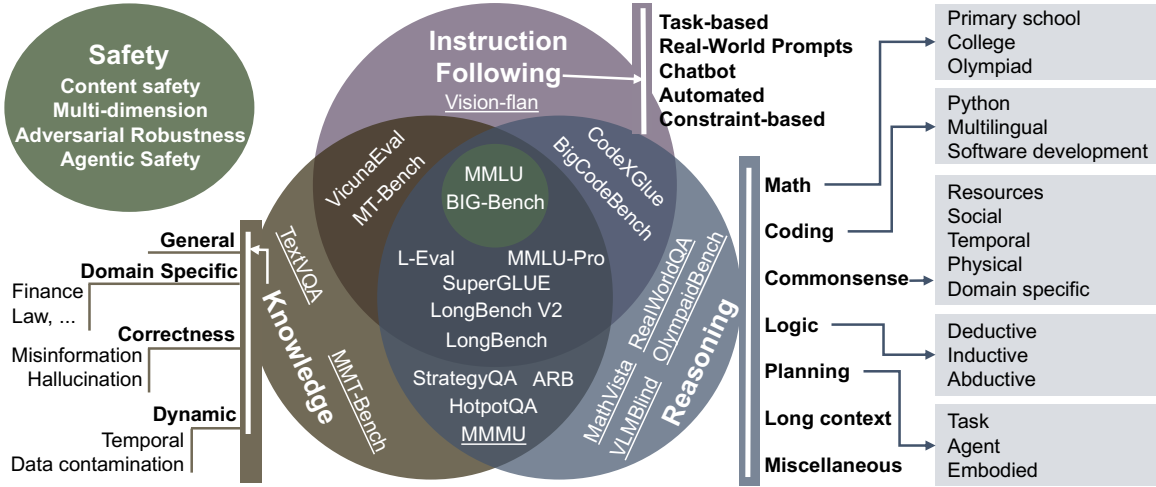


Fig. 6. Illustration of capability-based benchmark taxonomy involving: knowledge, reasoning, instruction following, multimodal, and safety. The interaction denotes the interplay of different abilities. Note that multi-modal benchmarks are marked with underline. The taxonomy and corresponding datasets are tentative and we will keep improve it.

TABLE 1  
Capability-based benchmark taxonomy.

<b>Knowledge</b>	<b>General</b>	[3–7]	<b>Reasoning</b>	<b>Commonsense:</b>		<b>Instruction Following</b>	<b>Task-based</b>	[121–126]
	<b>Domain:</b>			Resources	[77–80]		<b>Real-World Prompts</b>	[127–130]
	Finance	[8–10]		Social	[64–69]		<b>Automated</b>	[130–133]
	Law	[11–13]		Temporal	[70–72]		<b>Style-Control</b>	[134–137]
<b>Reasoning</b>	<b>Incorrect:</b>		<b>Reasoning</b>	Physical	[7, 57, 73–76]	<b>Safety</b>	<b>Constraint-based</b>	[138–140]
	Misinformation	[14]		Domain-Specific	[81–84]		<b>Content Safety</b>	[168–173]
	Hallucination	[15]		<b>Logic:</b>			<b>Multi-Dimension</b>	[174–188]
	<b>Dynamic:</b>			Deductive	[48–52]		<b>Adversarial Robustness</b>	[184, 189–206]
	Temporal	[16–21]		Inductive	[53–56]		<b>Agentic Safety</b>	[207–214]
	Data contamination	[22–24]		Abductive	[57–59]			
	<b>Math:</b>			<b>Planning:</b>				
	Primary school	[25–28]		Task	[95, 96, 98–102]			
	College	[29–31]		Agent	[103–108]			
	Olympiad	[32–34]		Embodied	[109–114]			
<b>Reasoning</b>	<b>Coding:</b>		<b>Reasoning</b>	<b>Long context</b>	[90–94]	<b>Safety</b>		
	Python	[35–40]		<b>Miscellaneous</b>	[117–120]			
	Multi-lingual	[41–44, 47]						
	Software Development	[45, 46]						

potential risk is labor-intensive. This has led to efforts to crowdsource<sup>3</sup> and automate scenario generation [3]. However, using LLMs to judge other LLMs can introduce error if not carefully validated [187], which will be further discussed in Section 3.4.2. Therefore, a clear trend in evaluation design for trustworthiness is moving beyond static question sets toward more interactive simulations. For example, some studies introduce role-playing games to simulate some scenarios, so that the involved agents may discover potential risks and produce training/testing data automatically [188].

### 2.5.3 Adversarial Robustness

In previous sections, we primarily focused on detecting whether models generated unsafe content. Benchmark datasets usually collect potentially harmful prompts through pattern matching or rule-based filtering to measure

the toxicity probability of model outputs. As LLMs advanced, researchers recognized that static prompts were insufficient to comprehensively expose risks. This led to a shift toward adversarial testing or red-teaming methods, where humans or automated algorithms iteratively refine prompts to bypass safeguards. Before benchmarks, we first briefly introduce several typical attack methods as one of the basic evaluation components. There are two groups of methods: white-box and black-box attacks [189]. For white-box attack methods, Gradient-Based Red Teaming (GBRT) [190] uses model gradients to optimize prompts that trigger policy violations. In contrast, many black-box strategies treat the model as an API and use search or another LLM to craft exploits. Recent methods include reinforcement learning to generate realistic but harmful queries [191] and persona-driven attacks, such as SoP [192], which creates multi-character role-play scenarios to exploit a model’s social

3. <https://github.com/openai/evals>

compliance. To highlight the importance of robust input processing, fuzzing techniques have been proposed to capture subtle prompt variations. GPTFuzz [193] mutates seed prompts and reveal sensitivity to slight input perturbations, while WordGame [194] conceals harmful requests behind scrambled text to bypass content filters. These diverse red-teaming approaches exploit different model vulnerabilities (from over-confidence and “distractibility” to context manipulation and timing), greatly expanding the adversarial toolkit.

Alongside attack methodologies, researchers have built evaluation datasets to benchmark LLM robustness under attack. We can roughly classify them into three groups. The first type adopts single-turn attack, similar with those introduced above yet with intentional design to induce the model into ignoring its safety guardrails (e.g., a universal adversarial suffix). Example datasets include AdvBench [195], ForbiddenQuestions [196], AART [197], AdvPromptSet [198], AttaQ [199], CPAD [200], and ALERT [201] that introduces a fine-grained risk taxonomy consisting of 6 macro and 32 micro categories. Second, datasets like AnthropicRedTeam [202], AnthropicHarmlessBase [184], and Bot-Adversarial Dialogue (BAD) dataset [203] leverage human or agent to curate adversarial dialogues with the goal of exposing model failures in multi-turn interactions. The third group of datasets aim to proactively spotting vulnerabilities of models. Recent literature deploys evolutionary red teaming processes to optimize attacks. A representative work of this kind is AutoDan [204] which employs an hierarchical genetic algorithm to evolve prompts. Unlike previous attacks [205] that require gradient-based optimization, AutoDan efficiently operates mutations and crossovers of attack prompts as paraphrasing and linguistic exchange of paragraph content, easily strengthening any manually designed attacks without losing their semantic meaningfulness. The more recent follow-up AutoDan-Turbo [206] further extends such a genetic process to evolve high-level attacking strategies, leading to a life-long learnable red teaming system that can be generally applicable to discover unforeseen threats to forthcoming LLMs.

### 2.5.4 *Agentic Safety*

The newest frontier in LLM safety evaluation is agentic safety, which assesses LLMs that act as autonomous agents, operating tools or navigating environments on behalf of users. These agents must not only avoid producing harmful content, but also avoid harmful actions. This introduces new safety challenges rooted from both users and environments, which involve handling multifaceted threats associated with user authorities, system mechanisms and runtime user-system interaction sessions [207]. In terms of the environment, many studies focus on web agents like Mind2Web-SC [208], AdvWeb [209], EIA [210]. To explore more domains, EICU-AC [211] targets the medical domain to evaluate access control of LLM agents based on user authorization when processing electronic health records. SafeOS [207] evaluates the robustness of OS agents, meanwhile, investigates a broad range of attacks including prompt injection, system sabotage attacks, and environment attacks. Agent-SafetyBench [212] encompasses 349 interactive environments (simulated scenarios) with 2,000 total test cases,

covering 8 categories of safety risks and 10 common failure modes for agent behavior. ASB [213] includes 10 scenarios and benchmark various attack tools, e.g., prompt injection, memory poisoning, and backdoor. Instead of building costly environments, R-judge [214] consists of 569 logs of multi-turn agent interactions (drawn from various simulated applications) with annotated risk events covering 27 scenario types and 10 distinct risk categories. The task is for an LLM to read the log and correctly flag any unsafe decisions or outcomes.

Clearly, agentic safety evaluation is inherently more challenging than static LLM evaluation, because it requires simulating an interactive environment. Besides, the benchmark has to define the risk taxonomy, evaluator for open-ended responses or actions, attack tools for robustness assessment, etc. Therefore, in the future more environments are expected to cover various domains. In these simulations, as agents are intended to handle long-horizon tasks, another evaluation focus is long-term robustness under distribution shift. An agent might start aligned, but after many steps or after successively encountering adversarial inputs, it could deviate from policy. Finally, a critical aspect of agentic safety is balancing utility with safety. If an agent is overly constrained, it may refuse to use its tools at all or become useless.

## 2.6 Integrated Capabilities: General-purpose Evaluation

Early benchmarks for LLMs often targeted isolated capabilities, e.g., logical reasoning or instruction following in separate tests. However, real-world tasks rarely exercise these skills in isolation. Recent evaluation efforts therefore emphasize integrated assessment from GLUE to MMLU, and to Big-Bench, measuring how well an LLM can combine knowledge, reasoning, instruction following, multi-modal understanding, and safety together. The goal is to move beyond concrete ability tests toward holistic evaluation for general-purpose performance or artificial general intelligence, mirroring the integrated demands of real-world use. After the discussion on each ability, we summarize their taxonomy in Figure 6. Now, we will discuss the overlaps of integrated or comprehensive benchmarks.

### 2.6.1 *Interplay Among Evaluation Capabilities*

In practice, while we have detailed the evaluation datasets and potential challenges for each individual capability in preceding sections, these capabilities are inherently intertwined. In this section, rather than trivially enumerating every possible combination, we will instead examine selected examples of capability interplay and then discuss generalized and holistic evaluation.

**Knowledge & reasoning.** Knowledge and reasoning are inherently intertwined. Effective reasoning fundamentally depends on a model’s underlying knowledge base. As illustrated in earlier sections on mathematical reasoning, it assumes mastery of basic mathematical theorems; coding reasoning requires knowledge of programming languages; commonsense reasoning often evaluates familiarity with commonsense knowledge; planning necessitates procedural knowledge; even purely logical reasoning rarely operates

in complete isolation from knowledge — even using arbitrary symbols or “neutral” facts to eliminate the influence of memorized world knowledge, it may risk oversimplifying language understanding and introduce artificial patterns that models exploit without generalizing to real-world reasoning [61]. Conversely, knowledge-intensive tasks frequently demand reasoning capabilities. For example, open-domain QA datasets like HotpotQA [215] and StrategyQA [216] usually require models to retrieve and interconnect multiple facts before reaching conclusions, i.e., multi-hop reasoning.

**Instruction following & Knowledge & Reasoning.** Broadly speaking, instructions, as user inputs to models, can encompass any task description. This implies that knowledge and reasoning capabilities can also be considered sub-skills of instruction following. For example, in practice, users often ask LLMs to perform multi-step tasks via natural language instructions, “Explain how to solve this math problem step by step.” or “Analyze the argument in the following paragraph.” Benchmarks like MT-Bench [130] combine instruction following with knowledge, requiring LLMs to generate responses across diverse topics while adhering to user directives. Similarly, VicunaEval [132] integrates inherent knowledge with precise instruction execution. Reasoning tasks can also be framed as instructions. They usually define a set of rules, where LLMs must follow logical steps to derive final states from initial conditions [117]. Relevant benchmarks also include CodeXGlue [45] and BigCodeBench [39]. These require LLMs to follow detailed programming instructions for tasks like code completion, unit testing, and documentation generation.

**Interaction with multi-modal understanding.** Multi-modal understanding is inherently orthogonal to other capabilities. All previously discussed evaluation benchmarks can be extended to additional modalities. A clear example is VQA benchmarks, where models are given an image and a related question. To succeed, the model must interpret visual content (detect objects, scenes, text in the image) and often use world knowledge or reasoning to answer the question. Benchmarks like OK-VQA [217] specifically target this intersection, requiring models to integrate external knowledge with visual comprehension. For multi-modal reasoning, specialized benchmarks emerge. Math-Vista [29] and MathVision [218] targets the evaluation of multi-modal math reasoning by combining diagrammatic representations with textual problem statements. MMMU [158] and MMMU-Pro [146] presents college-level questions that interleave text with heterogeneous visual inputs, demanding both domain-specific knowledge and advanced reasoning skills. This dataset pushes models to draw upon a broad base of subject knowledge while performing deliberate, expert-level reasoning across multiple disciplines. OlympiadBench [33] further pushes difficulty to Olympiad-level in math and physics context. Similarly, visual instruction tuning [219, 220] bridges visual understanding with instruction following, while Huang et al.[221] provide a comprehensive survey. In summary, multi-modal evaluation expands the scope of integrated assessment. It ensures LLMs’ general capabilities extends beyond text to interpret and reason about visual (or auditory, etc.) worlds in conjunction with language.

**Interaction with safety.** The safety capabilities of LLMs are also orthogonal to other competencies yet critically important for real-world deployment. Increasingly, benchmarks incorporate safety evaluations alongside knowledge and reasoning tasks. For instance, TruthfulQA [222] systematically tests models with 818 challenging questions spanning 38 domains (e.g., health, law, finance) to distinguish between truthful responses and fluent but factually incorrect answers that mimic human plausibility. This paradigm evaluates not just factual knowledge and linguistic proficiency, but crucially measures truthfulness alignment, prioritizing correct, honest responses over eloquently stated misconceptions and thereby integrating factual reasoning with safety metrics. For intersection with instruction-following capabilities, models that unconditionally obey user requests risk generating harmful outputs. Effective safety alignment necessitates the ability to override instructions when appropriate. Contemporary evaluations address this by incorporating refusal-worthy prompts (Section 2.5.1), where properly aligned models must demonstrate refusal competence or safe response redirection. Notably, SafeBench [223] provides a systematic framework for evaluating multi-modal LLM safety, extending these principles to complex, real-world deployment scenarios. This comprehensive approach ensures that safety mechanisms remain robust when models operate at the intersection of knowledge retrieval, reasoning, and instruction execution, a critical requirement for trustworthy AI systems.

## 2.6.2 Comprehensive Evaluation

Based on the above analysis, the field is progressively integrating knowledge, reasoning, instruction following, multi-modal understanding, and safety into integrated benchmark suites, moving beyond isolated skill testing, for a comprehensive measure of a model’s general-purpose capabilities. This is not only the abilities are inherently intertwined, but also real-world deployment requires the simultaneous application of these capabilities. Early initiatives like GLUE [224] and SuperGLUE [225] pioneered this approach by aggregating several common NLP tasks, enabling multifaceted evaluation of pre-trained language models like BERT during fine-tuning. Subsequent comprehensive benchmarks expanded the scope. MMLU [226] includes 57 subjects, including elementary mathematics, US history, computer science, and law. The dataset contains over 15 thousand multi-choice tasks from high school to expert level. MMLU-Pro [227] updates the MMLU framework with more challenging reasoning tasks, enhanced robustness, and reduced dataset noise. As the transition from task-oriented to capability-centric evaluation occurs, BIG-bench [62] curates over 200 diverse tasks covering mathematics, linguistics, commonsense reasoning, and social bias analysis among others. To address computational constraints, BIG-bench Lite provides a distilled 24-task subset for efficient performance measurement. Building on that idea, recent benchmark collections like HELM [228] and VHELM [229] explicitly report a profile of each model across many aspects, from accuracy on academic questions to robustness under input perturbations and fairness in responses. The aim is to identify not just “which model is best” but in what ways

a model is strong or weak, and how well it balances the competing demands of capability and alignment.

Another holistic evaluation frameworks put LLMs into agent roles, asking them to operate in interactive environments or multi-step decision problems, which tests a convergence of capabilities in scenarios closer to real-world deployment. Except for prior agent planning (Section 2.2.6) and agentic safety (Section 2.5.4), there are also some general evaluation involving one or multiple LLM agents to collaborate or compete. LLMs-as-an-Examiner [230] simulates peer review assessments. Each LLM operates like an examiner to generate queries and also judge the responses from other LLMs. By collaboratively determining the evaluation results, this process reduces biases and enhances fairness in evaluations. Auto-Arena [231] introduces discussion among LLM agents to automate this process. Inspired by educational assessment processes, AutoDetect [232] employs three LLM-powered agents — Examiner, Questioner, and Assessor — that collaborate to generate test scenarios and analyze model responses. AgentCQ [233] leverages LLMs to automate the creation and assessment of clarifying questions in conversational search systems, while LEGALAGENT [13] pushes the evaluation further to agent-based legal reasoning. There are also visual agent evaluation framework to evaluate the ability of MLLMs to perform complex real-world tasks like UI operation on mobile devices [234], robotic control in household tasks [235], card-based games [236], and navigation [116, 237, 238]. Other similar frameworks include IQA-Eval [239], ALI-Agent [240], ChatEval [241], MATEval [242], and AgentSims [243]. Unlike static datasets, this type of interactive benchmarks also test adaptability and decision-making. An agent can observe new information and must decide its next action. More practically, a model might initially answer a question incorrectly, but in an interactive setting it could be given feedback or detect the error and correct itself. Metrics for such evaluations can include success rate, efficiency (steps taken), and qualitative ratings of the agent’s behavior.

### 3 AUTO-EVALUATION

In Section 2, we introduced commonly used datasets categorized by five core capabilities and discussed their interplay. However, these static datasets lead to delayed updates of test sets, hindering their alignment with model progress. Additionally, they are still susceptible to performance overestimation due to data contamination. In this section, we first introduce several dynamic benchmarks and live leaderboards, followed by methods for automated dataset curation and evaluation. The Github page we will maintain and welcome any collaborators is [https://github.com/ALEX-nlp/Chapter3\\_Awesome\\_Paper\\_List](https://github.com/ALEX-nlp/Chapter3_Awesome_Paper_List).

#### 3.1 Dynamic benchmarks

Dynamic benchmarks aim at continuously updating the testing data to offer a fairer assessment. There are mainly two types of advantages. First, it considers the dynamic nature of world knowledge, thus preventing false negatives caused by outdated information and accurately assessing

model latest advancements. Early works highlight the timeliness of knowledge by introducing timestamp, where a piece of knowledge holds true only within its own timestamps. To obtain accurate timestamp, common sources for dataset curation include WIKIDATA [16], news articles [18], or existing datasets annotated by crowd-sourcing workers [17]. The target of such evaluation is similar with temporal commonsense reasoning in Section 2.2.4. Building on top of them, recent benchmarks target real-time evaluations. REALTIMEQA [19] evaluates models weekly on approximately 30 multiple-choice questions derived from recent news events. This benchmark highlights the importance of continual learning and real-time knowledge integration for accurate and timely responses. KOLA [20] and KNOT [21] take a step further by not only evaluating the coverage of the rapidly changing world knowledge but also the ability of models to integrate the new knowledge with the existing knowledge.

The second advantage of dynamic benchmarks is that consistently updating data can mitigate the data contamination issue. There are two type of methods. The first group still leverages the timeliness of knowledge. Since as long as the evaluation data pertains to the latest knowledge, the model would not have been exposed to it, thereby no testing data shall be seen during training. However, other researchers also point out that the risk lies in the difficulty of completely distinguishing new from old knowledge based on a specific cutoff date (such as the model’s release date). For instance, even if a movie is released after this date and the model should not have seen it, necessary information might have already been exposed to the model through early promotions and related activities [23]. Representative works include EvoWiki [23] and AntiLeak-Bench [24]. For example, EvoWiki is an evolving dataset that categorizes information into stable, evolved, and uncharted states. By comparing the information before and after the LLM’s cutoff date, stable data do not change and evolved data indicate an update. While, uncharted data only involve those events happened thereafter.

Instead of the timeliness of knowledge, another line of research highlights the dynamics of data — as long as the evaluation data keeps changing, they are hardly exposed to the model during evaluation. Dynabench [244] is an open-source platform that incorporates human-and-model-in-the-loop dataset creation. Unlike traditional static benchmarks, Dynabench enables annotators to craft examples that challenge current models, revealing their weaknesses and promoting the development of more robust systems. This dynamic approach directly integrates human feedback into the evaluation process. Livebench [22] aims at releasing new questions monthly, sourced from recent information such as math competitions, arXiv papers, and news articles, thereby ensuring that models are assessed on fresh, unseen data. It encompasses six categories: math, coding, reasoning, language comprehension, instruction following, and data analysis, each with tasks that have verifiable ground-truth answers. Focusing on specific coding task, Livecodebench [47] instead continuously collects new problems from coding competitions on platforms like LeetCode, AtCoder, and CodeForces. Beyond code generation, LiveCodeBench also assesses capabilities in self-repair, code execution, and test



output prediction, providing a holistic view of an LLM’s coding proficiency.

*Live Leaderboard* To facilitate a convenient evaluation, some researchers develop and maintain real-time evaluation platforms that are updated either manually or automatically. Chatbot Arena [245] is an open evaluation platform based on human preferences. Users can engage in side-by-side conversations with anonymous AI models and vote for their preferred responses, facilitating direct comparisons of AI capabilities in real-world scenarios. The platform employs the Elo rating system to rank models based on user votes. Since its launch in May 2023, Chatbot Arena has attracted millions of participants and collected over 800,000 votes, becoming a critical resource for live, community-driven LLM evaluation.

EvalPlus Leaderboard [246] is a platform for code generation. It utilizes enhanced benchmarks, such as HumanEval+ and MBPP+, which offer significantly more test cases than their original versions, to assess models’ code correctness and efficiency. By ranking models based on metrics like pass@1 using greedy decoding, the leaderboard provides insights into each model’s coding proficiency and robustness.

Open LLM Leaderboard [247] tracks, ranks, and evaluates open-source LLMs and chatbots. It provides a centralized resource for comparing the performance of various models across multiple benchmarks, facilitating informed decisions for researchers and developers in the AI community. Users can submit their models for automated evaluation on Hugging Face’s GPU cluster, ensuring standardized assessments. The leaderboard is continuously updated, reflecting the latest advancements.

C-Eval Leaderboard [248] is a comprehensive Chinese evaluation suite, consisting of 13,948 multiple-choice questions spanning 52 diverse disciplines, including humanities, science, and engineering, and is structured across four difficulty levels: middle school, high school, college, and professional. Notably, C-Eval includes a challenging subset known as C-Eval Hard, which focuses on subjects requiring advanced reasoning skills, such as advanced mathematics and college physics.

Clearly, to ensure the high quality of dynamic benchmarks, the key is how to automate dataset curation and evaluation, which will be detailed in the next section.

## 3.2 Automated Dataset Curation

Qualified human-annotated data requires substantial budgets and time cost, thus being particularly vulnerable to the rapid outdatedness and potential information leakage. Accordingly, more and more evaluation datasets are constructed in auto-synthesized manners. In this section, we summarize the common auto-synthesis strategies into three main branches: compilation, derivation and generation. (1) *Compilation* involves combining or selecting existing annotations to align with the intended use of the dataset. (2) *Derivation* utilizes existing datasets but modifies annotations or adds new components to serve specific purposes. (3) *Generation* involves partially or completely constructing datasets by automatically generating new contexts or annotations, often with the assistance of LLMs. We detail these three

strategies as below. Note that the strategies summarized here are not exclusive: the construction of one evaluated dataset can leverage multiple strategies.

### 3.2.1 Compilation

Compilation is the most simple and widely-used approach for building a new evaluation dataset. It can be further divided into Combination and Selection.

**Combination** integrates existing annotations into a new, single benchmark. These benchmarks are designed to evaluate the general capabilities of LLMs [22, 62, 145, 224, 225], or assess specific abilities in a more comprehensive and robust manner like math/STEM [31, 33, 249, 250], coding [38, 47], long-context understanding [90, 91, 251], information retrieval [252–255], etc. The biggest challenge for combination is how to construct the benchmark a hierarchical and reasonable taxonomy. The taxonomy in existing benchmarks are usually designed in the following dimensions:

- **Target abilities.** For example, MathVista [29] summarizes seven mathematical reasoning capabilities. MMBench [149] designs a three-level, twenty-subclass taxonomy tree to evaluate the perception and reasoning abilities of LLMs. Recently, MEGA-Bench [256] includes over 500 real-world tasks within the hierarchical taxonomy.
- **Discipline and/or difficulty.** Most benchmarks sourced from examinations or exercises [158, 226, 248, 249, 257] are usually categorized by disciplines and/or difficulties. For instance, MMLU [226] and MMMU [158] incorporate questions from 57 and 30 subjects, respectively. MATH [31] and M3Exam [257] divide their questions into 5 and 3 difficulty levels.

Given the critical importance and relatively limited scale of the taxonomy, it still heavily relies on human design. Traditionally, the taxonomy is entirely designed by the main contributors of the benchmarks. However, as the taxonomy scale expands, the responsibility for its expansion is distributed. For example, BIG-Bench [62] encourages the entire community to submit pull requests for new tasks. MEGA-Bench [256] initially provides a draft two-level taxonomy and invites all project members to contribute to its growth. When the taxonomy is completed, they either collect previous datasets (such as GLUE [224], SuperGLUE [225], BIG-bench [62], and MathVista [29]) or gather annotations from multiple sources like websites, textbooks, or real-world data (such as MMLU [226], MATH [31], OlympiadBench [33], C-Eval [248], M3Exam [257] and MMMU [158]).

**Selection** involves filtering annotations to create new benchmarks. Such process usually serves for three purposes:

- **Scale control.** Some benchmarks, especially those constructed in combination way, sample annotations randomly from previous datasets to control the scale of the curated datasets [91, 258].
- **Preliminary filtering.** The raw data crawled from real-world are sometimes noisy. It requires a preliminary selection to improve the recall of qualified data with the minimum time and budget cost. Simple but effective heuristic rules are usually adopted. For example, ELI5 [259] and BRIGHT [260] select

high-quality posts and/or answers (measured by views, votes, URL numbers) in the Reddit or Stack-Exchange. Benchmarks for tabular task like HybridQA [261] and FinQA [262] retain moderate-size tables (measured by row and column numbers) to balance the information amount and task difficulty. SciToolBench [263] picks out tools (Python functions) passing the unit tests to ensure the correctness of these tools.

- **Post-refinement.** When the annotations have been made from scratch or collected from previous datasets, an additional selection (after the main construction process) further benefit the dataset from various aspects like quality, diversity, difficulty, etc. Compared to pre-filtering, it requires more customized assessment and often involves LLMs/LVLMs as judges. For instance, MMLongBench-Doc [264] evaluates the document understanding abilities of LVLMs (instead of their intrinsic knowledge) and thus employs GPT-4o to remove sample candidates which can be directly answered without the access to documents. VisRAG-Bench [265] introduces Llama-3 to filter out context-dependent queries which are not appropriate for open-domain retrieval task. MMStar [148] evaluates the visual mathematical reasoning abilities and thus employs GPT-4 to remove sample candidates which can be answered by text-only information. HaluEval [15] and MMHal-Bench [266] introduce ChatGPT and LLaVA to rank and select high-quality hallucinated answers (*i.e.*, more plausible and close to the correct answers) from previous generations. For a simple approach to achieve strong reasoning during test-time inference, S1 [267] has conducted post-refinements regarding quality, difficulty, and diversity. For quality, they filter out low-quality examples by checking if they contain any string patterns with formatting issues, such as ASCII art diagrams, non-existent image references, or inconsistent question numbering. For difficulty, they first use Claude3.5 to select the correct responses, and then measure the problem difficulty through the token length of each response including the reasoning process. For diversity, they introduce the Mathematics Subject Classification (MSC) system (*e.g.*, geometry, dynamic systems, real analysis, etc.) and classify each question into these specific domains using Claude 3.5 Sonnet, keeping balanced distribution across different domains.

### 3.2.2 Derivation

Derivation is an automatic construction strategy somewhat between compilation and generation. It still heavily relies on existing annotations, but introduces significant modifications. We further categorize derivation into two subtypes, Transfer and Supplementary, as detailed below.

**Transfer** usually occurs when we evaluate the identical or highly-similar tasks/abilities under different settings. In such cases, creating new benchmarks from scratch is neither necessary nor economical. The common choice is to make new benchmarks by transferring from an existing, well-developed ones. For similar tasks, BEIR [252]

and ViDoRE [254] are information retrieval benchmarks collected from multiple QA datasets by simple conversion: (i) from question to query. (ii) merge passages as retrieval corpus. Being an open-domain QA dataset, OTT-QA [268] rewrites the queries in HybridQA [261] for decontextualization. To explore the long-context/multi-page settings, MP-DocVQA [269] and LongBench [91] increase the document lengths in previous datasets like DocVQA [151] by incorporating additional context pages or similar paragraphs. For the same tasks at different modalities, with the development of LVLMs, benchmarks for many critical capabilities and practical task in text domains are converted to visual domains and used to evaluate LVLMs. For example, Multimodal-Mind2Web [270] and SWE-bench Multimodal [271] trace back the webpage screenshots which are used as textual format in Mind2Web [272] and SWE-bench [38] to evaluate the agent and coding abilities of LVLMs. Similarly, Wiki-VISA [273] and M3DocVQA [274] render the Wikipedia URLs in NQ [5] and MultimodalQA [275] datasets towards the evaluation of visualized document understanding and grounding.

**Supplementary** usually occurs when we have the benchmark about some tasks and aim to evaluate their further/successive tasks. In such cases, it is a natural choice to build corresponding benchmarks in supplementary approach, *i.e.*, adding new annotations based on the original ones. To explore whether retrieval benefits code generation models, CodeRAG-Bench [276] is derived from coding benchmarks [42, 47] by augmenting collected documents as retrieval corpus. HellaSwag [74] is synthesized from ActivityNet [277] as a QA dataset in which the negative choices are adversarially synthesized. RuleBench [278] induces rules from multiple logic-related datasets and add these rules to form a new benchmark for inferential rule-following evaluation. MMLongBench-Doc [264] and MuirBench [279] add unanswerable ones by replacing keywords in original questions, thereby detecting potential hallucinations. Upon annotated events, MAVEN-FACT [280] automatically generate their factualities for Event Factuality Detection (EFD) task.

### 3.2.3 Generation

When there are no corresponding qualified annotations for reuse and/or edition, generation becomes the indispensable choice for automatic data construction.

**Rule-based Generation** is still widely-used in the era of LLMs due to its efficiency and deterministic, especially under the following scenarios. (1) *Mine real-world data.* The data for some tasks like math, coding and knowledge base are entailed in related informative platforms and knowledge-rich sources. It is natural to design automatic pipelines and extract these real-world high-quality data as evaluation benchmark. For example, LeanDojo [281] and SWE-Bench [38] extract proofs from Lean and pull requests from GitHub repositories, respectively. RealTimeQA [282] extracts questions from news websites which requires latest knowledge. AntiLeak-Bench [24] leverages the knowledge among entities from Wikidata and synthesizes QA pairs for contamination-free evaluation. Recently, CODEELO [283] extracts updated coding problems from CodeForces. (2) *Synthesize for certain capability evaluation.* To evaluate some

specialized capabilities of LLMs/LVLMs, it is also beneficial to create somewhat artificial but targeted datasets. To assess the comprehensive reasoning abilities across different modalities, MultimodalQA [275] creates cross-modal questions from single-modal questions by pre-defined, compositional templates. Similarly, RuleTaker [284] generates facts and rules in logic, performs forward inference to derive all its implications, and obtains questions expressed in (synthetic) English using simple natural language templates. MM-NIAH [285] concatenates interleaved image-text sequences from the OBELICS [286] dataset to create long-context documents, referred to as multimodal haystacks. POPE [287] employs templates which convert image instances with object detection annotations to QA pairs for object hallucination evaluation.

**LLM-based Generation** has been an important approach and research topic for automated dataset construction. The motivation of LLM-based generation can be categorized into four aspects. (1) Label generation. Here the raw corpus already exists. An LLM replaces human annotators by producing labels, rationales, or exemplar responses. Example methods cover multiple scenarios like role-playing [288], multi-agent communication [289], multi-turn interaction [258, 290], code generation [39, 291], tool-use [292], *etc.* (2) Context generation. This aims to complete the missing parts of existing benchmarks, e.g., generating responses or options in multi-choice questions. In such cases, the generations are exactly the evaluation targets of the benchmarks. For example, HaluEval [15] and FavaBench [293] focus on hallucination evaluation and thus leverages ChatGPT to generate contexts with potential errors. A series of benchmarks [294–297] focus on reward model evaluation and therefore generate a pair/group of LLM-generated responses which are fed to reward model. (3) Reference-based revision. This type of methods aim at generating new data based on provided reference. Representative works include WizardLM [298] and follow-up works. They treat an existing dataset as scaffolding and issue editing instructions: add constraints, deepen reasoning, inject noise, or rephrase, to an LLM, which rewrites each item into harder variants. Ying *et al.* [299] proposed two types of methods, mimicking and extending, to systematically update test sets to mitigate possible data contamination issue. This approach preserves topical relevance yet upgrades difficulty and coverage, functioning as controllable “data augmentation for evaluation”. (4) From-scratch generation. When no suitable seed corpus exists, carefully designed prompts elicit an LLM to invent both tasks and solutions. Self-Instruct [300] pioneers this method by supplying a handful of seed exemplars and letting the model extrapolate thousands of similar instruction–response pairs. LLM-as-Examiner [301] considers both evaluation breadth and depth to generate diverse evaluation data. For breadth of knowledge, they collect thousands of domain descriptions as instructions. For depth, they prompt LLMs to generate follow-up questions as well as the responses. Note that the first three generation approaches mentioned above can also be viewed as data synthesis methods — they reprocess existing data to maintain quality while ensuring flexibility.

### 3.3 Pipeline of Automated Dataset Curation

To conclude, towards a qualified benchmark, current works carefully consider and design the following steps: (1) Well-defined taxonomy. Under each topic and/or task type, personalized generation strategies or instructions are adopted by LLMs and significantly improves the coverage and quality of LLM generations (See more details about the taxonomy construction in previous discussion). For example, [145, 147, 302, 303] categorize multiple tasks and feed task-specific instructions to LLMs for generate more high-quality QA pairs. (2) Step decomposition. The auto-annotation of benchmark is usually decomposed into several subsequent steps. For instance, many QA-formatted benchmarks [289, 292, 296, 304] separately generate the questions/instructions, answers/responses. Regarding more complicated task, [305] synthesizes instances for *Summary of Haystack* task by four sub-steps: insight generation, document generation, query generation and summary generation. (3) Prompt strategy. Most earlier benchmarks [263, 292] draw inspirations from In-context Learning (ICL; [306]) and provide seed examples in prompts. These examples explicitly instruct LLMs to generate annotations with desired contents and formats. Moreover, more detailed prompts, higher quality the generation content, making prompts written in detailed instructions become more and more popular [273, 303]. (4) Verification. The preliminary generation from LLMs shall undergo verification procedure before use. For math and coding tasks, the easiest verification occurs when ground-truth answers are known and the generation result are deterministic or executable. In such cases, rule-based parsers or programming executor [263, 307, 308] are adopted for verification. Also, LLM-as-a-judge evaluators can efficiently assess the generation quality (More details in the coming section).

### 3.4 Evaluator

As LLMs have unified a variety of natural language processing tasks through natural language generation, a significant shift has occurred in how open-ended responses are evaluated. Traditionally, evaluation relied on task-specific metrics, calculated by directly comparing model outputs to reference texts. For example, in classification tasks [309], metrics such as accuracy, precision, recall, and F1 score are commonly employed; in ranking tasks [310], metrics like NDCG are typically used. Similarly, a similar approach was adopted for natural language generation tasks. BLEU [311], for instance, is an automatic metric that measures the quality of machine-generated translations by calculating the overlap of n-grams between the output and reference texts. A higher overlap indicates better alignment with the reference, suggesting higher translation quality. Likewise, ROUGE [312] evaluates the quality of summaries in the same way by calculating word overlap between the evaluated summary and human-generated ideal summaries. However, BLEU and ROUGE primarily rely on lexical matching, often overlooking lexical order and meaning. METEOR [313] improves upon BLEU by not only considering unigram overlap but also incorporating lexical stem and semantic matching, which better captures linguistic diversity in translations. Additionally, METEOR accounts for recall and the ordering of lexical

matches, allowing for a more accurate assessment of machine translation quality. Despite these improvements, such metrics still heavily rely on surface-level lexical overlap, which often fails to capture deeper semantic nuances, coherence, or logical consistency.

In response, many embedding-based evaluation methods have emerged to assess model-generated responses at the semantic level. BERTScore [314], for example, evaluates the quality of machine-generated text by comparing the semantic similarity between the generated text and reference using contextual embeddings from BERT. Unlike traditional metrics that focus solely on lexical overlap, BERTScore captures deeper meaning by comparing the cosine similarity of token embeddings, which makes it more effective in handling synonyms, paraphrasing, and variations in sentence structure. However, BERTScore still heavily relies on the availability of reference answers. In evaluation tasks where references are scarce or difficult to obtain, this dependence significantly limits its applicability. Furthermore, while BERTScore shifts from word-level to semantic-level evaluation, it struggles to capture aspects beyond semantics such as helpfulness and harmlessness. These limitations have highlighted the need for the development of reference-free and multi-aspect evaluation methods.

As the performance of LLM has progressively advanced, GPT and other models are introduced to replace BERT as evaluators with or without reference, a.k.a., the proposal of the LLM-as-a-judge concept [130]. Moreover, thanks to the strong knowledge memorization and instruction following capabilities, LLM-as-a-judge can even evaluate the responses from multiple aspects, like informative, engaging, etc. A typical example is GPTScore [315], which utilizes LLMs (including GPT-3 [316], OPT [317], and FLAN [318]) to evaluate text quality across multiple aspects without relying on reference responses.

Currently, the development of LLM-as-a-judge is continually progressing, and the definition of LLM-as-a-Judge has gradually taken on a clear and formal expression. A formal definition of LLM-as-a-Judge is as follows:

$$P_{\theta}(X^n, C) \rightarrow R$$

- $P_{\theta}$ : The LLM-as-a-Judge fulfilled by any LLM, which can be either foundation LLMs or fine-tuned version. The generation process is the auto-regressive process.
- $X^n$ : The samples to be evaluated. They can be of any available type, such as text, images, or videos. Here  $n$  represents the number of samples to be evaluated: When  $n = 1$ , it becomes a point-wise judgment, where the evaluation result is a score. When  $n = 2$ , it becomes a pair-wise judgment, where the evaluation result is a comparison. When  $n > 2$ , it becomes a list-wise judgment, where the evaluation result is presented as a ranking.
- $C$ : The context of the input  $x$ , which includes relevant evaluation examples, historical information in the dialogue, or the definition of evaluation criteria.
- $R$ : The final evaluation result obtained from the LLM-as-a-Judge can be a relative or absolute score with rationale or not.

Based on the above definition, in this section, we introduce four strategies for making LLM-based evaluators more effective and robust: suitable prompt context, multi-evaluator collaboration, human-LLM collaboration, and better base LLMs.

### 3.4.1 Suitable prompt context

Careful prompt design is crucial for guiding an LLM judge to produce accurate and consistent evaluations. By tailoring the prompt with context, examples, or structured reasoning steps, researchers aim to align the model’s judgments with human criteria. Below we review current methods that optimize the evaluation prompt from the perspectives of in-context samples, reasoning instruction, fine-grained criteria, and role-play augmentation.

**In-context Samples.** Demonstration plays an important role in the in-context learning. Many works [306, 319, 320] have discussed the effectiveness regarding the sample selection, order, etc. Focusing on evaluation, a few high-quality demonstrations can calibrate the LLMs’ expectations, guiding them in understanding and applying assessment standards. Methods like GPTScore [315] provide example answers with known quality to help the model learn how to assess text quality on the fly. Kotonya et al. [321] shows the effectiveness of combining multiple prompt design methods with zero-shot and one-shot in-context samples, and the CoT prompt-based method shows considerable potential for assessing the quality of generated summaries. Few-shot prompts make evaluation training-free and adaptable, but they can also introduce bias if the selected examples are unrepresentative. To mitigate such bias, ALLURE [322] iteratively refines the in-context examples by identifying erroneous evaluation outcomes, correcting them, and incorporating the revised results as updated examples. Alternatively, Song et al. [323] introduce two types of many-shot in-context learning prompts, Many-Shot with Reference (MSwR) and Many-Shot without Reference (MSoR), to combat position or symbol biases.

**Reasoning instruction.** Evaluation also requires the reasoning ability to verify or infer the relationship between the response and the question. We roughly classify existing methods into two groups: CoT and planning instruction. A representative work in the first group is G-Eval [324], which designs an auto-CoT framework that instructs LLMs to automatically generate evaluation steps given criteria before scoring. This framework with GPT-4 as the backbone model significantly improves the assessment of text summarization and dialogue generation tasks, achieving a high correlation with human evaluations. To further evaluate the effectiveness of G-Eval, Chiang et al. [325] examine how specific details in G-Eval’s evaluation process influence the correlation between ratings provided by LLMs and those given by humans. Their findings indicate that the auto CoT used in G-Eval does not always enhance alignment with human ratings. However, they also find that prompting the LLM to explain its own ratings consistently improves the correlation between ChatGPT’s evaluations and human judgments. Domain-specific evaluators like ICE-Score [326] for code generation go further. The prompt includes detailed evaluation steps, criteria, and task definitions, leading the LLM through a checklist (e.g. correctness, efficiency)

when scoring code. The second group of methods target the planning ability — one can improve the evaluation performance by reducing the evaluation difficulty through task decomposition. [327] proposes the Branch-Solve-Merge (BSM) framework to evaluate responses by dividing tasks into parallel sub-tasks, then solving each sub-task separately, and finally merging the results into an overall assessment. While, SocREval [328] introduces the Socratic method to leverage a sequence of probing questions to refine the reasoning instruction.

**Fine-grained criteria.** Another prompt strategy is to explicitly embed evaluation criteria or rubrics into the prompt, so the LLM judge assesses each aspect independently. This criteria decomposition makes the evaluation more transparent and objective. For example, some users prefer to informative responses while others may expect concise answers. Some researchers have explored fine-grained evaluations by indicating specific aspects (e.g., Fluency, Coherence, etc.) [301, 329, 330] and detailed rubrics [331] [332] via in-context learning. For instance, Jain *et al.* [333] investigate the efficacy of LLMs as multi-dimensional evaluators: coherence, relevance, consistency, and fluency, each with two example scores. Their findings indicate that the prompt design strategies perform on par with traditional evaluation frameworks in text summarization tasks. Similarly, FineSurE [334] exemplifies this by breaking summarization quality into dimensions like faithfulness, completeness, and conciseness; the LLM performs fact-checking and key fact alignment for each before outputting an overall judgment. Furthermore, HD-EVAL [335] enhances principle-driven prompting with hierarchical criteria. The authors first decompose the evaluation aspects using an LLM and assign scores to each sub-metric. Then, an aggregator combines these sub-metric scores into a total score, with human-labeled results used to train the aggregator. To investigate whether the evaluator can recognize and differentiate between various evaluation criteria, Hu *et al.* [336] summarize and define an explicit hierarchical classification system consisting of 11 criteria. Using these criteria to test the evaluation capabilities of models, they identify that LLMs often confuse different criteria. To address this issue, they train the evaluator using clearly defined criteria to mitigate the potential confusion of different evaluation standards by LLMs. More studies on tuning evaluators will be introduced later.

**Multi-turn & role-play augmentation.** To better align with human judgment, recent methods have introduced multi-turn or role-play instructions. AutoCalibrate [337] leverages a multi-stage prompt refinement process: the LLM is first prompted to draft initial evaluation criteria for a task, then revise them, and finally apply them. Another approach is to give the LLM a specific role or persona [338] like “You are a strict grammar teacher” or “You are a helpful peer reviewer”. This can inject diverse evaluative perspectives and make the LLM more adaptable to different contexts. However, there are also some concerns: overly narrow roles or poorly chosen criteria can bias the evaluation. The goal of all these prompt-based techniques is to supply *just enough contextual guidance* so that the LLM’s inherent knowledge is steered toward accurate judging, minimizing randomness or bias in its responses.

### 3.4.2 Multi-Evaluator Collaboration

Relying on the results from a single LLM judge may not be reliable due to the various biases inherent in LLMs. Typical biases include:

- **Position Bias.** Position bias refers to the tendency of LLMs to favor answers based on their position in the response. This bias is common in various natural language processing tasks [301, 339] as well as in human decision-making processes [340]. Even advanced LLMs like ChatGPT and GPT-4 encounter this issue when acting as evaluators [130, 341].
- **Knowledge Bias.** Knowledge bias occurs when the pre-trained data fails to include certain essential tasks or introduces potentially harmful knowledge, which can undermine the generative performance of LLMs. In the evaluation scenario, this bias occurs when the knowledge required for evaluation tasks exceeds the scope of the LLM judge’s training.
- **Style Bias.** Style bias in LLMs refers to the tendency to favor certain writing styles or tones due to the patterns in the pre-trained data. This bias can affect the LLM’s judgment, leading to assign higher scores to outputs that align with its preferred style, regardless of content quality.
- **Format Bias.** Format bias refers to the situation where a judge is fine-tuned without a reference but validated with a reference, or vice versa, resulting in a mismatched format. LLM judges perform poorly in these mismatched formats.

To overcome this limitation, several architectures and techniques use multiple LLMs [242] (or multiple instances of an LLM [330, 342]) that either cooperate or compete, and then combines their outputs. The intuition is that aggregating multiple perspectives can cancel out individual errors or biases and lead to more reliable outcomes. There are two main types of methods: cooperative approaches and aggregation approaches.

The first group is **aggregated multi-agent evaluation**, where models judge independently and their results are fused later. Representative works include Language-Model-as-an-Examiner [301], which let a panel of LLMs generate probing questions about a candidate answer and then independently evaluate the answer, aggregating their scores. This peer-questioning plus voting mimics how a committee of examiners might each test a student with different questions, leading to a well-rounded evaluation. The benefit of voting ensembles is their simplicity and parallelizability. However, if all models share a blind spot, the ensemble won’t fix it. Also, how to ensemble their results is critical in the final evaluation quality. 1) Beyond simple voting, more sophisticated aggregation methods assign different weights or roles to each evaluator. One idea is to weight judges by their past agreement with humans. PRE [343] conducts a “qualification exam” to select LLMs as reviewers, and then weights their ratings based on how well each aligns with human judgments. Differently, PiCO [344] treats the evaluation problem as a constrained optimization. Multiple models answer questions and evaluate each other’s answers, and an algorithm finds weights for each model’s opinion to maximize overall consistency within the group. 2) Apart from



weighting by quality, we can also assign different evaluation criteria to different models. AIME [345] gives each of several LLM judges a specific aspect to score (e.g. one model focuses only on factual accuracy, another only on fluency), and then concatenating or fusing these aspect-specific scores into an overall evaluation. Similarly, HD-Eval [335] uses a panel of evaluators where each handles a hierarchically decomposed subset of criteria. 3) Advanced aggregation schemes also borrow from consensus algorithms. For example, Gao *et al.* [346] applied Bayesian models to calibrate win rates when many LLM evaluators are voting, correcting biases in pairwise preference aggregation. Others construct a preference graph from multiple weak judges and then use graph algorithms to derive a final ranking that is more transitive-consistent [347].

Another group of method is **cooperative multi-agent evaluation**, where multiple LLMs interact, sharing information or engaging in debate, to reach a consensus. In these setups, each model might handle a different sub-task or provide feedback on others. For example, WideDeep [348] uses an architecture that lets models share information at a “neuro-level”, effectively merging their intermediate representations to improve joint decision-making. Other work borrows from human workflows. Xu *et al.* [349] simulate an academic review process — each agent drafts a solution, then reviews others’ work and revises its own answer based on received critiques. Similarly, ABSEval [350] assigns four distinct agent roles (answer synthesis, critique, execution, commonsense reasoning) that sequentially interact to evaluate an answer. By role assignment in a collaborative workflow, the evaluators complement each other’s strengths (one agent might catch logical errors, another factual errors, etc.). However, a notable risk is “groupthink”. If the models have similar biases or training backgrounds, their agreement may simply reinforce a shared bias rather than provide truly independent perspectives. Designing agent diversity (e.g. using different model architectures or prompt viewpoints) may be a potential solution. Therefore, another line of works is to have LLMs debate or challenge each other’s answers in a competitive fashion. In such frameworks, LLMs take on roles of debaters and a separate judge (which could itself be an LLM or an ensemble of LLMs) decides the winner of the debate. An example work is Auto-Arena [342], where candidate models engage in multi-round debates over a question, pointing out flaws in each other’s responses. Extensions of this idea, like the MORE and SAMRE architectures [351], involve multiple advocate agents and iterative rebuttal rounds, resembling a courtroom with opposing counsel and a verdict delivered after several back-and-forths. By contrast, decentralized debate structures let all models freely converse without a single controller. ChatEval [352] assigns diverse roles to multiple LLMs (e.g. one may emphasize precision, another creativity) and lets them discuss an open-ended question collectively. Similarly, PRD [353] has models not only rank each other’s answers but also discuss them, which helped reduce biases like self-enhancement (where a model favors responses similar to itself) and positional bias (favoring the first presented answer). Competitive debates tend to reveal flaws through contradiction and defense, leading to a more nuanced judgment. The challenge, however, is complexity

and cost: multi-round debates consume more computation, and if not carefully orchestrated, the interactions could go in circles or become incoherent. Nonetheless, adversarial multi-LLM evaluation is a promising way to stress-test answers and achieve consensus closer to human critical analysis.

For efficiency, a variant of multi-evaluator systems is the cascade approach, where judges are arranged in tiers of increasing strength or cost. The idea is to use cheaper (or less powerful) models to handle easy evaluations and reserve expensive state-of-the-art models for the tricky cases, thereby optimizing resource use while maintaining accuracy. Jung *et al.* [354] propose Cascaded Selective Evaluation, where a small judge model first evaluates; only if its confidence is low or a decision boundary is ambiguous, a larger model (like GPT-4) is called in. Similarly, CascadedEval [355] combines open-source fine-tuned judges with proprietary models in a pipeline, leveraging the strengths of each. The fine-tuned judge handles routine cases and the proprietary model corrects its failures. Such cascades illustrate a pragmatic collaboration between models of different caliber. The main challenge is designing a reliable gating mechanism to decide when to escalate to the next tier. If tuned well, cascaded systems can be both efficient and robust, effectively forming a safety net where the final tier (strongest model or even a human) only handles the most uncertain evaluations.

### 3.4.3 Human-LLM Collaboration

Despite advances in automated evaluation, human insight remains essential, especially for open-ended tasks where nuanced understanding or ethical considerations are critical. Human-LLM collaboration frameworks aim at combining both merits: the efficiency of LLM judgments and the reliability of human oversight. A straightforward solution is **humans as verifier**. The LLM judge operates almost autonomously, but a human performs a final check or adjustment on its outputs [325, 356].

Another line of work is **humans as assistant**. In this setup, LLMs generate initial judgments, which humans then refine before a final judgment is made. For example, HMCEval [357] proposed a human-machine collaborative framework for dialogue evaluation. It optimizes evaluation reliability while minimizing human effort. Through a sample assignment approach, it reduces human involvement by half while maintaining 99% evaluation accuracy, demonstrating a highly efficient solution for reliable dialogue evaluation. While, CoEval [358] lets LLMs first generate task-specific evaluation metrics, which are then judged by humans for their usefulness. Afterward, carefully selected metrics are input to the evaluator to obtain evaluation results, which are further refined by humans. EvalGen [359] tackles the problem of “criteria drift”, where an LLM’s tendency to unintentionally change its evaluation standards over many responses. In EvalGen, humans periodically provide feedback on the LLM’s judging criteria, keeping them aligned over time. This iterative refinement of the evaluation rubric, driven by human judgment, was found to improve the consistency and fairness of long-running automated evaluations. In addition to assisting in the generation of evaluation metrics, evaluators can also be employed

TABLE 2  
Comparisons of example evaluators in general or specific domains.

Method	Format	Critiques	Multi-rubrics	Data	Tuning	Example rubrics
Shepherd	pointwise	yes	Overall	Human	SFT	Error Analysis
Themis	pointwise	yes	Multiple	GPT-4	SFT	Cohesiveness; Likability; Clarity; Length; Engagement; etc.
PandaLM	pairwise	yes	Multiple w/o ratings	GPT-3.5	SFT	Relative Conciseness; Clarity; Comprehensiveness; Formality; Adherence to Instructions; etc.
JudgeLM	pointwise & pairwise	yes	Multiple w/o ratings	GPT-4	SFT	Helpfulness; Relevance; Accuracy; Level of Details of Responses
AUTO-J	pointwise & pairwise	yes	Multiple w/o ratings	GPT-4	SFT	Core Idea Capturing; Concise; Coverage; Harmlessness; Creativity; Engagement; Information Richness; etc.
Prometheus	pointwise & pairwise	yes	Multiple	GPT-4	SFT	Each sample is assigned a specific evaluation measure
TIGERScore	pointwise	yes	Multiple w/o ratings	GPT-4	SFT	Comprehension; Accuracy; Informativeness; Coherence; Fact Consistency; Fluency; Accuracy; etc.
CritiqueLLM	pointwise & pairwise	yes	Multiple	GPT-4	SFT	Accuracy; User Satisfaction; Logical Coherence; Creativity; Richness; Overall Score
HALU-J	pointwise	yes	Overall	GPT-4o	DPO	Hallucination
Safety-J	pointwise	yes	Overall	Human & GPT-4	SFT	Safety

to help generate test samples. Ribeiro and Lundberg [360] introduce AdaTest, a human-LLM collaborative approach for automatically generating unit tests to identify and fix bugs in NLP models. AdaTest significantly improves bug detection efficiency, making users 5-10x more effective than traditional methods. Additionally, Rastogi et al. [361] enhance the AdaTest auditing tool with human-AI collaboration, creating AdaTest++ to rigorously evaluate commercial language models like GPT-3 and Azure’s sentiment analysis. Their tool leverages human strengths in sensemaking and hypothesis testing, effectively identifying a wide range of failure modes. Another work by Wang et al. [362] introduced a calibration framework to correct known biases of LLM judges via human guidance.

The benefit of human-in-the-loop methods is a high assurance of quality: before any score is finalized, a person has vetted the process. Besides, Human-LLM collaboration can also be used to continuously improve the evaluator over time. By analyzing where the LLM’s judgments disagree with humans, developers can refine prompts or fine-tune the model. While, the downside is scalability — it requires human labor for each evaluation or each batch of evaluations, so it may not be as fast or cheap as fully automated methods. Thus, these approaches are often more useful for high-stakes settings (e.g. medical evaluation) where accuracy outweighs speed. Or, in open-ended tasks where the “ground truth” is subjective and context-dependent, human guidance helps keep the automated judge aligned with social values and the specific goals of the evaluation.

#### 3.4.4 Better base LLMs

All of the aforementioned methods enhance the LLM’s evaluation capability without modifying the LLM parameters themselves. In this section, we focus on training LLMs specially for evaluation usage. There are general-purpose evaluators as well as domain-specific evaluators focusing on particular issues (e.g. safety compliance or factual accuracy). Below, we survey these advances from the perspectives of evaluation data curation and tuning techniques, followed by

highlighting some representative systems. Finally, we outline key trends and trade-offs between fine-tuned evaluators and prompting-based evaluation.

**Evaluation data construction.** In Section 3.2, we have detailed automated dataset curation methods including compilation, derivation, and generation. Therefore, we will not repeat those methods and only focus on the construction of evaluation data. The difference is that the training data for evaluators need to include LLMs’ responses, which we classified into “context generation” in Section 3.2.3. To annotate these responses, there are mainly two types of approaches: manually-labeled and auto-synthetic.

To obtain manual labels, a straightforward solution to hire experts for annotation [363]. This usually results in high-quality, nuanced feedback, but it is costly and slow to scale. To lower the cost, some works leverage existing resources like online community feedback or crowdsourced annotations. Shepherd [363] collects user feedback from two well-known communities: Stack Exchange and the Pushshift Reddit. They treated the title and subtitle of a post as a question, the top-level comments as answers, and the replies to these comments as critiques. The quality of these critiques can be evaluated based on the net upvotes and downvotes. Similarly, Vu et al. [364] curated a large and diverse set of over 100 quality assessment tasks, encompassing more than 5 million human evaluations from publicly released human feedback. Such community-sourced critiques provide diverse, real-world error examples. However, these relies written by human and the standards for their feedback may differ from benchmarking’s needs.

For auto-synthetic data, a common strategy is to have a powerful model directly generate assessment scores given the LLM’s response. For example, Auto-J [329] implement a “divide-and-conquer” approach, where GPT-4 generates two critiques for each response, which are then merged into a more thorough critique before providing a final rating. Except for pointwise ratings, many works favor pairwise comparisons. The above Auto-J [329] integrates both. For pairwise data, they provide two responses to the evaluator

and ask them to identify the criteria where the evaluations differ between the two. PandaLM [341] re-formulates and completes the samples from Alpaca 52K as tuple (instruction, input, response1, response2), and ask GPT to generate output tuples (evaluation result, evaluation reason, reference response). JudgeLM [365] adopts a similar way but leverages GPT-4 acted as a “teacher” judge. Except for the data format (i.e., comparison or ratings), scoring rubrics are also critical. Prometheus [366] uses GPT-4 to enhance a set seed of manual rubrics and generate more. These newly generated rubrics not only provide clearer standards, but also are more favorable for GPT judge, thereby high-quality evaluation data. According to various rubrics, the corresponding error analysis or critiques provide additional interpretability. TIGERScore [367] builds a dataset called MetricInstruct with instruction prompts that ask for error analysis, where each entry includes a model output and a list of errors (with types and severity) as the label.

The above two methods each has its own merits. Often the best results come from combining human expertise with LLM generation. For example, InstructScore [368] and TIGERScore [367] uses explicit human instruction (defining what to evaluate) together with GPT-4’s implicit knowledge to label data. While, Safety-J [369] leverages human refine or review initial safety critiques.

**Tuning techniques.** Given an evaluation dataset, the next step is to train the LLM to produce desired judgments or critiques via standard post-training techniques, e.g., supervised finetuning (SFT) [363, 365, 366, 370] or direct preference optimization (DPO) [371–373]. By contrast, traditional RL is less commonly used for training evaluators, since obtaining a numeric reward for a correct evaluation is not trivial — one would need a “meta-evaluator”. However, RLHF (human feedback) is introduced for improvements. For example, Safety-J [369] employs an iterative preference learning loop, which uses its own critiques to perform meta-evaluation and then prefers revisions that improve its performance. Over iterations, this is akin to the model reinforcing behaviors that lead to more accurate safety judgments. Of course, when letting the evaluator evaluate its own outputs and improve, the strategy is related to self-RL (e.g. “Self-Refine” and “Self-Reward” methods).

Complementary to the above tuning techniques, some tricks were introduced to make an evaluator be stable and unbiased in its judgments — some irrelevant factors like the order in which answers are presented may be captured during tuning. To mitigate positional bias, JudgeLM [365] conducted swap-augmentation (shuffling answer order), and used reference support/drop techniques to teach the judge to rely on content rather than position or formatting. For robustness, GPT-4-based evaluators have been shown to exhibit variability if prompts are paraphrased<sup>4</sup>. To counter this, recent research [372] generated many paraphrased instructions and fine-tuned models to give consistent preferences.

Each of the above models has a unique emphasis. Table 2 lists some representative evaluators including general and some specialized evaluators, e.g., Safety-J [369] for safety judgement or Halu-J [371] for assessing hallucinations,

which prompt GPT-4 to generate multiple pieces of evidence for each instance as well as the final critique based on evidence.

## 4 OPEN CHALLENGES AND FUTURE DIRECTIONS

In this section, we discuss three core challenges that characterize the path towards generalizable evaluation in the era of LLMs.

### 4.1 Challenges in Capability-Based Evaluation

As LLMs unify various tasks and show human-like abilities, we conclude the transition from task-centric to capability-based evaluation. Section 2 provides a comprehensive survey of capability-based benchmarks including isolated and integrated evaluation. Based on them, we observe two core challenges.

**On one hand, how to achieve optimal balance between the efficiency and generalization of evaluation?**

We can see that those comprehensive benchmarks face inherent scalability challenges. Unlike training data which benefits from scaling laws, benchmark expansion cannot indefinitely cover all desired competencies. Even for the agent-based evaluation, they also face their own efficiency-generalization trade-offs due to dependency on environment design. The core challenge lies in selecting optimal task combinations that maximize evaluation efficiency while enabling reliable prediction of model full capability spectrum with limited test data. Preliminary solutions incorporating interpretability techniques like MUI [2] have been proposed, but these represent only initial steps.

**On the other hand, should the evaluation focus on fine-grained capabilities or comprehensive integration?**

Integrated datasets enable multi-dimensional analysis of model capabilities to identify strengths and weaknesses as guidance for training, yet they often overlook the tightly coupled nature of these competencies. While, agent-based evaluation naturally integrates multiple capabilities’ testing via some environments. The high competency threshold for meaningful participation often excludes smaller models. Meanwhile, it typically lacks granular interpretability, thus offering little guidance for model optimization.

### 4.2 Challenges in Automated Evaluation

As LLM capabilities expand, creating suitable evaluation data by hand and judging the model’s responses become a bottleneck. Automation promises to keep pace with rapid model progress and reduce our reliance on expensive human labeling. For the transition from manual to automated evaluation, Section 3 provides a comprehensive survey of dataset curation and evaluators. Now, we discuss the core challenges from the two aspects.

Recent progress in automated benchmarking shows a striking dependence on LLMs for data curation, yet **generating harder, more diverse, and genuinely high-quality test data remains challenging**. First, the quality ceiling of synthetic data is bounded by the current capability of the generator model. As one tries to raise difficulty — longer contexts, more intricate reasoning, multimodal grounding

4. <https://eugeneyan.com/writing/llm-evaluators/>

— the fidelity of LLM-generated content drops sharply. Although LLMs will keep improving, evaluation difficulty will rise in tandem, preserving the difficulty–quality trade-off. In practice, many researchers now resort to teacher–student distillation: crafting challenging prompts with a stronger “teacher” model and using them to train or benchmark smaller systems. Ultimately, however, a true closed loop of continuous model improvement demands generation techniques that surpass current capability ceiling, not merely mirror it. Second, verifying the quality of synthetic data is itself non-trivial. If one needs an even stronger “super-LLM” to vet examples, the workflow becomes circular: how does one generate or validate data for that super-LLM? When existing human-curated corpora are exhausted, the field risks a “chicken-and-egg” bottleneck in which no component can improve without better data from the other. Third, although “diversity” is widely acknowledged as crucial, there is no unified measurement or definition. Recent work has proposed counting domains, capability categories, or difficulty levels [267], and Shypula *et al.* [374] introduce a new metric for data diversity and quality. Yet we still lack a fine-grained formalism that links specific diversity dimensions to learning efficiency. An even deeper question is whether optimal diversity should be model-specific? The data needed to expose weaknesses in a retrieval-augmented LLM may differ from that required for a domain-specific LLM. Addressing these gaps will be pivotal for next-generation evaluation pipelines, which we will further discuss it later.

For evaluators, a straightforward question is **prompt-based or tuned evaluators**. Compared with prompt-based evaluation, tuned evaluators offer substantial advantages in cost and throughput, yet they cannot fully replace a strong foundation model when factual knowledge or sophisticated reasoning is required. When fine-tuning data are sparse or poorly curated, a tuned judge is prone to new biases, and its generalization seldom matches that of the underlying base model. Even so, tuned evaluators remain highly promising.

In specific, the current trend of evaluation data curation favors large-scale synthetic labeling using LLMs, sometimes combined with human annotations. This looks like knowledge distillation, tuned evaluators ultimately will be capped by the teacher model. But, through some well-designed tricks and human involvements, tuned evaluators can be more robust. Furthermore, some studies let an evaluator critique its own judgments and retrain on those critiques, forming a self-refinement loop that can yield ever-improving judges. Thus, we vision a rise of specialized “judge” models that focus on particular concerns, e.g., safety, bias, factuality, reasoning [375], or on domains such as math [376]. These niche evaluators incorporate domain knowledge (retrieval for factuality, step-by-step solution checking for math) that a general judge might not possess. The trend suggests an ensemble of evaluators, each an expert in checking a certain aspect, could be used together for thoroughly evaluation. Finally, **fine-grained, explainable judgments also attract increasing research attention**. This not only builds user trust but also transforms evaluation into a form of error analysis. It enables using the judgments to directly improve the generative model, thus closing the loop between evaluation and revision.

### 4.3 Challenges in Generalizable Evaluation

The core challenge in the era of LLMs is ensuring that our evaluation method keep up with the essentially unbounded capabilities of future LLMs. Traditional evaluation is bounded in the sense that it uses a fixed set of test examples and metrics, often reflecting the existing capabilities of LLMs. But actually LLMs are moving targets — their abilities grow with scale and training, while evaluation can not be expanded infinitely considering the efficiency. We are increasingly observing that an evaluation which a new model excels at might no longer be discriminative (the model “outgrew” the test), or conversely, a model might possess latent capabilities that the evaluation fails to reveal. This mis-match between what models can do and what we measure them on is widening. Thus, a core future direction is designing generalizable evaluations that anticipate and extrapolate to new model behaviors, rather than being one-off, static tests.

One aspect of this is **forecasting model capabilities** from the perspective of evaluation method. If we had reliable ways to predict how a model will perform on a broad range of tasks before actually testing it (or before the model even exists), we could design better benchmarks and safety checks proactively. Scaling laws can be regarded as a typical early work. They provide empirical relationships between model size, training compute, or data and performance, so that we may know the LLMs’ future performance even during the early training stage. The BIG-Bench [1] was also motivated by extrapolating performance, tasks in BIG-Bench were chosen to be beyond the reach of smaller models, with the expectation that progress would be measurable as models scale up. A recent work proposed to reduce the number of tasks by training a generic assessor for predictive performance [377]. Indeed, its results indicated that some tasks improve smoothly with model size while others show discontinuous leaps at certain scales. Understanding these patterns (why some abilities suddenly “activate” at a threshold) is crucial for forecasting. If we can identify predictors in smaller models or early training phases that correlate with later emergent capabilities, we could flag potential breakthroughs in advance. These predictions, while not perfect, help benchmark designers include tasks that will remain challenging at the next generation of models, thereby “future-proofing” evaluations to some extent. That is, generalizable evaluation cares about “How predictable are LLM capabilities?” We may use integrated ability and cross-scenario data to forecast performance of yet-unseen models.

Another direction of generalizable evaluation is dealing with the **inherent coverage problem** from the perspective of datasets. An LLM’s possible behaviors are virtually infinite (suppose our ultimate target is AGI or ASI), but any test set is finite. How can we ensure that a finite evaluation set meaningfully probes the vast space of model competence? One idea is to focus on maximizing diversity and coverage with minimal data (as discussed above). To do so, instead of one-test-set-for-all, future evaluations might be adaptive or model-specific. For example, an evaluator could iteratively find areas where the model’s performance is problematic and add more tests there, until performance stabilizes. This

resembles adaptive testing in education, where questions are chosen based on a student’s previous answers to pinpoint their proficiency. In the LLM context, we may need to keep generating follow-up questions with suitable difficulty to map out the boundaries of its capabilities. If the model easily handles all math questions but struggles with certain logic puzzles, the system would concentrate evaluation on the latter to fully characterize the weakness. Some preliminary work in the safety domain in this direction includes adversarial testing and red teaming methods, where an auxiliary model or algorithm tries to find inputs that make the model fail. Going forward, model-specific diversity could become standard [2]. Each new model might be evaluated with a tailored set of stress tests chosen to cover its potential blind spots (as identified by prior models or preliminary runs). The goal is to achieve broad coverage (knowledge, reasoning, multi-modal, instruction following, safety, etc.) with as few test items as possible by targeting representative challenges rather than exhaustively enumerating trivial cases. This not only makes evaluation more efficient but also more generalizable: a well-chosen small test suite could predict performance on a much larger distribution of tasks because it captures the essential difficulties.

The third direction of generalizable evaluation seeks to **predict as-yet-uncovered abilities given limited testing sets** from the perspective of metric. An preliminary attempt is the Model Utilization Index (MUI) [2]. MUI augments traditional, outcome-oriented scores by incorporating mechanism interpretability techniques, whereas classical metrics concern what result the model produced on a fixed test set, MUI additionally measures how much internal effort the model expended to obtain that result. Extensive experiments reveal an intuitive law: performance score is inversely correlated with MUI. The basic idea is when judging a human’s overall proficiency we weigh both outcome and effort, where equal performance achieved with less effort (lower MUI) signals greater competence. Nevertheless, this line of work remains constrained by the present limits of interpretability research. Neuron-localization methods, for example, have been criticized for imperfectly disentangling functional sub-skills, potentially undermining MUI’s precision. Sparse-Autoencoder (SAE) approaches, while more expressive, currently lack off-the-shelf generalizability; training a SAE for every new foundation model is prohibitively expensive. Moreover, both families of the above techniques require white-box access and are inapplicable to closed-source LLMs. Despite these hurdles, the marriage of interpretability and evaluation presents a promising path forward. By looking inside the model we may transcend the intrinsic ceiling of finite test sets, inferring latent strengths or weaknesses that static outcome metrics miss. In short, explainable-aware metrics such as MUI demonstrate how one can “see the whole from a part”, uncovering a model’s true potential with limited external data.

The last intriguing direction of generalizable evaluation is **using a model’s minor signals or reasoning traces to discover hidden capabilities or weaknesses**, probably from the perspective of evaluators. As LLMs increasingly can show their work (through CoT prompting, rationale outputs, or just the open-ended response itself), we have new data to judge what the model “knows” or where it

falls short. Some recent works focus on evaluating reasoning traces [378]. Anthropic’s study [379] shows that CoT explanations are not always faithful. Although a reasoning model occasionally discloses which prompts or intermediate deductions it used, in most cases the verbalized CoT only partially reflects the model’s actual computation. Even so, CoT monitoring remains valuable. Because unexpected behaviors, especially ones that unfold over several steps, often leave detectable artefacts in the trace, a fine-grained analysis can still surface hidden patterns. In other words, by inspecting the style, structure, or subtle irregularities in a model’s explanation, evaluators can uncover clues about latent strengths or systematic flaws that would be invisible in a simple right/wrong score. Such process-oriented evaluation does more than mark an answer incorrect; it reveals why it is wrong, and that diagnostic insight generalizes to many other inputs, not just the specific question posed. By treating the model’s own explanations as data to be checked for factual alignment and logical validity, the boundary between outcome evaluation and process evaluation begins to blur. If the underlying reasoning process is demonstrably sound — even on problems we did not explicitly test — we gain confidence in the model’s broader reliability.

## 5 CONCLUSION

LLMs are improving at a pace that outstrips conventional evaluation pipelines. In this survey, we mapped that tension onto two transitions and highlight the core limitation. 1) From tasks to capabilities, we re-organize benchmarks around five core abilities, knowledge, reasoning, instruction following, multi-modality and safety. This yet raises two open questions: Efficiency vs. generality and Granularity vs. integration. 2) From human-curated to LLM-automated evaluation. Automation is essential for keeping pace, but it introduces its own difficulties like generating harder, more diverse, high-quality data and tuning explainable, fine-grained LLM judges. 3) Toward generalizable evaluation, the core obstacle is a coverage gap: finite test sets cannot scale with unbounded model abilities. We thus discuss the potential directions including predictive evaluation, adaptive datasets, generalizable metrics, and see-the-whole-from-a-part evaluator. Addressing these challenges demands a hybrid toolbox. Only by scaling our evaluations as aggressively as we scale our models can we ensure that performance claims remain meaningful, reliable and fair. In the future, because the field evolves month-by-month, we will maintain a living repository<sup>5</sup>, and warmly invite contributions that refine, correct or extend the taxonomy presented here.

## REFERENCES

- [1] B. bench authors, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *Transactions on Machine*

<sup>5</sup>. Benchmarks for core capabilities are at <https://github.com/ALEX-nlp/Benchmark-of-core-capabilities/tree/main>, and auto-evaluation methods are at [https://github.com/ALEX-nlp/Chapter3\\_Awesome\\_Paper\\_List](https://github.com/ALEX-nlp/Chapter3_Awesome_Paper_List)



- Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=uyTL5Bvosj>
- [2] Y. Cao, J. Ying, Y. Wang, X. Qiu, X. Huang, and Y. Jiang, "Revisiting llm evaluation through mechanism interpretability: a new metric and model utility law," *arXiv preprint arXiv:2504.07440*, 2025.
  - [3] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2463–2473.
  - [4] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611.
  - [5] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019.
  - [6] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1533–1544.
  - [7] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4149–4158. [Online]. Available: <https://aclanthology.org/N19-1421/>
  - [8] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "Financebench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.
  - [9] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, Y. Xu, H. Kang, Z. Kuang, C. Yuan, K. Yang, Z. Luo, T. Zhang, Z. Liu, G. XIONG, Z. Deng, Y. Jiang, Z. Yao, H. Li, Y. Yu, G. Hu, H. Jiajia, X.-Y. Liu, A. Lopez-Lira, B. Wang, Y. Lai, H. Wang, M. Peng, S. Ananiadou, and J. Huang, "Finben: An holistic financial benchmark for large language models," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: <https://openreview.net/forum?id=loDHzstVP6>
  - [10] Y. Nie, B. Yan, T. Guo, H. Liu, H. Wang, W. He, B. Zheng, W. Wang, Q. Li, W. Sun, Y. Wang, and D. Tao, "Cfinbench: A comprehensive chinese financial benchmark for large language models," *arXiv preprint arXiv:2407.02301*, 2024, available at: <https://doi.org/10.48550/arXiv.2407.02301>.
  - [11] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, A. Huang, S. Zhang, K. Chen, Z. Yin, Z. Shen, J. Ge, and V. Ng, "LawBench: Benchmarking legal knowledge of large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7933–7962. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.452/>
  - [12] H. Li, Y. Chen, Q. Ai, Y. WU, R. Zhang, and Y. LIU, "Lexeval: A comprehensive chinese legal benchmark for evaluating large language models," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: <https://openreview.net/forum?id=8RaxRs5VDf>
  - [13] H. Li, J. Chen, J. Yang, Q. Ai, W. Jia, Y. Liu, K. Lin, Y. Wu, G. Yuan, Y. Hu *et al.*, "Legalagentbench: Evaluating llm agents in legal domain," *arXiv preprint arXiv:2412.17259*, 2024.
  - [14] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. [Online]. Available: <https://aclanthology.org/2022.acl-long.229/>
  - [15] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.397/>
  - [16] W. Chen, X. Wang, and W. Y. Wang, "TIMEQA: A benchmark for temporal question answering," in *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://arxiv.org/abs/2108.06314>
  - [17] M. Zhang and E. Choi, "SituatingQA: Incorporating extra-linguistic contexts into QA," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7371–7387. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.586/>
  - [18] A. Liska, D. Kumar, A. Singhal *et al.*, "STREAMINGQA: A benchmark for adaptation to new knowledge over time in question answering models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: <https://arxiv.org/abs/2205.11388>
  - [19] J. Kasai, K. Choi *et al.*, "REALTIMEQA: A dynamic benchmark for real-time question answering," in *Proceedings of the 2022 Annual Meeting of the Association for*

- Computational Linguistics (ACL)*, 2022, pp. 2034–2045.
- [20] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li, C. Li, Z. Zhang, Y. Bai, Y. Liu, A. Xin, K. Yun, L. GONG, N. Lin, J. Chen, Z. Wu, Y. Qi, W. Li, Y. Guan, K. Zeng, J. Qi, H. Jin, J. Liu, Y. Gu, Y. Yao, N. Ding, L. Hou, Z. Liu, X. Bin, J. Tang, and J. Li, “KoLA: Carefully benchmarking world knowledge of large language models,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=AqN23oqraW>
- [21] Y. Liu, Z. Yao, X. Lv, Y. Fan, S. Cao, J. Yu, L. Hou, and J. Li, “Untangle the KNOT: Interweaving conflicting knowledge and reasoning skills in large language models,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 17186–17204. [Online]. Available: <https://aclanthology.org/2024.lrec-main.1493/>
- [22] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum, “Livebench: A challenging, contamination-free LLM benchmark,” *CoRR*, vol. abs/2406.19314, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.19314>
- [23] W. Tang, Y. Cao, Y. Deng, J. Ying, B. Wang, Y. Yang, Y. Zhao, Q. Zhang, X. Huang, Y. Jiang *et al.*, “Evowiki: Evaluating llms on evolving knowledge,” *arXiv preprint arXiv:2412.13582*, 2024.
- [24] X. Wu, L. Pan, Y. Xie, R. Zhou, S. Zhao, Y. Ma, M. Du, R. Mao, A. T. Luu, and W. Y. Wang, “Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge,” *arXiv preprint arXiv:2412.13670*, 2024.
- [25] Y. Wang, X. Liu, and S. Shi, “Deep neural solver for math word problems,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 845–854. [Online]. Available: <https://aclanthology.org/D17-1088>
- [26] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, “MathQA: Towards interpretable math word problem solving with operation-based formalisms,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2357–2367. [Online]. Available: <https://aclanthology.org/N19-1245>
- [27] S.-y. Miao, C.-C. Liang, and K.-Y. Su, “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*, D. Jurafsky, J. Chai, N. Schuster, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 975–984. [Online]. Available: <https://aclanthology.org/2020.acl-main.92>
- [28] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, “Training verifiers to solve math word problems,” *arXiv preprint arXiv:2110.14168*, 2021.
- [29] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K. Chang, M. Galley, and J. Gao, “Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=KUNzEQMWU7>
- [30] T. Sawada, D. Paleka, A. Havrilla, P. Tadepalli, P. Vidas, A. Kranias, J. J. Nay, K. Gupta, and A. Komatsuzaki, “ARB: advanced reasoning benchmark for large language models,” *CoRR*, vol. abs/2307.13692, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2307.13692>
- [31] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, “Measuring mathematical problem solving with the math dataset,” *NeurIPS*, 2021.
- [32] B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu, and B. Chang, “Omni-math: A universal olympiad level mathematic benchmark for large language models,” *CoRR*, vol. abs/2410.07985, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2410.07985>
- [33] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun, “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)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 3828–3850. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.211>
- [34] E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. de Oliveira Santos, O. Järvinen, M. Barnett, R. Sandler, M. Vrzala, J. Sevilla, Q. Ren, E. Pratt, L. Levine, G. Barkley, N. Stewart, B. Grechuk, T. Grechuk, S. V. Enugandla, and M. Wildon, “Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai,” *CoRR*, vol. abs/2411.04872, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2411.04872>
- [35] D. Hendrycks, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, and J. Steinhardt, “Measuring coding challenge competence with apps,” *NeurIPS*, 2021.
- [36] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry,

- Q. Le *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.
- [37] Y. Lai, C. Li, Y. Wang, T. Zhang, R. Zhong, L. Zettlemoyer, S. W. tau Yih, D. Fried, S. Wang, and T. Yu, "Ds-1000: A natural and reliable benchmark for data science code generation," *ArXiv*, vol. abs/2211.11501, 2022.
- [38] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan, "SWE-bench: Can language models resolve real-world github issues?" in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=VTF8yNQm66>
- [39] T. Y. Zhuo, M. C. Vu, J. Chim, H. Hu, W. Yu, R. Widyasari, I. N. B. Yusuf, H. Zhan, J. He, I. Paul *et al.*, "Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions," *arXiv preprint arXiv:2406.15877*, 2024.
- [40] D. Huang, J. M. Zhang, Y. Qing, and H. Cui, "Effibench: Benchmarking the efficiency of automatically generated code," *CoRR*, vol. abs/2402.02037, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2402.02037>
- [41] F. Cassano, J. Gouwar, D. Nguyen, S. Nguyen, L. Phipps-Costin, D. Pinckney, M. Yee, Y. Zi, C. J. Anderson, M. Q. Feldman, A. Guha, M. Greenberg, and A. Jangda, "Multipl-e: A scalable and polyglot approach to benchmarking neural code generation," *IEEE Trans. Software Eng.*, vol. 49, no. 7, pp. 3675–3691, 2023. [Online]. Available: <https://doi.org/10.1109/TSE.2023.3267446>
- [42] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, "Evaluating large language models trained on code," *CoRR*, 2021.
- [43] W. Yan, H. Liu, Y. Wang, Y. Li, Q. Chen, W. Wang, T. Lin, W. Zhao, L. Zhu, H. Sundaram, and S. Deng, "Codescope: An execution-based multilingual multitask multidimensional benchmark for evaluating llms on code understanding and generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 5511–5558. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.301>
- [44] L. Chai, S. Liu, J. Yang, Y. Yin, K. Jin, J. Liu, T. Sun, G. Zhang, C. Ren, H. Guo, Z. Wang, B. Wang, X. Wu, B. Wang, T. Li, L. Yang, S. Duan, and Z. Li, "Mceval: Massively multilingual code evaluation," *CoRR*, vol. abs/2406.07436, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.07436>
- [45] S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. B. Clement, D. Drain, D. Jiang, D. Tang, G. Li, L. Zhou, L. Shou, L. Zhou, M. Tufano, M. Gong, M. Zhou, N. Duan, N. Sundaresan, S. K. Deng, S. Fu, and S. Liu, "Codexglue: A machine learning benchmark dataset for code understanding and generation," *CoRR*, vol. abs/2102.04664, 2021.
- [46] J. Guo, Z. Li, X. Liu, K. Ma, T. Zheng, Z. Yu, D. Pan, Y. Li, R. Liu, Y. Wang, S. Guo, X. Qu, X. Yue, G. Zhang, W. Chen, and J. Fu, "Codeeditorbench: Evaluating code editing capability of large language models," *CoRR*, vol. abs/2404.03543, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2404.03543>
- [47] N. Jain, K. Han, A. Gu, W. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, "Livecodebench: Holistic and contamination free evaluation of large language models for code," *CoRR*, vol. abs/2403.07974, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2403.07974>
- [48] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, W. Zhou, J. Coady, D. Peng, Y. Qiao, L. Benson, L. Sun, A. Wardle-Solano, H. Szabó, E. Zubova, M. Burtell, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, A. R. Fabbri, W. Kryscinski, S. Yavuz, Y. Liu, X. V. Lin, S. Joty, Y. Zhou, C. Xiong, R. Ying, A. Cohan, and D. Radev, "FOLIO: natural language reasoning with first-order logic," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 22 017–22 031. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.1229>
- [49] O. Tafjord, B. Dalvi, and P. Clark, "Proofwriter: Generating implications, proofs, and abductive statements over natural language," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 3621–3634. [Online]. Available: <https://doi.org/10.18653/v1/2021.findings-acl.317>
- [50] A. Saparov and H. He, "Language models are greedy reasoners: A systematic formal analysis of chain-of-thought," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: <https://openreview.net/forum?id=qFVBzXxR2V>
- [51] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang, "Logiqa: A challenge dataset for machine reading comprehension with logical reasoning," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 2020, pp. 3622–3628. [Online]. Available: <https://doi.org/10.24963/ijcai.2020/501>

- [52] W. Yu, Z. Jiang, Y. Dong, and J. Feng, "Reclor: A reading comprehension dataset requiring logical reasoning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: <https://openreview.net/forum?id=HJgT4tvB>
- [53] F. Chollet, "On the measure of intelligence," *arXiv preprint arXiv:1911.01547*, 2019.
- [54] Y. Xu, W. Li, P. Vazezipoor, S. Sanner, and E. B. Khalil, "Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations," *Trans. Mach. Learn. Res.*, vol. 2024, 2024. [Online]. Available: <https://openreview.net/forum?id=E8m8oySvPJ>
- [55] R. Alur, R. Bodík, G. Juniwal, M. M. K. Martin, M. Raghothaman, S. A. Seshia, R. Singh, A. Solar-Lezama, E. Torlak, and A. Udupa, "Syntax-guided synthesis," in *Formal Methods in Computer-Aided Design, FMCAD 2013, Portland, OR, USA, October 20-23, 2013*. IEEE, 2013, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/6679385/>
- [56] K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton, "CLUTRR: A diagnostic benchmark for inductive reasoning from text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 4505–4514. [Online]. Available: <https://doi.org/10.18653/v1/D19-1458>
- [57] W. Zhao, J. T. Chiu, C. Cardie, and A. M. Rush, "Abductive commonsense reasoning exploiting mutually exclusive explanations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 14883–14896. [Online]. Available: <https://doi.org/10.18653/v1/2023.acl-long.831>
- [58] M. Del and M. Fishel, "True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4," in *Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics, \*SEM@ACL 2023, Toronto, Canada, July 13-14, 2023*, A. Palmer and J. Camacho-Collados, Eds. Association for Computational Linguistics, 2023, pp. 314–322. [Online]. Available: <https://doi.org/10.18653/v1/2023.starsem-1.28>
- [59] N. Young, Q. Bao, J. Bensemann, and M. Witbrock, "Abductionrules: Training transformers to explain unexpected inputs," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 218–227. [Online]. Available: <https://doi.org/10.18653/v1/2022.findings-acl.19>
- [60] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, C. Shao, Y. Yan, Q. Yang, Y. Song, S. Ren, X. Hu, Y. Li, J. Feng, C. Gao, and Y. Li, "Towards large reasoning models: A survey of reinforced reasoning with large language models," 2025. [Online]. Available: <https://arxiv.org/abs/2501.09686>
- [61] H. Liu, Z. Fu, M. Ding, R. Ning, C. Zhang, X. Liu, and Y. Zhang, "Logical reasoning in large language models: A survey," *arXiv preprint arXiv:2502.09100*, 2025.
- [62] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoen, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, and et al., "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Trans. Mach. Learn. Res.*, vol. 2023, 2023. [Online]. Available: <https://openreview.net/forum?id=uyTL5Bvosj>
- [63] M. Luo, S. Kumbhar, M. Shen, M. Parmar, N. Varshney, P. Banerjee, S. Aditya, and C. Baral, "Towards logigluue: A brief survey and A benchmark for analyzing logical reasoning capabilities of language models," *CoRR*, vol. abs/2310.00836, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2310.00836>
- [64] H. Rashkin, A. Bosselut, M. Sap, K. Knight, and Y. Choi, "Modeling naive psychology of characters in simple commonsense stories," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, I. Gurevych and Y. Miyao, Eds. Association for Computational Linguistics, 2018, pp. 2289–2299. [Online]. Available: <https://aclanthology.org/P18-1213/>
- [65] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. F. Allen, "A corpus and evaluation framework for deeper understanding of commonsense stories," *CoRR*, vol. abs/1604.01696, 2016. [Online]. Available: <http://arxiv.org/abs/1604.01696>
- [66] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," *Communications of the ACM*, vol. 64, no. 9, pp. 99–106, 2021.
- [67] A. S. Gordon, Z. Kozareva, and M. Roemmele, "Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning," in *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, E. Agirre, J. Bos, and M. T. Diab, Eds. The Association for Computer Linguistics, 2012, pp. 394–398. [Online]. Available: <https://aclanthology.org/S12-1052/>
- [68] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 6720–6731. [Online]. Available: [http://openaccess.thecvf.com/content\\_](http://openaccess.thecvf.com/content_)

- CVPR\_2019/html/Zellers\_From\_Recognition\_to\_Cognition\_Visual\_Commonsense\_Reasoning\_CVPR\_2019\_paper.html
- [69] L. Du, X. Ding, K. Xiong, T. Liu, and B. Qin, “e-care: a new dataset for exploring explainable causal reasoning,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 432–446. [Online]. Available: <https://doi.org/10.18653/v1/2022.acl-long.33>
- [70] Q. N. Ben Zhou, Daniel Khoshabi and D. Roth, ““going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding,” in *EMNLP*, 2019.
- [71] M. T. Phu, M. V. Nguyen, and T. H. Nguyen, “Fine-grained temporal relation extraction with ordered-neuron LSTM and graph convolutional networks,” in *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT 2021, Online, November 11, 2021*, W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, Eds. Association for Computational Linguistics, 2021, pp. 35–45. [Online]. Available: <https://doi.org/10.18653/v1/2021.wnut-1.5>
- [72] X. Wang, Y. Chen, N. Ding, H. Peng, Z. Wang, Y. Lin, X. Han, L. Hou, J. Li, Z. Liu, P. Li, and J. Zhou, “MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 926–941. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.60/>
- [73] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, “PIQA: reasoning about physical commonsense in natural language,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 7432–7439. [Online]. Available: <https://doi.org/10.1609/aaai.v34i05.6239>
- [74] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “Hellaswag: Can a machine really finish your sentence?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [75] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, “SWAG: A large-scale adversarial dataset for grounded commonsense inference,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 93–104. [Online]. Available: <https://doi.org/10.18653/v1/d18-1009>
- [76] S. Zhang, R. Rudinger, K. Duh, and B. V. Durme, “Ordinal common-sense inference,” *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 379–395, 2017. [Online]. Available: [https://doi.org/10.1162/tacl\\_a\\_00068](https://doi.org/10.1162/tacl_a_00068)
- [77] C. Elkan and R. Greiner, “D. b. lenat and r. v. guha, building large knowledge-based systems: Representation and inference in the cyc project,” *Artif. Intell.*, vol. 61, no. 1, pp. 41–52, 1993. [Online]. Available: [https://doi.org/10.1016/0004-3702\(93\)90092-P](https://doi.org/10.1016/0004-3702(93)90092-P)
- [78] M. Sicilia, E. García-Barriocanal, S. Sánchez-Alonso, and M. E. Rodríguez, “On integrating learning object metadata inside the opencyc knowledge base,” in *Proceedings of the IEEE International Conference on Advanced Learning Technologies, ICALT 2004, Joensuu, Finland, August 30 - September 1, 2004*, Kinshuk, C. Looi, E. Sutinen, D. G. Sampson, I. Aedo, L. Uden, and E. Kähkönen, Eds. IEEE Computer Society, 2004. [Online]. Available: <https://doi.org/10.1109/ICALT.2004.1357711>
- [79] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, “Open mind common sense: Knowledge acquisition from the general public,” in *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings*, ser. Lecture Notes in Computer Science, R. Meersman and Z. Tari, Eds., vol. 2519. Springer, 2002, pp. 1223–1237. [Online]. Available: [https://doi.org/10.1007/3-540-36124-3\\_77](https://doi.org/10.1007/3-540-36124-3_77)
- [80] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 4444–4451. [Online]. Available: <https://doi.org/10.1609/aaai.v31i1.11164>
- [81] H. Zhang, X. Liu, H. Pan, Y. Song, and C. W. Leung, “ASER: A large-scale eventuality knowledge graph,” in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Y. Huang, I. King, T. Liu, and M. van Steen, Eds. ACM / IW3C2, 2020, pp. 201–211. [Online]. Available: <https://doi.org/10.1145/3366423.3380107>
- [82] Y. Ma, Z. Wang, M. Li, Y. Cao, M. Chen, X. Li, W. Sun, K. Deng, K. Wang, A. Sun, and J. Shao, “MMEKG: multi-modal event knowledge graph towards universal representation across modalities,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, V. Basile, Z. Kozareva, and S. Stajner, Eds. Association for Computational Linguistics, 2022, pp. 231–239. [Online]. Available: <https://doi.org/10.18653/v1/2022.acl-demo.23>
- [83] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, “ATOMIC: an atlas of machine commonsense for if-then reasoning,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in*



- Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 3027–3035. [Online]. Available: <https://doi.org/10.1609/aaai.v33i01.33013027>
- [84] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi, “Visualcomet: Reasoning about the dynamic context of a still image,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12350. Springer, 2020, pp. 508–524. [Online]. Available: [https://doi.org/10.1007/978-3-030-58558-7\\_30](https://doi.org/10.1007/978-3-030-58558-7_30)
- [85] E. Cambria, R. Speer, C. Havasi, and A. Hussain, “Senticnet: A publicly available semantic resource for opinion mining,” in *Commonsense Knowledge, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*, ser. AAAI Technical Report, vol. FS-10-02. AAAI, 2010. [Online]. Available: <http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2216>
- [86] S. Heindorf, Y. Scholten, H. Wachsmuth, A. N. Ngomo, and M. Potthast, “Causenet: Towards a causality graph extracted from the web,” in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, M. d’Aquino, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, Eds. ACM, 2020, pp. 3023–3030. [Online]. Available: <https://doi.org/10.1145/3340531.3412763>
- [87] C. Yu, W. Wang, X. Liu, J. Bai, Y. Song, Z. Li, Y. Gao, T. Cao, and B. Yin, “Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery,” in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 1173–1191. [Online]. Available: <https://doi.org/10.18653/v1/2023.findings-acl.76>
- [88] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [89] S. Chen, S. Wong, L. Chen, and Y. Tian, “Extending context window of large language models via positional interpolation,” *CoRR*, vol. abs/2306.15595, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2306.15595>
- [90] C. An, S. Gong, M. Zhong, M. Li, J. Zhang, L. Kong, and X. Qiu, “L-eval: Instituting standardized evaluation for long context language models,” 2023.
- [91] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li, “LongBench: A bilingual, multitask benchmark for long context understanding,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3119–3137. [Online]. Available: <https://aclanthology.org/2024.acl-long.172>
- [92] X. Zhang, Y. Chen, S. Hu, Z. Xu, J. Chen, M. Hao, X. Han, Z. Thai, S. Wang, Z. Liu, and M. Sun, “Infinitebench: Extending long context evaluation beyond 100K tokens,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15262–15277. [Online]. Available: <https://aclanthology.org/2024.acl-long.814>
- [93] Y. Bai, S. Tu, J. Zhang, H. Peng, X. Wang, X. Lv, S. Cao, J. Xu, L. Hou, Y. Dong, J. Tang, and J. Li, “Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks,” *CoRR*, vol. abs/2412.15204, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2412.15204>
- [94] X. Li, Y. Cao, Y. Ma, and A. Sun, “Long context vs. rag for llms: An evaluation and revisits,” 2024. [Online]. Available: <https://arxiv.org/abs/2501.01880>
- [95] Q. Lyu, L. Zhang, and C. Callison-Burch, “Goal-oriented script construction,” in *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, A. Belz, A. Fan, E. Reiter, and Y. Sripada, Eds. Association for Computational Linguistics, 2021, pp. 184–200. [Online]. Available: <https://doi.org/10.18653/v1/2021.inlg-1.19>
- [96] M. Koupaei and W. Y. Wang, “Wikihow: A large scale text summarization dataset,” *CoRR*, vol. abs/1810.09305, 2018. [Online]. Available: <http://arxiv.org/abs/1810.09305>
- [97] X. Li, Y. Cao, M. Chen, and A. Sun, “Take a break in the middle: Investigating subgoals towards hierarchical script generation,” in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 10129–10147. [Online]. Available: <https://doi.org/10.18653/v1/2023.findings-acl.644>
- [98] K. Valmeekam, M. Marquez, A. O. Hernandez, S. Sreedharan, and S. Kambhampati, “Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets_and_Benchmarks.html)

- [99] Y. Shen, K. Song, X. Tan, W. Zhang, K. Ren, S. Yuan, W. Lu, D. Li, and Y. Zhuang, "Taskbench: Benchmarking large language models for task automation," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2024/hash/085185ea97db31ae6dcac7497616fd3e-Abstract-Datasets\\_and\\_Benchmarks\\_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/085185ea97db31ae6dcac7497616fd3e-Abstract-Datasets_and_Benchmarks_Track.html)
- [100] H. S. Zheng, S. Mishra, H. Zhang, X. Chen, M. Chen, A. Nova, L. Hou, H. Cheng, Q. V. Le, E. H. Chi, and D. Zhou, "NATURAL PLAN: benchmarking llms on natural language planning," *CoRR*, vol. abs/2406.04520, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.04520>
- [101] O. Styles, S. Miller, P. Cerda-Mardini, T. Guha, V. Sanchez, and B. Vidgen, "Workbench: a benchmark dataset for agents in a realistic workplace setting," *CoRR*, vol. abs/2405.00823, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2405.00823>
- [102] S. Huang, W. Zhong, J. Lu, Q. Zhu, J. Gao, W. Liu, Y. Hou, X. Zeng, Y. Wang, L. Shang, X. Jiang, R. Xu, and Q. Liu, "Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 4363–4400. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.259>
- [103] S. Yao, H. Chen, J. Yang, and K. Narasimhan, "Webshop: Towards scalable real-world web interaction with grounded language agents," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html)
- [104] J. Xie, K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su, "Travelplanner: A benchmark for real-world planning with language agents," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=l5XQzNkAOe>
- [105] Y. Wu, X. Tang, T. M. Mitchell, and Y. Li, "Smartplay: A benchmark for llms as intelligent agents," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=S2oTVrlcp3>
- [106] K. Valmeekam, S. Sreedharan, M. Marquez, A. O. Hernandez, and S. Kambhampati, "On the planning abilities of large language models (A critical investigation with a proposed benchmark)," *CoRR*, vol. abs/2302.06706, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2302.06706>
- [107] K. Valmeekam, A. O. Hernandez, S. Sreedharan, and S. Kambhampati, "Large language models still can't plan (A benchmark for llms on planning and reasoning about change)," *CoRR*, vol. abs/2206.10498, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2206.10498>
- [108] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, and et al, "Do as i can, not as i say: Grounding language in robotic affordances," 2022. [Online]. Available: <https://arxiv.org/abs/2204.01691>
- [109] M. Côté, Á. Kádár, X. Yuan, B. Kybartas, T. Barnes, E. Fine, J. Moore, M. J. Hausknecht, L. E. Asri, M. Adada, W. Tay, and A. Trischler, "Textworld: A learning environment for text-based games," in *Computer Games - 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers*, ser. Communications in Computer and Information Science, T. Cazenave, A. Saffidine, and N. R. Sturtevant, Eds., vol. 1017. Springer, 2018, pp. 41–75. [Online]. Available: [https://doi.org/10.1007/978-3-030-24337-1\\_3](https://doi.org/10.1007/978-3-030-24337-1_3)
- [110] M. Savva, J. Malik, D. Parikh, D. Batra, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, and V. Koltun, "Habitat: A platform for embodied AI research," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 9338–9346. [Online]. Available: <https://doi.org/10.1109/ICCV.2019.00943>
- [111] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain et al., "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *arXiv preprint arXiv:2108.03272*, 2021.
- [112] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," 2020. [Online]. Available: <https://arxiv.org/abs/1912.01734>
- [113] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. J. Martinez, H. Yin, M. Lingelbach, M. Hwang, A. Hiranaka, S. Garlanka, A. Aydin, S. Lee, J. Sun, M. Anvari, M. Sharma, D. Bansal, S. Hunter, K. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, Y. Li, S. Savarese, H. Gweon, C. K. Liu, J. Wu, and L. Fei-Fei, "BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation," *CoRR*, vol. abs/2403.09227, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2403.09227>
- [114] M. Shridhar, X. Yuan, M. Côté, Y. Bisk, A. Trischler,

- and M. J. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=0IOX0YcCdTn>
- [115] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Learning-from-demonstrations benchmark for generalizable manipulation skills," *CoRR*, abs/2107.14483, 2021b. URL <https://arxiv.org/abs/2107.14483>, vol. 14483, 2021.
- [116] Z. Wang, M. Wu, Y. Cao, Y. Ma, M. Chen, and T. Tuytelaars, "Navigating the nuances: A fine-grained evaluation of vision-language navigation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 4681–4704.
- [117] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [118] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, "Vision language models are blind," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 18–34.
- [119] M. Sharma, "Exploring and improving the spatial reasoning abilities of large language models," *CoRR*, vol. abs/2312.01054, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2312.01054>
- [120] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *CoRR*, vol. abs/2112.03227, 2021. [Online]. Available: <https://arxiv.org/abs/2112.03227>
- [121] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [122] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3470–3487. [Online]. Available: <https://aclanthology.org/2022.acl-long.244>
- [123] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=9Vrb9D0Wl4>
- [124] Y. K. Chia, P. Hong, L. Bing, and S. Poria, "InstructEval: Towards holistic evaluation of instruction-tuned large language models," in *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, A. V. Miceli-Barone, F. Barez, S. Cohen, E. Voita, U. Germann, and M. Lukasik, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 35–64. [Online]. Available: <https://aclanthology.org/2024.scalellm-1.4>
- [125] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*, 2021.
- [126] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, and X. Shen, "Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5085–5109. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.340>
- [127] D. Eccleston, "Sharegpt," 2023, accessed: 2025-01-22. [Online]. Available: <https://sharegpt.com/>
- [128] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin, "Free dolly: Introducing the world's first truly open instruction-tuned llm," *Company Blog of Databricks*, 2023.
- [129] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. M. Nguyen, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. A. Glushkov, A. V. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. J. Mattick, "Openassistant conversations - democratizing large language model alignment," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: <https://openreview.net/forum?id=VSJotgbPHF>
- [130] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.
- [131] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpaca-eval: An automatic evaluator of instruction-following models," [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval), 5 2023.
- [132] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%\*

- chatgpt quality," March 2023. [Online]. Available: <https://lmsys.org/blog/2023-03-30-vicuna/>
- [133] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica, "From crowdsourced data to high-quality benchmarks: Arena-hard and benchmark-builder pipeline," *arXiv preprint arXiv:2406.11939*, 2024.
- [134] T. Li, A. Angelopoulos, and W.-L. Chiang, "Does style matter? disentangling style and substance in chatbot arena," *LMSYS Blog*, 2024, \*Equal contribution. [Online]. Available: <https://lmsys.org/blog/2024-08-28-style-control/>
- [135] G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the judge? a study on judgement bias," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8301–8327. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.474>
- [136] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-controlled alpaca-eval: A simple way to debias automatic evaluators," *arXiv preprint arXiv:2404.04475*, 2024.
- [137] R. Park, R. Rafailov, S. Ermon, and C. Finn, "Disentangling length from quality in direct preference optimization," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 4998–5017. [Online]. Available: <https://aclanthology.org/2024.findings-acl.297>
- [138] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," *arXiv preprint arXiv:2311.07911*, 2023.
- [139] Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu, and W. Wang, "Follow-bench: A multi-level fine-grained constraints following benchmark for large language models," *arXiv preprint arXiv:2310.20410*, 2023.
- [140] B. Y. Lin, Y. Deng, K. Chandu, F. Brahman, A. Ravichander, V. Pyatkin, N. Dziri, R. L. Bras, and Y. Choi, "Wildbench: Benchmarking llms with challenging tasks from real users in the wild," *arXiv preprint arXiv:2406.04770*, 2024.
- [141] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv:2306.13394*, 2023.
- [142] J. Li, W. Lu, H. Fei, M. Luo, M. Dai, M. Xia, Y. Jin, Z. Gan, D. Qi, C. Fu *et al.*, "A survey on benchmarks of multimodal large language models," *arXiv preprint arXiv:2408.08632*, 2024.
- [143] J. Huang and J. Zhang, "A survey on evaluation of multimodal large language models," *arXiv preprint arXiv:2408.15769*, 2024.
- [144] X.AI, "Grok-1.5 vision preview," <https://x.ai/blog/grok-1.5v>, 2024.
- [145] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, J. Lei, Q. Lu, R. Chen, P. Xu, R. Zhang, H. Zhang, P. Gao, Y. Wang, Y. Qiao, P. Luo, K. Zhang, and W. Shao, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," *arXiv:2404.16006*, 2024.
- [146] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun *et al.*, "Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark," *arXiv preprint arXiv:2409.02813*, 2024.
- [147] M. Chen, Y. Cao, Y. Zhang, and C. Lu, "Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16 449–16 469. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.960>
- [148] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *arXiv:2403.20330*, 2024.
- [149] Y. Liu, H. Duan, B. L. Yuanhan Zhang, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" *arXiv:2307.06281*, 2023.
- [150] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou, "Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models," *arXiv:2310.14566*, 2023.
- [151] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *WACV*, 2021.
- [152] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *CVPR*, 2019.
- [153] J. Tang, Q. Liu, Y. Ye, J. Lu, S. Wei, C. Lin, W. Li, M. F. F. B. Mahmood, H. Feng, Z. Zhao, Y. Wang, Y. Liu, H. Liu, X. Bai, and C. Huang, "Mtvqa: Benchmarking multilingual text-centric visual question answering," *arXiv:2405.11985*, 2024.
- [154] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," *arXiv:2203.10244*, 2022.
- [155] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *ECCV*, 2016.
- [156] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna, "Blink: Multimodal large language models can see but not perceive," in *European Conference on Computer Vision*. Springer, 2024, pp. 148–166.
- [157] X. Wang, Y. Zhou, X. Liu, H. Lu, Y. Xu, F. He, J. Yoon, T. Lu, G. Bertasius, M. Bansal *et al.*, "Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences," *arXiv preprint arXiv:2401.10529*, 2024.
- [158] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang,

- S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of CVPR*, 2024.
- [159] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, "Mantis: Interleaved multi-image instruction tuning," *Transactions on Machine Learning Research*, 2024.
- [160] J. Li, K. Pan, Z. Ge, M. Gao, W. Ji, W. Zhang, T.-S. Chua, S. Tang, H. Zhang, and Y. Zhuang, "Fine-tuning multimodal llms to follow zero-shot demonstrative instructions," in *ICLR*, 2024.
- [161] F. Wang, X. Fu, J. Y. Huang, Z. Li, Q. Liu, X. Liu, M. D. Ma, N. Xu, W. Zhou, K. Zhang *et al.*, "Muirbench: A comprehensive benchmark for robust multi-image understanding," in *ICLR*, 2025.
- [162] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024.
- [163] V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, M. Malinowski, Y. Yang, C. Doersch *et al.*, "Perception test: A diagnostic benchmark for multimodal video models," in *NeurIPS*, 2024.
- [164] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," in *NeurIPS*, 2023.
- [165] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv:2405.21075*, 2024.
- [166] X. Ma, Y. Gao, Y. Wang, R. Wang, X. Wang, Y. Sun, Y. Ding, H. Xu, Y. Chen, Y. Zhao *et al.*, "Safety at scale: A comprehensive survey of large model safety," *arXiv preprint arXiv:2502.05206*, 2025.
- [167] P. Röttger, F. Pernisi, B. Vidgen, and D. Hovy, "Safety prompts: a systematic review of open datasets for evaluating and improving large language model safety," *CoRR*, vol. abs/2404.05399, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2404.05399>
- [168] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," *arXiv preprint arXiv:2203.09509*, 2022.
- [169] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.301/>
- [170] Z. Lin, Z. Wang, Y. Tong, Y. Wang, Y. Guo, Y. Wang, and J. Shang, "ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4694–4702. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.311/>
- [171] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 678–24 704, 2024.
- [172] H. Sun, G. Xu, J. Deng, J. Cheng, C. Zheng, H. Zhou, N. Peng, X. Zhu, and M. Huang, "On the safety of conversational models: Taxonomy, dataset, and benchmark," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3906–3923. [Online]. Available: <https://aclanthology.org/2022.findings-acl.308/>
- [173] E. Fleisig, A. Amstutz, C. Atalla, S. L. Blodgett, H. Daumé III, A. Olteanu, E. Sheng, D. Vann, and H. Wallach, "FairPrism: Evaluating fairness-related harms in text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6231–6251. [Online]. Available: <https://aclanthology.org/2023.acl-long.343/>
- [174] D. Ganguli, N. Schiefer, M. Favaro, and J. Clark. (2023) Challenges in evaluating AI systems. [Online]. Available: <https://www.anthropic.com/index/evaluating-ai-systems>
- [175] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li, "Decodingtrust: A comprehensive assessment of trustworthiness in GPT models," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2023/hash/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets_and_Benchmarks.html)
- [176] S. Ghosh, P. Varshney, E. Galinkin, and C. Parisien, "AEGIS: online adaptive AI content safety moderation with ensemble of LLM experts," *CoRR*, vol. abs/2404.05993, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2404.05993>
- [177] T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwag, K. Huang, L. He, B. Wei, D. Li, Y. Sheng, R. Jia, B. Li, K. Li, D. Chen, P. Henderson, and P. Mittal, "Sorry-bench: Systematically evaluating large language model safety refusal behaviors," *CoRR*, vol. abs/2406.14598, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.14598>
- [178] W. Wang, Z. Tu, C. Chen, Y. Yuan, J. Huang,

- W. Jiao, and M. R. Lyu, "All languages matter: On the multilingual safety of large language models," *CoRR*, vol. abs/2310.00905, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2310.00905>
- [179] X. Yuan, J. Li, D. Wang, Y. Chen, X. Mao, L. Huang, H. Xue, W. Wang, K. Ren, and J. Wang, "S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models," *CoRR*, vol. abs/2405.14191, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2405.14191>
- [180] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15 537–15 553. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.830>
- [181] Y. Mou, S. Zhang, and W. Ye, "Sg-bench: Evaluating LLM safety generalization across diverse tasks and prompt types," *CoRR*, vol. abs/2410.21965, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2410.21965>
- [182] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, and J. Shao, "Salad-bench: A hierarchical and comprehensive safety benchmark for large language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 3923–3954. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.235>
- [183] W. Zhang, X. Lei, Z. Liu, M. An, B. Yang, K. Zhao, K. Wang, and S. Lian, "Chisafetybench: A chinese hierarchical safety benchmark for large language models," *CoRR*, vol. abs/2406.10311, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.10311>
- [184] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, and et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *CoRR*, vol. abs/2204.05862, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2204.05862>
- [185] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning ai with shared human values," *CoRR*, 2020.
- [186] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: Evaluating safeguards in llms," in *EACL (Findings)*, 2024.
- [187] H. Koh, D. Kim, M. Lee, and K. Jung, "Can llms recognize toxicity? a structured investigation framework and toxicity metric," *arXiv preprint arXiv:2402.06900*, 2024.
- [188] X. Pang, S. Tang, R. Ye, Y. Xiong, B. Zhang, Y. Wang, and S. Chen, "Self-alignment of large language models via monopolylogue-based social scene simulation," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 39 416–39 447.
- [189] T. Raheja, N. Pochhi, and F. Curie, "Recent advancements in llm red-teaming: Techniques, defenses, and ethical considerations," *arXiv preprint arXiv:2410.09097*, 2024.
- [190] N. Wichers, C. Denison, and A. Beirami, "Gradient-based language model red teaming," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 2862–2881.
- [191] A. F. Hardy, H. Liu, B. Lange, and M. J. Kochenderfer, "Astprompter: Weakly supervised automated language model red-teaming to identify likely toxic prompts," *arXiv preprint arXiv:2407.09447*, 2024.
- [192] Y. Yang, Z. Xiao, X. Lu, H. Wang, H. Huang, G. Chen, and Y. Chen, "Sop: Unlock the power of social facilitation for automatic jailbreak attack," *arXiv preprint arXiv:2407.01902*, 2024.
- [193] J. Yu, X. Lin, Z. Yu, and X. Xing, "Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts," *arXiv preprint arXiv:2309.10253*, 2023.
- [194] T. Zhang, B. Cao, Y. Cao, L. Lin, P. Mitra, and J. Chen, "Wordgame: Efficient & effective llm jailbreak via simultaneous obfuscation in query and response," *arXiv preprint arXiv:2405.14023*, 2024.
- [195] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *CoRR*, vol. abs/2307.15043, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2307.15043>
- [196] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1671–1685.
- [197] B. Radharapu, K. Robinson, L. Aroyo, and P. Lahoti, "AART: ai-assisted red-teaming with diverse data generation for new llm-powered applications," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*, M. Wang and I. Zitouni, Eds. Association for Computational Linguistics, 2023, pp. 380–395. [Online]. Available: <https://doi.org/10.18653/v1/2023.emnlp-industry.37>
- [198] D. Esiobu, X. E. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. M. Smith, "ROBBIE: robust bias evaluation of large generative language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 3764–3814. [Online]. Available: <https://doi.org/10.18653/v1/2023.emnlp-main.230>
- [199] G. Kour, M. Zalmanovici, N. Zwerdling, E. Goldbraich, O. Fandina, A. Anaby Tavor, O. Raz, and E. Farchi, "Unveiling safety vulnerabilities of large



- language models,” in *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 111–127. [Online]. Available: <https://aclanthology.org/2023.gem-1.10/>
- [200] C. Liu, F. Zhao, L. Qing, Y. Kang, C. Sun, K. Kuang, and F. Wu, “Goal-oriented prompt attack and safety evaluation for llms,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.11830>
- [201] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, and B. Li, “Alert: A comprehensive benchmark for assessing large language models’ safety through red teaming,” 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2404.08676>
- [202] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, and et al., “Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned,” *CoRR*, vol. abs/2209.07858, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2209.07858>
- [203] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, “Bot-adversarial dialogue for safe conversational agents,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 2950–2968. [Online]. Available: <https://aclanthology.org/2021.naacl-main.235/>
- [204] X. Liu, N. Xu, M. Chen, and C. Xiao, “Autodan: Generating stealthy jailbreak prompts on aligned large language models,” in *ICLR*, 2024.
- [205] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models,” *arXiv preprint arXiv:2307.15043*, 2023.
- [206] X. Liu, P. Li, E. Suh, Y. Vorobeychik, Z. Mao, S. Jha, P. McDaniel, H. Sun, B. Li, and C. Xiao, “Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms,” in *ICLR*, 2025.
- [207] W. Luo, S. Dai, X. Liu, S. Banerjee, H. Sun, M. Chen, and C. Xiao, “Agrail: A lifelong agent guardrail with effective and adaptive safety detection,” *arXiv preprint arXiv:2502.11448*, 2025.
- [208] Z. Xiang, L. Zheng, Y. Li, J. Hong, Q. Li, H. Xie, J. Zhang, Z. Xiong, C. Xie, C. Yang *et al.*, “Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning,” *arXiv preprint arXiv:2406.09187*, 2024.
- [209] C. Xu, M. Kang, J. Zhang, Z. Liao, L. Mo, M. Yuan, H. Sun, and B. Li, “Advweb: Controllable black-box attacks on vlm-powered web agents,” *arXiv preprint arXiv:2410.17401*, 2024.
- [210] Z. Liao, L. Mo, C. Xu, M. Kang, J. Zhang, C. Xiao, Y. Tian, B. Li, and H. Sun, “Eia: Environmental injection attack on generalist web agents for privacy leakage,” *arXiv preprint arXiv:2409.11295*, 2024.
- [211] W. Shi, R. Xu, Y. Zhuang, Y. Yu, J. Zhang, H. Wu, Y. Zhu, J. Ho, C. Yang, and M. D. Wang, “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, pp. 22 315–22 339.
- [212] Z. Zhang, S. Cui, Y. Lu, J. Zhou, J. Yang, H. Wang, and M. Huang, “Agent-safetybench: Evaluating the safety of llm agents,” *arXiv preprint arXiv:2412.14470*, 2024.
- [213] H. Zhang, J. Huang, K. Mei, Y. Yao, Z. Wang, C. Zhan, H. Wang, and Y. Zhang, “Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents,” *arXiv preprint arXiv:2410.02644*, 2024.
- [214] T. Yuan, Z. He, L. Dong, Y. Wang, R. Zhao, T. Xia, L. Xu, B. Zhou, F. Li, Z. Zhang *et al.*, “R-judge: Benchmarking safety risk awareness for llm agents,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 1467–1490.
- [215] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2369–2380.
- [216] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, “Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 346–361, 2021.
- [217] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, “Ok-vqa: A visual question answering benchmark requiring external knowledge,” in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.
- [218] K. Wang, J. Pan, W. Shi, Z. Lu, M. Zhan, and H. Li, “Measuring multimodal mathematical reasoning with math-vision dataset,” *arXiv:2402.14804*, 2024.
- [219] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- [220] Z. Xu, C. Feng, R. Shao, T. Ashby, Y. Shen, D. Jin, Y. Cheng, Q. Wang, and L. Huang, “Vision-flan: Scaling human-labeled tasks in visual instruction tuning,” in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 15 271–15 342.
- [221] J. Huang, J. Zhang, K. Jiang, H. Qiu, and S. Lu, “Visual instruction tuning towards general-purpose multimodal model: A survey,” *arXiv preprint arXiv:2312.16602*, 2023.
- [222] S. Lin, J. Hilton, and O. Evans, “Truthfulqa: Measuring how models mimic human falsehoods,” *arXiv preprint arXiv:2109.07958*, 2021.
- [223] Z. Ying, A. Liu, S. Liang, L. Huang, J. Guo, W. Zhou, X. Liu, and D. Tao, “Safebench: A safety evaluation framework for multimodal large language models,” *arXiv preprint arXiv:2410.18927*, 2024.

- [224] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupala, and A. Alishahi, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: <https://aclanthology.org/W18-5446/>
- [225] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.
- [226] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.
- [227] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, 2024.
- [228] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *Transactions on Machine Learning Research*, 2022.
- [229] T. Lee, H. Tu, C. H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. Roberts, M. Yasunaga, H. Yao, C. Xie *et al.*, "Vhelm: A holistic evaluation of vision language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 140 632–140 666, 2024.
- [230] Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, J. Zhang, J. Li, and L. Hou, "Benchmarking foundation models with language-model-as-an-examiner," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 78 142–78 167. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/f64e55d03e2fe61aa4114e49cb654acb-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/f64e55d03e2fe61aa4114e49cb654acb-Paper-Datasets_and_Benchmarks.pdf)
- [231] R. Zhao, W. Zhang, Y. K. Chia, W. Xu, D. Zhao, and L. Bing, "Auto-arena: Automating llm evaluations with agent peer battles and committee discussions," 2024. [Online]. Available: <https://arxiv.org/abs/2405.20267>
- [232] J. Cheng, Y. Lu, X. Gu, P. Ke, X. Liu, Y. Dong, H. Wang, J. Tang, and M. Huang, "Autodetect: Towards a unified framework for automated weakness detection in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, 2024. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.397>
- [233] C. Siro, Y. Yuan, M. Aliannejadi, and M. de Rijke, "AGENT-CQ: automatic generation and evaluation of clarifying questions for conversational search with llms," *CoRR*, vol. abs/2410.19692, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2410.19692>
- [234] J. Zhang, J. Wu, Y. Teng, M. Liao, N. Xu, X. Xiao, Z. Wei, and D. Tang, "Android in the zoo: Chain-of-action-thought for gui agents," *arXiv:2403.02713*, 2024.
- [235] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *CVPR*, 2020.
- [236] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma *et al.*, "Fine-tuning large vision-language models as decision-making agents via reinforcement learning," *arXiv:2405.10292*, 2024.
- [237] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.
- [238] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," in *CVPR*, 2020.
- [239] R. Li, B. Wang, R. Li, and X. Du, "IQA-EVAL: automatic evaluation of human-model interactive question answering," *CoRR*, vol. abs/2408.13545, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2408.13545>
- [240] J. Zheng, H. Wang, A. Zhang, T. D. Nguyen, J. Sun, and T. Chua, "Ali-agent: Assessing llms' alignment with human values via agent-based evaluation," *CoRR*, vol. abs/2405.14125, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2405.14125>
- [241] C. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better llm-based evaluators through multi-agent debate," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=FQepisCUWu>
- [242] Y. Li, S. Zhang, R. Wu, X. Huang, Y. Chen, W. Xu, G. Qi, and D. Min, "Mateval: A multi-agent discussion framework for advancing open-ended text evaluation," in *International Conference on Database Systems for Advanced Applications*, ser. Lecture Notes in Computer Science, M. Onizuka, J. Lee, Y. Tong, C. Xiao, Y. Ishikawa, S. Amer-Yahia, H. V. Jagadish, and K. Lu, Eds., vol. 14856, Springer. Springer, 2024, pp. 415–426. [Online]. Available: [https://doi.org/10.1007/978-981-97-5575-2\\_31](https://doi.org/10.1007/978-981-97-5575-2_31)
- [243] J. Lin, H. Zhao, A. Zhang, Y. Wu, H. Ping, and Q. Chen, "Agentsims: An open-source sandbox for large language model evaluation," *CoRR*, vol. abs/2308.04026, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.04026>
- [244] A. Dulny, A. Hotho, and A. Krause, "Dynabench: A benchmark dataset for learning dynamical systems from low-resolution data," in *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part I*, ser. Lecture Notes in Computer Science, D. Koutra,

- C. Plant, M. Gomez-Rodriguez, E. Baralis, and F. Bonchi, Eds., vol. 14169. Springer, 2023, pp. 438–455. [Online]. Available: [https://doi.org/10.1007/978-3-031-43412-9\\_26](https://doi.org/10.1007/978-3-031-43412-9_26)
- [245] W. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. Stoica, “Chatbot arena: An open platform for evaluating llms by human preference,” in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=3MW8GKNyZl>
- [246] J. Liu, C. S. Xia, Y. Wang, and L. Zhang, “Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html)
- [247] A. Myrzakhan, S. M. Bsharat, and Z. Shen, “Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena,” *CoRR*, vol. abs/2406.07545, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.07545>
- [248] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, Y. Fu, M. Sun, and J. He, “C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: [http://papers.nips.cc/paper\\_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html)
- [249] W. Chen, M. Yin, M. Ku, P. Lu, Y. Wan, X. Ma, J. Xu, X. Wang, and T. Xia, “TheoremQA: A theorem-driven question answering dataset,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7889–7901. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.489/>
- [250] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang, “SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models,” in *Proceedings of the Forty-First International Conference on Machine Learning*, 2024.
- [251] D. Song, S. Chen, G. H. Chen, F. Yu, X. Wan, and B. Wang, “Milebench: Benchmarking mllms in long context,” *arXiv preprint arXiv:2404.18532*, 2024.
- [252] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://openreview.net/forum?id=wCu6T5xFjeJ>
- [253] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “MTEB: Massive text embedding benchmark,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. [Online]. Available: <https://aclanthology.org/2023.eacl-main.148/>
- [254] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, and P. Colombo, “Colpali: Efficient document retrieval with vision language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.01449>
- [255] J. Ying, Y. Cao, K. Xiong, L. Cui, Y. He, and Y. Liu, “Intuitive or dependent? investigating llms’ behavior style to conflicting prompts,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 4221–4246.
- [256] J. Chen, T. Liang, S. Siu, Z. Wang, K. Wang, Y. Wang, Y. Ni, W. Zhu, Z. Jiang, B. Lyu, D. Jiang, X. He, Y. Liu, H. Hu, X. Yue, and W. Chen, “Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.10563>
- [257] W. Zhang, M. Aljunied, C. Gao, Y. K. Chia, and L. Bing, “M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 5484–5505, 2023.
- [258] X. Wang, Z. Wang, J. Liu, Y. Chen, L. Yuan, H. Peng, and H. Ji, “Mint: Evaluating llms in multi-turn interaction with tools and language feedback,” 2024. [Online]. Available: <https://arxiv.org/abs/2309.10691>
- [259] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, “ELI5: Long form question answering,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3558–3567. [Online]. Available: <https://aclanthology.org/P19-1346/>
- [260] H. Su, H. Yen, M. Xia, W. Shi, N. Muennighoff, H. Yu Wang, H. Liu, Q. Shi, Z. S. Siegel, M. Tang, R. Sun, J. Yoon, S. O. Arik, D. Chen, and T. Yu, “Bright: A realistic and challenging benchmark for reasoning-intensive retrieval,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.12883>
- [261] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, “HybridQA: A dataset of multi-hop question answering over tabular and textual data,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1026–1036. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.91/>
- [262] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova,

- D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang, "FinQA: A dataset of numerical reasoning over financial data," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3697–3711. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.300/>
- [263] Y. Ma, Z. Gou, J. Hao, R. Xu, S. Wang, L. Pan, Y. Yang, Y. Cao, and A. Sun, "SciAgent: Tool-augmented language models for scientific reasoning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15701–15736. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.880>
- [264] Y. Ma, Y. Zang, L. Chen, M. Chen, Y. Jiao, X. Li, X. Lu, Z. Liu, Y. Ma, X. Dong, P. Zhang, L. Pan, Y.-G. Jiang, J. Wang, Y. Cao, and A. Sun, "Mmlongbench-doc: Benchmarking long-context document understanding with visualizations," 2024. [Online]. Available: <https://arxiv.org/abs/2407.01523>
- [265] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu, and M. Sun, "Visrag: Vision-based retrieval-augmented generation on multi-modality documents," 2024. [Online]. Available: <https://arxiv.org/abs/2410.10594>
- [266] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning large multimodal models with factually augmented rlhf," 2023. [Online]. Available: <https://arxiv.org/abs/2309.14525>
- [267] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto, "s1: Simple test-time scaling," *arXiv preprint arXiv:2501.19393*, 2025.
- [268] W. Chen, M. wei Chang, E. Schlinger, W. Wang, and W. Cohen, "Open question answering over tables and text," *Proceedings of ICLR 2021*, 2021.
- [269] R. Tito, D. Karatzas, and E. Valveny, "Hierarchical multimodal transformers for multi-page docvqa," 2023. [Online]. Available: <https://arxiv.org/abs/2212.05935>
- [270] B. Zheng, B. Gou, J. Kil, H. Sun, and Y. Su, "Gpt-4v(ision) is a generalist web agent, if grounded," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: <https://openreview.net/forum?id=pieckJ2DIB>
- [271] J. Yang, C. E. Jimenez, A. L. Zhang, K. Lieret, J. Yang, X. Wu, O. Press, N. Muennighoff, G. Synnaeve, K. R. Narasimhan, D. Yang, S. I. Wang, and O. Press, "Swe-bench multimodal: Do ai systems generalize to visual software domains?" 2024. [Online]. Available: <https://arxiv.org/abs/2410.03859>
- [272] X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su, "Mind2web: Towards a generalist agent for the web," 2023.
- [273] X. Ma, S. Zhuang, B. Koopman, G. Zuccon, W. Chen, and J. Lin, "Visa: Retrieval augmented generation with visual source attribution," 2024. [Online]. Available: <https://arxiv.org/abs/2412.14457>
- [274] J. Cho, D. Mahata, O. Irsoy, Y. He, and M. Bansal, "M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding," 2024. [Online]. Available: <https://arxiv.org/abs/2411.04952>
- [275] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant, "Multimodal{qa}: complex question answering over text, tables and images," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=ee6W5UgQLa>
- [276] Z. Z. Wang, A. Asai, X. V. Yu, F. F. Xu, Y. Xie, G. Neubig, and D. Fried, "Coderag-bench: Can retrieval augment code generation?" 2024. [Online]. Available: <https://arxiv.org/abs/2406.14497>
- [277] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1710722>
- [278] W. Sun, C. Zhang, X. Zhang, X. Yu, Z. Huang, P. Chen, H. Xu, S. He, J. Zhao, and K. Liu, "Beyond instruction following: Evaluating inferential rule following of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.08440>
- [279] F. Wang, X. Fu, J. Y. Huang, Z. Li, Q. Liu, X. Liu, M. D. Ma, N. Xu, W. Zhou, K. Zhang, T. L. Yan, W. J. Mo, H.-H. Liu, P. Lu, C. Li, C. Xiao, K.-W. Chang, D. Roth, S. Zhang, H. Poon, and M. Chen, "Muirbench: A comprehensive benchmark for robust multi-image understanding," 2024. [Online]. Available: <https://arxiv.org/abs/2406.09411>
- [280] C. Li, H. Peng, X. Wang, Y. Qi, L. Hou, B. Xu, and J. Li, "MAVEN-FACT: A large-scale event factuality detection dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11140–11158. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.651/>
- [281] K. Yang, A. Swope, A. Gu, R. Chalamala, P. Song, S. Yu, S. Godil, R. Prenger, and A. Anandkumar, "Le-anDojo: Theorem proving with retrieval-augmented language models," in *Neural Information Processing Systems (NeurIPS)*, 2023.
- [282] J. Kasai, K. Sakaguchi, yoichi takahashi, R. L. Bras, A. Asai, X. V. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui, "Realtime QA: What's the answer right now?" in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: <https://openreview.net/forum?id=HfKOIPCvsv>
- [283] S. Quan, J. Yang, B. Yu, B. Zheng, D. Liu, A. Yang, X. Ren, B. Gao, Y. Miao, Y. Feng, Z. Wang, J. Yang, Z. Cui, Y. Fan, Y. Zhang, B. Hui, and J. Lin, "Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings," 2025. [Online]. Available: <https://arxiv.org/abs/2501.01257>

- [284] P. Clark, O. Tafjord, and K. Richardson, "Transformers as soft reasoners over language," 2020. [Online]. Available: <https://arxiv.org/abs/2002.05867>
- [285] W. Wang, S. Zhang, Y. Ren, Y. Duan, T. Li, S. Liu, M. Hu, Z. Chen, K. Zhang, L. Lu *et al.*, "Needle in a multimodal haystack," *arXiv preprint arXiv:2406.07230*, 2024.
- [286] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh, "Obelics: An open web-scale filtered dataset of interleaved image-text documents," 2023.
- [287] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 292–305. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.20/>
- [288] N. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, J. Yang, M. Zhang, Z. Zhang, W. Ouyang, K. Xu, W. Huang, J. Fu, and J. Peng, "RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 14743–14777. [Online]. Available: <https://aclanthology.org/2024.findings-acl.878/>
- [289] G. Li, H. A. Al Kader Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: communicative agents for "mind" exploration of large language model society," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.
- [290] G. Bai, J. Liu, X. Bu, Y. He, J. Liu, Z. Zhou, Z. Lin, W. Su, T. Ge, B. Zheng, and W. Ouyang, "MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7421–7454. [Online]. Available: <https://aclanthology.org/2024.acl-long.401/>
- [291] A. Gu, B. Rozière, H. Leather, A. Solar-Lezama, G. Synnaeve, and S. I. Wang, "Cruxeval: A benchmark for code reasoning, understanding and execution," 2024. [Online]. Available: <https://arxiv.org/abs/2401.03065>
- [292] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, and M. Sun, "Toolllm: Facilitating large language models to master 16000+ real-world apis," 2023.
- [293] A. Mishra, A. Asai, V. Balachandran, Y. Wang, G. Neubig, Y. Tsvetkov, and H. Hajishirzi, "Fine-grained hallucination detection and editing for language models," in *First Conference on Language Modeling*, 2024. [Online]. Available: <https://openreview.net/forum?id=dJMTn3QOWO>
- [294] Y. Liu, Z. Yao, R. Min, Y. Cao, L. Hou, and J. Li, "Rm-bench: Benchmarking reward models of language models with subtlety and style," 2024. [Online]. Available: <https://arxiv.org/abs/2410.16184>
- [295] L. Li, Y. Wei, Z. Xie, X. Yang, Y. Song, P. Wang, C. An, T. Liu, S. Li, B. Y. Lin, L. Kong, and Q. Liu, "Vlrewardbench: A challenging benchmark for vision-language generative reward models," 2024. [Online]. Available: <https://arxiv.org/abs/2411.17451>
- [296] Z. Jin, H. Yuan, T. Men, P. Cao, Y. Chen, K. Liu, and J. Zhao, "Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment," 2024. [Online]. Available: <https://arxiv.org/abs/2412.13746>
- [297] T. Zhu, Q. Liu, F. Wang, Z. Tu, and M. Chen, "Unraveling cross-modality knowledge conflicts in large vision-language models," *arXiv preprint arXiv:2410.03659*, 2024.
- [298] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, Q. Lin, and D. Jiang, "Wizardlm: Empowering large pre-trained language models to follow complex instructions," in *The Twelfth International Conference on Learning Representations*, 2024.
- [299] J. Ying, Y. Cao, Y. Bai, Q. Sun, B. Wang, W. Tang, Z. Ding, Y. Yang, X. Huang, and S. Yan, "Automating dataset updates towards reliable and timely evaluation of large language models," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 17106–17132. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/1e89c12621c0315373f20f0aeabe5dbe-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/1e89c12621c0315373f20f0aeabe5dbe-Paper-Datasets_and_Benchmarks_Track.pdf)
- [300] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.
- [301] Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu *et al.*, "Benchmarking foundation models with language-model-as-an-examiner," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [302] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang, "Agentbench: Evaluating llms as agents," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=zAdUB0aCTQ>
- [303] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," 2023. [Online]. Available: <https://arxiv.org/abs/2307.16125>
- [304] Y. Jiao, M. Zhong, S. Li, R. Zhao, S. Ouyang,

- H. Ji, and J. Han, "Instruct and extract: Instruction tuning for on-demand information extraction," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10030–10051. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.620/>
- [305] P. Laban, A. Fabbri, C. Xiong, and C.-S. Wu, "Summary of a haystack: A challenge to long-context LLMs and RAG systems," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 9885–9903. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.552/>
- [306] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, and Z. Sui, "A survey on in-context learning," 2024. [Online]. Available: <https://arxiv.org/abs/2301.00234>
- [307] X. Yue, X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mammoth: Building math generalist models through hybrid instruction tuning," 2023. [Online]. Available: <https://arxiv.org/abs/2309.05653>
- [308] D. Zhang, Z. Hu, S. Zhoubian, Z. Du, K. Yang, Z. Wang, Y. Yue, Y. Dong, and J. Tang, "Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models," 2024. [Online]. Available: <https://arxiv.org/abs/2401.07950>
- [309] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.
- [310] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A theoretical analysis of ndcg type ranking measures," in *Conference on learning theory*. PMLR, 2013, pp. 25–54.
- [311] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [312] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
- [313] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [314] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
- [315] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.
- [316] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [317] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin et al., "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.
- [318] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma et al., "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [319] X. Ye, S. Iyer, A. Celikyilmaz, V. Stoyanov, G. Durrett, and R. Pasunuru, "Complementary explanations for effective in-context learning," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4469–4484. [Online]. Available: <https://aclanthology.org/2023.findings-acl.273/>
- [320] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," *arXiv preprint arXiv:2111.02080*, 2021.
- [321] N. Kotonya, S. Krishnasamy, J. Tetreault, and A. Jaimes, "Little giants: Exploring the potential of small LLMs as evaluation metrics in summarization in the Eval4NLP 2023 shared task," in *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, D. Deutsch, R. Dror, S. Eger, Y. Gao, C. Leiter, J. Opitz, and A. Rücklé, Eds. Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 202–218. [Online]. Available: <https://aclanthology.org/2023.eval4nlp-1.17>
- [322] H. Hasanbeig, H. Sharma, L. Betthausen, F. V. Frujeri, and I. Momennejad, "Allure: A systematic protocol for auditing and improving llm-based evaluation of text using iterative in-context-learning," *arXiv preprint arXiv:2309.13701*, 2023.
- [323] M. Song, M. Zheng, and X. Luo, "Can many-shot in-context learning help long-context llm judges? see more, judge better!" *arXiv preprint arXiv:2406.11629*, 2024.
- [324] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.
- [325] C.-H. Chiang and H.-y. Lee, "A closer look into using large language models for automatic evaluation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8928–8942. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.599>
- [326] T. Y. Zhuo, "Ice-score: Instructing large language models to evaluate code," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 2232–2242.
- [327] S. Saha, O. Levy, A. Celikyilmaz, M. Bansal, J. Weston, and X. Li, "Branch-solve-merge improves large language model evaluation and generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,



- K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 8352–8370. [Online]. Available: <https://aclanthology.org/2024.naacl-long.462>
- [328] H. He, H. Zhang, and D. Roth, “SocREval: Large language models with the socratic method for reference-free reasoning evaluation,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2736–2764. [Online]. Available: <https://aclanthology.org/2024.findings-naacl.175/>
- [329] J. Li, S. Sun, W. Yuan, R.-Z. Fan, H. Zhao, and P. Liu, “Generative judge for evaluating alignment,” *arXiv preprint arXiv:2310.05470*, 2023.
- [330] Z. Yu, C. Gao, W. Yao, Y. Wang, W. Ye, J. Wang, X. Xie, Y. Zhang, and S. Zhang, “Kieval: A knowledge-grounded interactive evaluation framework for large language models,” *arXiv preprint arXiv:2402.15043*, 2024.
- [331] M. Gao, J. Ruan, R. Sun, X. Yin, S. Yang, and X. Wan, “Human-like summarization evaluation with chatgpt,” *arXiv preprint arXiv:2304.02554*, 2023.
- [332] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, “Prometheus 2: An open source language model specialized in evaluating other language models,” *arXiv preprint arXiv:2405.01535*, 2024.
- [333] S. Jain, V. Keshava, S. M. Sathyendra, P. Fernandes, P. Liu, G. Neubig, and C. Zhou, “Multi-dimensional evaluation of text summarization with in-context learning,” *arXiv preprint arXiv:2306.01200*, 2023.
- [334] H. Song, H. Su, I. Shalymov, J. Cai, and S. Mansour, “Finesure: Fine-grained summarization evaluation using llms,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 906–922.
- [335] Y. Liu, T. Yang, S. Huang, Z. Zhang, H. Huang, F. Wei, W. Deng, F. Sun, and Q. Zhang, “Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition,” *arXiv preprint arXiv:2402.15754*, 2024.
- [336] X. Hu, M. Gao, S. Hu, Y. Zhang, Y. Chen, T. Xu, and X. Wan, “Are llm-based evaluators confusing nlg quality criteria?” *arXiv preprint arXiv:2402.12055*, 2024.
- [337] Y. Liu, T. Yang, S. Huang, Z. Zhang, H. Huang, F. Wei, W. Deng, F. Sun, and Q. Zhang, “Calibrating llm-based evaluator,” in *LREC/COLING*, 2024.
- [338] Y. R. Dong, T. Hu, and N. Collier, “Can llm be a personalized judge?” *arXiv preprint arXiv:2406.11657*, 2024.
- [339] G. Aslanyan and U. Porwal, “Position bias estimation for unbiased learning-to-rank in ecommerce search,” in *String Processing and Information Retrieval: 26th International Symposium, SPIRE 2019, Segovia, Spain, October 7–9, 2019, Proceedings 26*. Springer, 2019, pp. 47–64.
- [340] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.
- [341] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie *et al.*, “Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization,” *arXiv preprint arXiv:2306.05087*, 2023.
- [342] R. Zhao, W. Zhang, Y. K. Chia, D. Zhao, and L. Bing, “Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions,” *arXiv preprint arXiv:2405.20267*, 2024.
- [343] Z. Chu, Q. Ai, Y. Tu, H. Li, and Y. Liu, “Pre: A peer review based large language model evaluator,” *arXiv preprint arXiv:2401.15641*, 2024.
- [344] K.-P. Ning, S. Yang, Y.-Y. Liu, J.-Y. Yao, Z.-H. Liu, Y.-H. Tian, Y. Song, and L. Yuan, “Pico: Peer review in llms based on the consistency optimization,” *arXiv preprint arXiv:2402.01830*, 2024.
- [345] B. Patel, S. Chakraborty, W. A. Suttle, M. Wang, A. S. Bedi, and D. Manocha, “Aime: Ai system optimization via multiple llm evaluators,” *arXiv preprint arXiv:2410.03131*, 2024.
- [346] Y. Gao, G. Xu, D. Z. Wang, and A. Cohan, “Bayesian calibration of win rate estimation with llm evaluators,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 4757–4769.
- [347] Z. Hu, J. Zhang, Z. Xiong, A. Ratner, H. Xiong, and R. Krishna, “Language model preference evaluation with multiple weak evaluators,” *arXiv preprint arXiv:2410.12869*, 2024.
- [348] X. Zhang, B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li, “Wider and deeper llm networks are fairer llm evaluators,” *arXiv preprint arXiv:2308.01862*, 2023.
- [349] Z. Xu, S. Shi, B. Hu, J. Yu, D. Li, M. Zhang, and Y. Wu, “Towards reasoning in large language models via multi-agent peer review collaboration,” *arXiv preprint arXiv:2311.08152*, 2023.
- [350] S. Liang, B. Zhang, J. Zhao, and K. Liu, “Abseval: An agent-based framework for script evaluation,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 12 418–12 434.
- [351] C. Bandi and A. Harrasse, “Adversarial multi-agent evaluation of large language models through iterative debates,” *arXiv preprint arXiv:2410.04663*, 2024.
- [352] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, “Chateval: Towards better llm-based evaluators through multi-agent debate,” *arXiv preprint arXiv:2308.07201*, 2023.
- [353] R. Li, T. Patel, and X. Du, “Prd: Peer rank and discussion improve large language model based evaluations,” *arXiv preprint arXiv:2307.02762*, 2023.
- [354] J. Jung, F. Brahman, and Y. Choi, “Trust or escalate: Llm judges with provable guarantees for human agreement,” *arXiv preprint arXiv:2407.18370*, 2024.
- [355] H. Huang, Y. Qu, J. Liu, M. Yang, and T. Zhao, “An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers,” *arXiv e-prints*, pp. arXiv–2403, 2024.
- [356] Q. Pan, Z. Ashktorab, M. Desmond, M. S. Cooper, J. Johnson, R. Nair, E. Daly, and W. Geyer, “Human-centered design recommendations for llm-as-a-judge,” in *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, 2024, pp. 16–29.

- [357] Y. Zhang, P. Ren, and M. de Rijke, "A human-machine collaborative framework for evaluating malevolence in dialogues," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 5612–5623. [Online]. Available: <https://aclanthology.org/2021.acl-long.436/>
- [358] Q. Li, L. Cui, L. Kong, and W. Bi, "Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation," *arXiv preprint arXiv:2310.19740*, 2023.
- [359] S. Shankar, J. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, and I. Arawjo, "Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–14.
- [360] M. T. Ribeiro and S. Lundberg, "Adaptive testing and debugging of NLP models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3253–3267. [Online]. Available: <https://aclanthology.org/2022.acl-long.230/>
- [361] C. Rastogi, M. Tulio Ribeiro, N. King, H. Nori, and S. Amershi, "Supporting human-ai collaboration in auditing llms with llms," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 913–926.
- [362] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu *et al.*, "Large language models are not fair evaluators," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9440–9450.
- [363] T. Wang, P. Yu, X. E. Tan, S. O'Brien, R. Pasunuru, J. Dwivedi-Yu, O. Golovneva, L. Zettlemoyer, M. Fazel-Zarandi, and A. Celikyilmaz, "Shepherd: A critic for language model generation," *arXiv preprint arXiv:2308.04592*, 2023.
- [364] T. Vu, K. Krishna, S. Alzubi, C. Tar, M. Faruqui, and Y.-H. Sung, "Foundational autoraters: Taming large language models for better automatic evaluation," *arXiv preprint arXiv:2407.10817*, 2024.
- [365] L. Zhu, X. Wang, and X. Wang, "Judgelm: Fine-tuned large language models are scalable judges," *arXiv preprint arXiv:2310.17631*, 2023.
- [366] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne *et al.*, "Prometheus: Inducing fine-grained evaluation capability in language models," in *The Twelfth International Conference on Learning Representations*, 2023.
- [367] D. Jiang, Y. Li, G. Zhang, W. Huang, B. Y. Lin, and W. Chen, "Tigerscore: Towards building explainable metric for all text generation tasks," *arXiv preprint arXiv:2310.00752*, 2023.
- [368] W. Xu, D. Wang, L. Pan, Z. Song, M. Freitag, W. Wang, and L. Li, "INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5967–5994. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.365/>
- [369] Y. Liu, Y. Zheng, S. Xia, J. Li, Y. Tu, C. Song, and P. Liu, "Safety-j: Evaluating safety with critique," *arXiv preprint arXiv:2407.17075*, 2024.
- [370] W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike, "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.
- [371] B. Wang, S. Chern, E. Chern, and P. Liu, "Halu-j: Critique-based hallucination judge," *arXiv preprint arXiv:2407.12943*, 2024.
- [372] X. Hu, L. Lin, M. Gao, X. Yin, and X. Wan, "Themis: A reference-free nlg evaluation language model with flexibility and interpretability," *arXiv preprint arXiv:2406.18365*, 2024.
- [373] T. Wu, W. Yuan, O. Golovneva, J. Xu, Y. Tian, J. Jiao, J. Weston, and S. Sukhbaatar, "Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge," *arXiv preprint arXiv:2407.19594*, 2024.
- [374] A. Shypula, S. Li, B. Zhang, V. Padmakumar, K. Yin, and O. Bastani, "Evaluating the diversity and quality of llm generated content," *arXiv preprint arXiv:2504.12522*, 2025.
- [375] N. Chen, Z. Hu, Q. Zou, J. Wu, Q. Wang, B. Hooi, and B. He, "Judgelm: Large reasoning models as a judge," *arXiv preprint arXiv:2504.00050*, 2025.
- [376] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui, "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, pp. 9426–9439.
- [377] L. Pacchiardi, L. G. Cheke, and J. Hernández-Orallo, "100 instances is all you need: predicting the success of a new llm on unseen data by testing on a few instances," 2024. [Online]. Available: <https://arxiv.org/abs/2409.03563>
- [378] D. Chen, Q. Yu, P. Wang, W. Zhang, B. Tang, F. Xiong, X. Li, M. Yang, and Z. Li, "xverify: Efficient answer verifier for reasoning model evaluations," *arXiv preprint arXiv:2504.10481*, 2025.
- [379] Y. Chen, J. Benton, A. Radhakrishnan, J. U. C. Denison, J. Schulman, A. Somani, P. Hase, M. W. F. R. V. Mikulik, S. Bowman, J. L. J. Kaplan *et al.*, "Reasoning models don't always say what they think," *Anthropic Research*, 2025.