OmniBench: A Comprehensive Benchmark Integrating Real-World, Time-sensitive, and Multi-Hop Questions with a Multi-Dimensional Hybrid Evaluation Framework

Wenjie Wang*, Yufeng Jiang*, Ge Sun*, Chenghang Dong*, Zheng Jun, Li Mengjie, Lixin Chen, Huan Wang†, Haoyu Wang, Bin Chen

Ant Group

{xiaowen.wwj, qingyue, dongchenghang.dch, zange.sg, muxing.zj, daisy.lmj, cora.clx, huan.wh, sunyi.why, burningchen.cb}@antgroup.com

Abstract

Recently, with the increasing capabilities of Large Language Models (LLMs), AI applications have gradually emerged to solve various problems in people's daily lives, so accurately measuring their performance and reliability is paramount. However, existing benchmarks predominantly rely on closed-ended, multiple-choice or short-answer question formats. While useful for assessment, these formats exhibit a significant gap compared to the diverse and open-ended nature of questions posed by real-world users. To bridge this gap, we produce OmniBench, a comprehensive open-domain benchmark. OmniBench is uniquely composed of authentic, user-generated questions harvested from real-world interactions on various websites and applications, covering 16 rigorously defined knowledge domains and 5 crucial user intents derived from a large-scale analysis of the mass corpus. Crucially, we propose three automated data construction pipelines that enable the continuous and periodic updating of the benchmark dataset. This approach not only ensures that the questions can keep up with current events, but also effectively mitigates the critical issue of data contamination prevalent in static benchmarks. Moreover, a multi-dimensional hybrid evaluation framework named OmniEval is proposed for evaluating the responses. This framework combines diverse metrics and evaluation methods to capture nuanced aspects of answer performance. Extensive validation demonstrates that this evaluation framework exhibits strong alignment with human judgments, ensuring the reliability of the benchmark results.

Introduction

The enhanced capabilities of Large Language Models(LLMs) such as o3(OpenAI 2025) and Gemini 2.5(Gemini 2025) have spurred a flourishing of AI applications. However, effectively evaluating these applications' performance remains a significant challenge, particularly for answering open-domain questions encountered in people's daily life. Widely adopted datasets such as MMLU(Hendrycks et al. 2020), GPQA(Rein et al. 2023) and Humanity's Last Exam(Phan et al. 2025) have become cornerstones to evaluate knowledge mastery across broad subjects. However, these established benchmarks exhibit three critical limitations: **1). Format Gap**: The overwhelming reliance on closed-ended multiple-choice (MCQ) or short-answer question formats creates a significant disparity from authentic user interactions, and the evaluation methods cannot assess the open-ended answer accurately. **2). Narrow Evaluation Scope**: These benchmarks focus narrowly on isolated capabilities, unlike real-world user queries that require integrated knowledge retrieval and complex reasoning. **3). Outdated and Data Contamination**: Infrequent updates leave static datasets vulnerable to data leakage and unable to keep up with current events. Open-LLM-Leaderboard(Myrzakhan, Bsharat, and Shen 2024) reduces bias in MCQs by converting datasets like CommonsenseQA(Talmor et al. 2019) to open-response formats. However, it still falls short in brideging the gap with authentic user requests and is vulnerable to data contamination on CommonsenseQA. AlignBench(Liu et al. 2024) and Wildbench(Lin et al. 2024) use open-ended questions. Yet, their evaluation methods risk bias: AlignBench relies on manually annotated reference answers, which may be incomplete, especially for long responses. Wildbench uses checklists, which may be inadequate for highly open-ended or novel questions, leading to inaccurate evaluation results.

To address these limitations, we introduce OmniBench, a comprehensive benchmark featuring open-ended, time-sensitive (weekly-updated), and multi-hop questions across 16 knowledge domains and 5 core user intents. We establish three automated pipelines for dataset construction and regular updates: Open-ended QA construction involves continuous data collection from popular websites and apps, such as Baidu, Rednote, and Zhihu, and processing through multiple steps, including query selection, attribute annotation, checklist generation, and human review; Time-sensitive QA pipeline generates questions and checklists from the latest news and government policy documents, with weekly updates. Multi-hop questions aim to evaluate the integrated capability of retrieval and reasoning. In this pipeline, we first extract triplets from the given documents, then use search tools to gather supplementary documents about the objects

*These authors contributed equally.

†Corresponding Author.

Benchmark	Dataset		Evaluation Method			
	Question Format	Automated Construction	Metric	Checklist-based	Rubric-based	Fact-checking
MMLU	Multiple-choice	✗	Accuracy	✗	✗	✗
HLE	Multiple-choice, Short-answer	✗	Accuracy	✗	✗	✗
OLL	Open-ended	✗	Accuracy	✗	✓	✗
AlignBench	Open-ended	✗	Multi-dimentional	✗	✓	✗
Wildbench	Open-ended	✓	Multi-dimentional	✓	✓	✗
OmniBench(ours)	**Open-ended**	✓	**Multi-dimentional**	✓	✓	✓

Table 1: Comparison between OmniBench and other representative benchmarks. "HLE" represents Humanity's Last Exam and "OLL" respresents Open-LLM-Leaderboard.

within these triplets. Finally, LLMs are used to generate multi-hop questions by synthesising information across documents.

In this paper, we propose OmniEval, a multi-dimensional hybrid framework with two orthogonal metric categories: Basic Metrics (Factuality, Completeness, Relevance) evaluating essential qualities of an answer; Advanced Metrics (Richness, Practicability, Insight, Inspiration) dynamically adapting to the five core user intents for higher-order value assessment. We apply the most suitable evaluation strategy to each dimension. Checklist-based evaluation combined with a Likert scale assesses Factuality and Completeness, where the evaluation criteria are unambiguous and easily judged by LLMs. Fact-checking evaluates factual claims extending beyond the predefined checklist in the dimension of Factuality. For all the remaining indicators, predefining an exhaustive checklist is difficult and would severely constrain evaluation accuracy. Therefore, we employ a rubric-based method to assess responses, applying specific rules rather than relying on mechanical checklist matching. OmniEval achieves a strong correlation with human judgment (Kendall's τ-b = 0.851) after the additional experiments. Table1 presents a comparison between OmniBench and other benchmarks in terms of dataset and evaluation method.

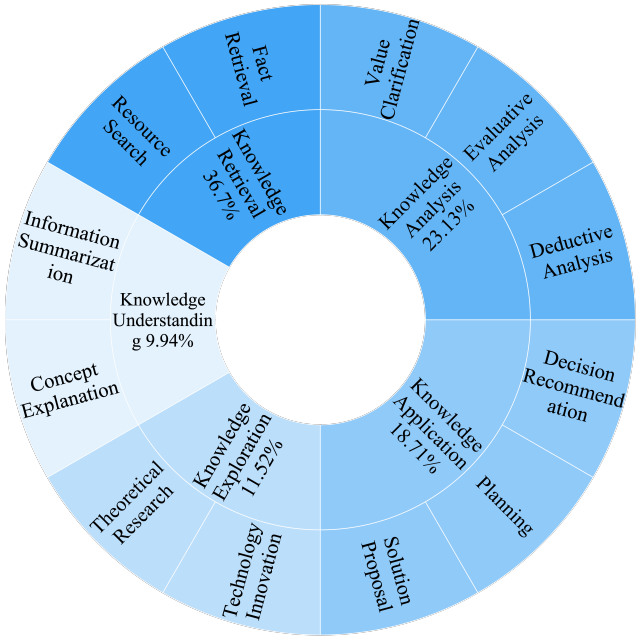

Figure 1: The user intent and sub-intent categories of OmniBench.

OmniBench

Dataset Statistics

To effectively assess AI applications' capability in addressing real-world problems, evaluation datasets must accurately reflect authentic human information needs. Through analysis of 60,000 real user questions, we established a multi-dimensional taxonomy of **Knowledge Domain** and **User Intent** that captures the diversity of cognitive processes and knowledge requirements in real-world scenarios. OmniBench contains 1,267 Chinese data samples (examples in Appendix A).

OmniBench systematically covers real-world scenarios through 16 knowledge domains (e.g., Digital technology, Arts and Entertainment) and 86 sub-domains, ensuring comprehensive knowledge breadth coverage. The distribution of domains and sub-domains is shown in Appendix B. Fact retrieval versus higher-order cognitive tasks such as complex reasoning and knowledge application, impose divergent capability requirements on AI applications. Following the concept of Bloom's Taxonomy, we classified five intents: Knowledge Retrieval, Knowledge Understanding, Knowledge Application, Knowledge Analysis, and Knowledge Exploration, corresponding to Remember, Understand, Apply, Evaluate, and Create in Bloom's Taxonomy. These were subdivided into 12 sub-intents to capture nuanced cognitive processes, providing a theoretical basis for evaluation metric design. The distribution of intents and sub-intents is shown in Figure 1.

Existing benchmarks target objective, closed-ended questions, whereas real-world queries require handling open-ended tasks like recommendations and planning. Therefore, we construct three specialized QA datasets: 1) Open-ended QA (79%) for real-world capability evaluation ; 2) Time-sensitive QA (4%) with weekly news updates assessing timeliness ; and 3) Multi-hop QA (17%) testing deep search/reasoning in complex knowledge network. Figure2 depicts the automated pipeline, incorporating manual validation at Step4 for quality control. Some question generation

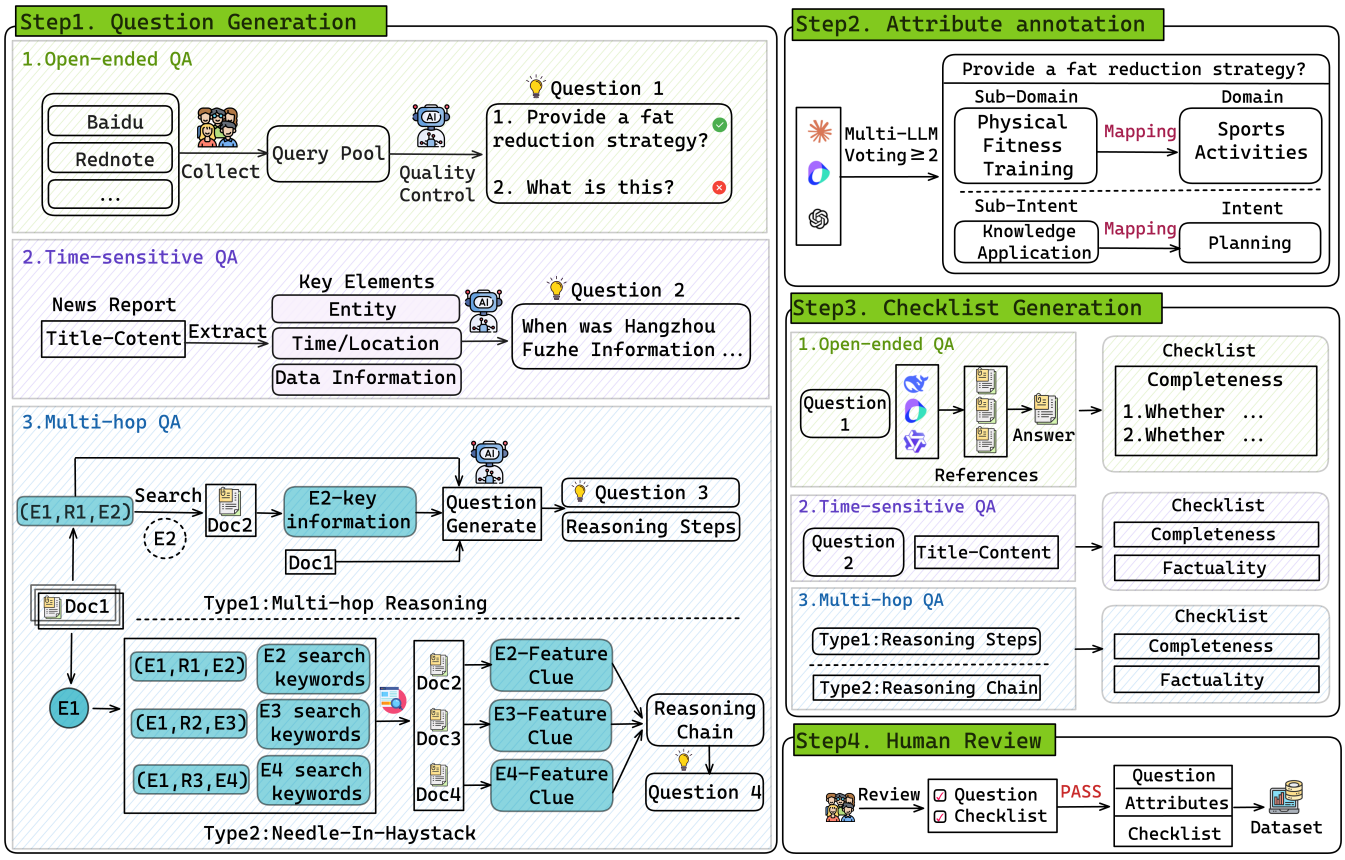

Figure 2: The OmniBench production pipeline includes: (1) Question Generation, (2) Attribute Annotation, (3) Checklist Generation, and (4) Human Review. Steps 1 and 3 use customized strategies for each data type, while steps 2 and 4 maintain consistent procedures across all categories.

prompt examples are provided in Appendix F.

Open-Ended QA

Data Collection Real user questions were collected from popular websites and apps (e.g., Baidu, Rednote, Zhihu), with LLM-based quality control based on the following principles: 1) Clear subject, 2) Explicit intent, and 3) Exclusion of code generation, identity recognition, security risks, and creative writing.

Attribute Annotation This section involves annotation across two dimensions: sub-intent, sub-domain. We employ a multi-LLM voting mechanism using three LLMs (Doubao-128k-pro, Claude-3.7, and GPT-4o), selecting results with at least 2 votes as final labels. Samples without consensus are excluded. Top-level intents and domains are obtained directly from predefined mappings.

Checklist Generation OmniBench uses checklist-based evaluation for Completeness and Factuality metrics (see Section Evaluation Metric for details). The construction process comprises three steps:

- Generate multiple references using multiple LLMs with RAG capabilities (e.g., Doubao, Qwen, DeepSeek-R1).
- Claude-3.7 integrates multiple references to generate standard answers by: 1) Retaining consensus content, 2) Integrating complementary information, and 3) Eliminating redundancies. This process ensures answer completeness and rigor.
- Generate a checklist based on key points and factual claims identified in the standard answer. Each checklist item must verify one independent element, be clearly stated, and contain no factual errors. Template: `whether + [verb] + [checkpoint] + [verb] + [fact point]`. Non-compliant items are manually revised.

Time-Sensitive QA

Question Generation Extract key elements from news content centered around the news titles. The key elements include entities time or location information, and data information, etc. Generate questions centered around the key elements, ensuring that each question has a unique and definitive objective answer within the content.

Checklist Generation Since the questions derive from the article's key elements, generating the checklist involves inputting both the questions and the news content into an LLM to extract checklist items.

Multi-Hop QA

Multi-Hop Reasoning

- **Triplet Extraction**: As shown in Figure 2, we use a LLM like GPT-4 to extract relational triplet (E_1, R_1, E_2) from the high-quality document Doc_1, where E_1, R_1, E_2 represent Entity1, Relation1, and Entity2 respectively. For example, from TV series document "Ren Changxia" on Baidu Baike (Doc_1), we extract triple *(Ren Changxia, director, Shen Haofang)*.
- **Document retrieval for** E_2: Use search tools to obtain the authoritative document (Doc_2) about tail entity E_2. For example, Shen Haofang's personal introduction on Baidu Baike.
- **Question Generation**: Feed Doc_1, Doc_2, and triplet into the LLM to generate questions that avoid explicitly naming E_2, instead referencing it through E_1 and R_1. Questions focus on key information about E_2 contained within Doc_2. Such as *"Which film directed by the director of TV series 'Ren Changxia' was released in 2004?"*
- **Checklist Generation**: The checklist is generated for the search information required at each hop of the question-answering process, with the answer for each hop being drawn from either Doc_1 or Doc_2. For example, the reasoning steps include:
  - **Step1**: Infer from Doc_1 that the director of "Ren Changxia" is Shen Haofang.
  - **Step2**: Extract from Doc_2 the films directed by Shen Haofang and their release years.
  - **Step3**: Compare the years to identify the the film released in 2004 is "Courtyard No.5".

  Checklist items cover key factual points in each reasoning step to evaluate whether AI applications can establish complete multi-hop reasoning chains.

Needle In A Haystack

- **Multiple relational triplet extraction**: We designate center entity E_c (serving as the answer entity) from high-quality document Doc_1 and extract heterogeneous relational triplets: $[\{(E_c, R_i, E_j)\} \mid i \in [1, n]]$. Tail entities must be concrete persons, locations, events, competitions, or literary works with semantically independent relations. For example, from film document "She's Got No Name" (Doc_1), we extract triplets: {*(She's Got No Name, director, Peter Chan)*, *(She's Got No Name, leading actress, Zhang Ziyi)*, *(She's Got No Name, selected for, 77th Cannes International Film Festival)*}.
- **Tail Entity Feature Clue Generation**: From Doc_1, we extract tail entity descriptions and generate 2-3 search keywords (e.g., "Peter Chan director"). Through these search keywords, relevant documents $Doc_2 - Doc_n$ pertaining to tail entities are retrieved, and distinctive feature clues of tail entities are extracted to support the following reasoning chain: Tail entity feature clue $\rightarrow$ Tail entity $\rightarrow$ Relation $\rightarrow$ Center entity. For example, Zhang Ziyi's feature clue is *"Graduated from the Performance Department of Central Academy of Drama, nominated for Academy Award for Best Actress for "Crouching Tiger, Hidden Dragon"*.
- **Question Generation**: Multiple logical reasoning chains from tail entity feature clues to the center entity are integrated, with tail entity deliberately obfuscated, as shown in Figure 2. Through clue descriptions of tail entities and their relationships with center entity, the target center entity is identified. Examples of generated question can be found in the Appendix A.
- **Checklist Generation**: The target answer corresponds to the "center entity". The checklist encompasses both the target answer and the tail entities from each relational triplet.

Multi-Dimensional Hybrid Evaluation Framework

We introduce **OmniEval**, a multi-dimensional hybrid evaluation framework designed for a comprehensive, precise, and human-aligned assessment of complex real-world questions. By strategically employing a diverse set of state-of-the-art judge models (e.g., GPT-4o, Claude-3.7-Sonnet, Gemini-2.5-Pro) for these distinct tasks, OmniEval ensures a robust, reliable, and multifaceted evaluation process. An example of OmniEval evaluation pipeline is shown in Figure 3.

Evaluation Metric

Multi-Dimensional Since an answer may excel in certain aspects while falling short in others, our metric system is designed to be hierarchical and multi-dimensional, enabling an independent diagnosis of LLM performance across different capabilities. The system comprises two categories of top-level orthogonal metric: basic and advanced metric. Basic metrics represent the fundamental requirements an answer must satisfy, while advanced metrics constitute more sophisticated enhancements built upon this foundation. Secondly, within each top-level metric category, there are also second-level orthogonal metric categories.

- **Basic metrics** are designed as fundamental capability metrics that assess the performance of the basic information and the quality of the core content of the model's answer, including:
  - **Factuality** (*Fact.*): The degree to which the information provided by the model's response conforms to the objective facts.
  - **Completeness** (*Comp.*): The comprehensive extent to which the content provided by the model's response covers all the information required for the question.
  - **Relevance** (*Rel.*): The correspondence between the model's response and the semantic intent of the question.
- **Advanced metrics** advanced metrics are customized as advanced capability metrics that judge the performance of additional and higher-order cognitive levels beyond the primary capability, based on the categorization of user intents mentioned before, including:
  - **Richness** (*Rich.*): The comprehensiveness and breadth of the model responses.
  - **Practicality** (*Prac.*): The actionability and practical application value of the model responses.

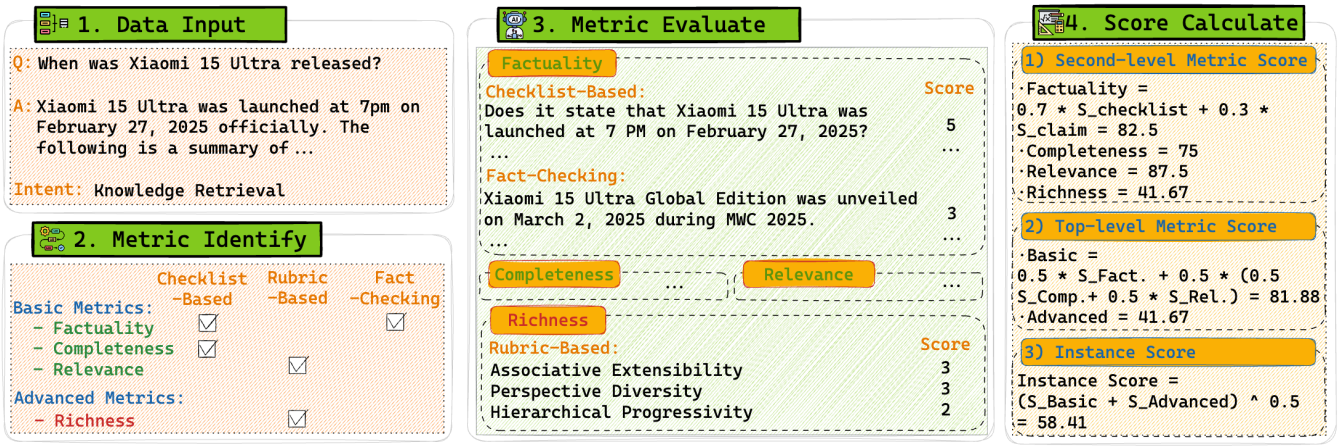

Figure 3: The OmniEval evaluation pipeline includes: (1) Data Input, (2) Metric Identify, (3) Metric Evaluate, (4) Score Calculate. The multi-dimensional metrics to be evaluated are identified from the input question's intent, and are evaluated with Hybrid LLM-as-a-Judge methods. Some evaluating prompt examples are provided in Appendix F.

- **Insight** (*Insi.*): The deep understanding and insight into the nature of the problem of the model responses.
- **Inspiration** (*Insp.*): The ability of the model's responses to stimulate users to think deeply and expand their cognition.

Intent-Oriented Another core design principle of our evaluation metrics is intent-oriented. We believed that LLMs should not be judged using a one-size-fits-all method, but rather should adapt dynamically to the specific intent of the user's question. From the user's perspective, the gold standard for a really 'useful' answer adaptively depends on the user's underlying purpose. Therefore, in our evaluation metric system, it is determined that the basic metrics must be evaluated for all questions of different intent categories, while the advanced metrics are supported to be combined specifically and then would be selected based on the intent categories as defined in the previous section. The table of metric mapping relationships is presented in Appendix C.

Fine-grained Atomic Dimensions We also pre-defined fine-grained atomic dimensions for the Relevance metric and all the advanced metrics, representing different aspects of the metrics to be evaluated. For example, the Relevance metric can be divided into two dimensions: Core Intent Responsiveness and Extended Content Redundancy. Furthermore, for those advanced metrics that are intent-specific, the definitions of the corresponding atomic dimensions will differ due to the different intent of the question to be evaluated. More details on all specific fine-grained atomic dimensions can be found in Appendix C. The design of these fine-grained atomic dimensions aims to systematically deconstruct the core connotations of each macro metric.

Evaluation Methodology

Checklist-Based We applied the Checklist-Based method to evaluate the Factuality and Completeness metrics. A checklist is defined as a list of instance-specific and verifiable items formulated as closed questions. Existing checklist methods use binary questions (yes/no) for each item, assuming model responses either fully meet or completely fail requirements. But we found most checklist items cannot be accurately judged this rigidly. For example, a response may be partially correct with minor errors or cover most points while missing details. In such cases, binary judgments are too simplistic to accurately evaluate LLMs. Therefore, to address this issue, we proposed an extension of the traditional binary judgments for Checklist-Based evaluation to a multi-level scoring system, based on a Likert scale. This system utilizes the 3-point scale for the Factuality metric (1, 3 and 5 representing fully incorrect, partially correct and fully correct), and a 5-point scale (1-5) for the Completeness metric, allowing the judge to provide more granular scores for each check item and improve the precision of evaluation results. And, to save resource consumption, we prompted the judge LLM to evaluate all check items under a given metric in a single pass instead of only one item at once.

Fact-Checking The necessity for a dedicated Fact-Checking process stems from two core limitations of Checklist-Based method: their impracticality for highly open-ended questions and their inability to capture "wild" factual claims outside the scope of the pre-defined checklists. Therefore, for the Factuality metric, we adopted the Fact-Checking method as a complementary evaluation approach. For the close-ended questions in Knowledge Retrieval or Knowledge Understanding, where objective facts can be pre-defined, the Factuality metric is assessed by both Checklist-Based and Fact-Checking methods. However, for another different set of questions (encompassing some from Knowledge Retrieval and Knowledge Understanding, and all from Knowledge Application, Knowledge Analysis, and Knowledge Exploration), their Factuality Score is derived solely from the Fact-Checking method. The Fact-Checking method consists of 4 main steps: Claim Extract → Query Rewrite → Query Search → Claim Judge. First, we prompt a judge model with high performance to extract as many verifiable, objective factual claims from the model's response

as possible, that are beyond the scope of the pre-defined checklist. Each claim is then rewritten as a valid and concise query. Relevant reference documents are obtained for each query from web search. Finally, the judge model evaluates whether each factual claim is correct based on the reference documents. The scoring criteria are the same as those of the Checklist-Based method for the Factuality metric.

Rubric-Based We applied the Rubric-Based method to the evaluation of the Relevance metric in the basic metrics and all the advanced metrics. Unlike the Factuality and Completeness metric, which can be efficiently broken down into a series of fixed and objective checklists, evaluating the Relevance metric and other advanced metrics such as Richness is highly holistic and dependent. Pre-defining an exhaustive checklist for these metrics is not only extremely difficult, but also severely limits the flexibility of the evaluation. Therefore, we adopted a rubric-based method, prompting the judge LLM to evaluate these metrics based on specific rubrics in the atomic dimensions instead of mechanical checklists. For each atomic dimension, we also developed specific fine-grained scoring criteria, using a 5-point scoring system. Generally to say, 5 corresponds to excellent, 4 to good, 3 to fair, 2 to poor, and 1 to bad. Details on some specific scoring criteria are presented in the prompts in Appendix F.

Overall Score Calculation First, for each second-level metric, depending on how it is evaluated, the average scores of n items within a checklist, multiple factual claims, or multiple atomic dimensions are computed separately, and the 5-point scores are converted to percentiles:

$$S_{\text{metric}} = (\frac{1}{n}\sum_{i=1}^{n} s_i - 1)/4 * 100$$

In particular, for the Factuality metric, if the intent of the question is Knowledge Retrieval or Knowledge Understanding, it is necessary to aggregate its Checklist-Based Score and Fact-Checking Score according to a weight w_1 (we set $w_1 = 0.7$ in our experiment):

$$S_{\textit{Fact.}} = w_1 \cdot S_{\text{Checklist-Based}} + (1 - w_1) \cdot S_{\text{Fact-Checking}}$$

While if the intent of the question is Knowledge Application, Knowledge analysis, and Knowledge Exploration, its Factuality Score is just the Fact-Checking Score.

Then, the scores of each second-level metric are aggregated into the score of top-level metric respectively. The Basic Metric score aggregates as a weighted average(we set $w_2 = 0.5$ in our experiment), while the Advanced Metric score uses a simple arithmetic mean:

$$S_{\text{Basic}} = w_2 \cdot S_{\textit{Fact.}} + (1 - w_2) \cdot \frac{n_{\textit{Comp.}}}{n_{\textit{Comp.}} + n_{\textit{Rel.}}} \cdot S_{\textit{Comp.}} + (1 - w_2) \cdot \frac{n_{\textit{Rel.}}}{n_{\textit{Comp.}} + n_{\textit{Rel.}}} \cdot S_{\textit{Rel.}}$$

$$S_{\text{Advanced}} = \frac{1}{m}\sum_{j=1}^{m} S_j$$

Next, the score of each instance is the geometric mean of its Basic Metric Score and Advanced Metric Score:

$$S_{\text{Instance}} = \sqrt{S_{\text{Basic}} \cdot S_{\text{Advanced}}}$$

Finally, for the whole OmniBench dataset, we provide the metric-wise average score as the Overall Score:

$$S_{\text{Overall}} = \frac{1}{7}\sum_{j=1}^{7}(\frac{1}{N_j}\sum_{i=1}^{N_j} S_{i,metric_j})$$

Experiment

Experiment Setup

We compared multiple AI applications—including Chinese models (Doubao, Yuanbao, Deepseek, Kimi) and international ones (OpenAI o4-mini, Gemini, Perplexity)—in general knowledge domain. Evaluations were run under two modes: Deep Thinking + Internet and Internet-only. This evaluated the impact of deep reasoning on Q&A performance and highlighted differences in information retrieval and integration among AI applications.

Result Analysis

Advanced metrics driving evaluation differentiation Table2 shows all AI products meet baseline proficiency in basic metrics (Factuality, Completeness, Relevance), with limited differentiation among them (e.g., a maximum difference of 16.78 points in completeness under deep thinking mode). In contrast, advanced metrics like insight demonstrate significant divergence, evidenced by a maximum score gap of 43.25 points in deep thinking mode. This underscores that advanced metrics are the primary drivers of substantial score differentials. DeepSeek's leading performance across advanced dimensions positions it at the forefront in average score under the deep thinking mode.

Chinese AI products show clear localization advantages In the Chinese context, Chinese AI applications perform better than foreign applications in dealing with general-purpose domain problems. This performance gap arises from several factors: Chinese AI applications are better optimized for Chinese contexts, ensuring more accurate understanding of questions in Chinese; they likely include more Chinese-language corpora in their training data, giving them a natural edge in Chinese general knowledge. On the other hand, during the Retrieval-Augmented-Generation (RAG) process, AI applications in China can access more effective and higher-quality Chinese retrieval documents, thereby enhancing their ability to provide superior answers.

Output Length and Deep Thinking Mode Influence Answer Quality Longer responses provide more comprehensive and multifaceted explanations, thereby achieving higher scores on advanced metrics. For instance, in deep thinking mode, applications generating longer responses significantly outperformed shorter-response counterparts on advanced metrics (except for Gemini). However, increased response length often introduces more factual inaccuracies

Applications	Overall	Length	Basic Metric			Advanced Metric				Dataset		
			Fact.	Comp.	Rel.	Rich.	Prac.	Insi.	Insp.	O-E	M-H	T-S
Deep Thinking & Internet Enable												
Doubao	84.62	1405.48	69.54	**93.18**	91.97	86.80	89.74	87.70	73.39	86.57	**77.59**	53.73
Yuanbao	83.82	1456.96	66.87	92.48	92.97	88.90	88.21	85.30	72.03	85.42	76.88	**67.49**
DeepSeek	**86.66**	1654.11	68.90	90.05	93.47	**90.11**	**92.04**	**89.58**	**82.49**	**88.68**	76.94	55.35
Kimi	65.15	1117.61	72.95	79.34	84.97	57.45	59.11	59.91	42.29	66.57	67.94	50.85
Perplexity	65.42	965.40	**74.49**	80.09	89.18	52.13	59.67	59.11	43.28	67.95	61.23	39.32
Openai	58.88	751.54	69.64	76.40	80.41	46.49	53.01	46.33	39.87	61.42	55.23	38.30
Gemini	80.12	1514.70	71.88	85.25	**96.63**	77.39	79.74	81.85	68.13	82.57	71.77	44.77
Only Internet Enable												
Doubao	65.86	774.00	**74.02**	81.88	93.33	48.48	62.66	57.67	42.97	66.41	64.05	52.02
Yuanbao	**80.19**	1995.66	70.94	**89.52**	91.00	**72.18**	81.35	**79.93**	**76.39**	81.33	**65.37**	**58.79**
DeepSeek	77.659	1400.73	69.33	86.22	92.23	66.68	**81.47**	74.25	73.39	**82.40**	65.18	45.79
Kimi	65.629	769.70	73.44	81.25	**93.78**	47.76	62.58	56.83	43.72	68.20	60.14	49.46

Table 2: Evaluation result with different modes and datasets. "Length" represents average response length, "O-E" represents open-ended QA dataset, "M-H" represents multi-hop QA dataset, and "T-S" represents time-sensitive QA dataset.

or hallucinations, resulting in lower factuality scores compared to concise answers. Deep thinking mode always produces longer responses through detailed reasoning or planing, which generally elevated advanced metrics at the cost of some factual errors. Consequently, it excels in open-ended analytical tasks but proves excessive for simple fact queries. AI applications may need a dynamic switching mechanism to optimizing output quality according to task requirements. See Appendix E for more details on the output length and deep thinking issues.

Multi-hop and time-sensitive problems highlight key limitations for AI applications These AI applications are notably weaker at tackling multi-hop and time-sensitive issues compared to open-ended problems. When tasks require multi-step retrieval and reasoning, current AI applications usually lose key information and perform poorly in the dimensions of Factuality, Completeness, and Richness, due to their inability to maintain coherent reasoning chains or gather sufficient high-value information. This reveals the technical bottlenecks in dealing with complex information chains and real-time data. See Appendix E for more details on multi-hop and time-sensitive issues.

Human Alignment

In order to verify OmniEval is aligned with human preferences, we conducted an extensive pairwise comparison study. More than 100 real users were recruited to evaluate 2 AI application pairs: DeepSeek vs Kimi and DeepSeek vs Perplexity. For each data instance, up to six evaluators were presented with two anonymized responses and asked to select the more "useful" one, or declare a tie, from a real user perspective. The final human preference label of each data instance (A Win, B Win or Tie) was determined by a majority vote of the evaluators.

For our analysis, we converted OmniEval's continuous scores into discrete pairwise outcomes to compare them against human judgments. Recognizing that minor score differences can be statistically insignificant, we introduced a tie-margin: a score difference of 4 points or less (on our 100-point scale) was classified as a 'Tie'. As this value not only reflects a typical margin of human cognitive annotation but also accounts for the inherent noise and uncertainty of automated evaluation framework. If the difference exceeded this threshold, the model with the higher score was declared the winner. We then measured the correlation using Kendall's tau-b coefficient. A detailed empirical justification for this 4-point threshold and the formal definition of the tau-b coefficient are provided in Appendix D.

Comparisons	τ-b	p-value
DeepSeek vs. Kimi	0.778	2.404e-250
DeepSeek vs. Perplexity	0.811	1.634e-272
Overall of 2 comparison pairs	0.851	$\approx$0.000

Table 3: Correlations with Human

As shown in Table 3, the analysis yields an overall Kendall's τ-b of 0.851, indicating high correlation between OmniEval rankings and human preferences at a significant level. This also confirms OmniEval's reliability as an automated evaluation framework and demonstrates its sensitivity in discerning meaningful quality differences between model responses.

Conclusion

In this paper, we introduce OmniBench, a comprehensive benchmark for open-ended, time-sensitive, and multi-hop questions. By analyzing 60,000 real-user queries, we define 16 knowledge domains and 5 user intents. Using three automated pipelines, we generate and regularly update the corresponding datasets. We also propose OmniEval, a multi-dimensional hybrid evaluation framework for holistic answer assessment. We hope OmniBench advances the evaluation of open-ended QA in real-user scenarios and guides AI applications toward providing better answers to real-world problems.

References

Gemini. 2025. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-03-25.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Lin, B. Y.; Deng, Y.; Chandu, K.; Brahman, F.; Ravichander, A.; Pyatkin, V.; Dziri, N.; Bras, R. L.; and Choi, Y. 2024. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. arXiv:2406.04770.

Liu, X.; Lei, X.; Wang, S.; Huang, Y.; Feng, Z.; Wen, B.; Cheng, J.; Ke, P.; Xu, Y.; Tam, W. L.; Zhang, X.; Sun, L.; Gu, X.; Wang, H.; Zhang, J.; Huang, M.; Dong, Y.; and Tang, J. 2024. AlignBench: Benchmarking Chinese Alignment of Large Language Models. arXiv:2311.18743.

Myrzakhan, A.; Bsharat, S. M.; and Shen, Z. 2024. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. arXiv:2406.07545.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-04-16.

Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; Zhang, C. B. C.; Shaaban, M.; Ling, J.; Shi, S.; Choi, M.; Agrawal, A.; Chopra, A.; Khoja, A.; Kim, R.; Ren, R.; Hausenloy, J.; Zhang, O.; Mazeika, M.; Dodonov, D.; Nguyen, T.; Lee, J.; Anderson, D.; Doroshenko, M.; Stokes, A. C.; Mahmood, M.; Pokutnyi, O.; Iskra, O.; Wang, J. P.; Levin, J.-C.; Kazakov, M.; Feng, F.; Feng, S. Y.; Zhao, H.; Yu, M.; Gangal, V.; Zou, C.; Wang, Z.; Popov, S.; Gerbicz, R.; Galgon, G.; Schmitt, J.; Yeadon, W.; Lee, Y.; Sauers, S.; Sanchez, A.; Giska, F.; Roth, M.; Riis, S.; Utpala, S.; Burns, N.; Goshu, G. M.; Naiya, M. M.; Agu, C.; Giboney, Z.; Cheatom, A.; Fournier-Facio, F.; Crowson, S.-J.; Finke, L.; Cheng, Z.; Zampese, J.; Hoerr, R. G.; Nandor, M.; Park, H.; Gehrunger, T.; Cai, J.; McCarty, B.; Garretson, A. C.; Taylor, E.; Sileo, D.; Ren, Q.; Qazi, U.; Li, L.; Nam, J.; Wydallis, J. B.; Arkhipov, P.; Shi, J. W. L.; Bacho, A.; Willcocks, C. G.; Cao, H.; Motwani, S.; de Oliveira Santos, E.; Veith, J.; Vendrow, E.; Cojoc, D.; Zenitani, K.; Robinson, J.; Tang, L.; Li, Y.; Vendrow, J.; Fraga, N. W.; Kuchkin, V.; Maksimov, A. P.; Marion, P.; Efremov, D.; Lynch, J.; Liang, K.; Mikov, A.; Gritsevskiy, A.; Guillod, J.; Demir, G.; Martinez, D.; Pageler, B.; Zhou, K.; Soori, S.; Press, O.; Tang, H.; Rissone, P.; Green, S. R.; Brüssel, L.; Twayana, M.; Dieuleveut, A.; Imperial, J. M.; Prabhu, A.; Yang, J.; Crispino, N.; Rao, A.; Zvonkine, D.; Loiseau, G.; Kalinin, M.; Lukas, M.; Manolescu, C.; Stambaugh, N.; Mishra, S.; Hogg, T.; Bosio, C.; Coppola, B. P.; Salazar, J.; Jin, J.; Sayous, R.; Ivanov, S.; Schwaller, P.; Senthilkuma, S.; Bran, A. M.; Algaba, A.; den Houte, K. V.; Sypt, L. V. D.; Verbeken, B.; Noever, D.; Kopylov, A.; Myklebust, B.; Li, B.; Schut, L.; Zheltonozhskii, E.; Yuan, Q.; Lim, D.; Stanley, R.; Yang, T.; Maar, J.; Wykowski, J.; Oller, M.; Sahu, A.; Ardito, C. G.; Hu, Y.; Kamdoum, A. G. K.; Jin, A.; Vilchis, T. G.; Zu, Y.; Lackner, M.; Koppel, J.; Sun, G.; Antonenko, D. S.; Chern, S.; Zhao, B.; Arsene, P.; Cavanagh, J. M.; Li, D.; Shen, J.; Crisostomi, D.; Zhang, W.; Dehghan, A.; Ivanov, S.; Perrella, D.; Kaparov, N.; Zang, A.; Sucholutsky, I.; Kharlamova, A.; Orel, D.; Poritski, V.; Ben-David, S.; Berger, Z.; Whitfill, P.; Foster, M.; Munro, D.; Ho, L.; Sivarajan, S.; Hava, D. B.; Kuchkin, A.; Holmes, D.; Rodriguez-Romero, A.; Sommerhage, F.; Zhang, A.; Moat, R.; Schneider, K.; Kazibwe, Z.; Clarke, D.; Kim, D. H.; Dias, F. M.; Fish, S.; Elser, V.; Kreiman, T.; Vilchis, V. E. G.; Klose, I.; Anantheswaran, U.; Zweiger, A.; Rawal, K.; Li, J.; Nguyen, J.; Daans, N.; Heidinger, H.; Radionov, M.; Rozhoň, V.; Ginis, V.; Stump, C.; Cohen, N.; Poświata, R.; Tkadlec, J.; Goldfarb, A.; Wang, C.; Padlewski, P.; Barzowski, S.; Montgomery, K.; Stendall, R.; Tucker-Foltz, J.; Stade, J.; Rogers, T. R.; Goertzen, T.; Grabb, D.; Shukla, A.; Givré, A.; Ambay, J. A.; Sen, A.; Aziz, M. F.; Inlow, M. H.; He, H.; Zhang, L.; Kaddar, Y.; Ängquist, I.; Chen, Y.; Wang, H. K.; Ramakrishnan, K.; Thornley, E.; Terpin, A.; Schoelkopf, H.; Zheng, E.; Carmi, A.; Brown, E. D. L.; Zhu, K.; Bartolo, M.; Wheeler, R.; Stehberger, M.; Bradshaw, P.; Heimonen, J.; Sridhar, K.; Akov, I.; Sandlin, J.; Makarychev, Y.; Tam, J.; Hoang, H.; Cunningham, D. M.; Goryachev, V.; Patramanis, D.; Krause, M.; Redenti, A.; Aldous, D.; Lai, J.; Coleman, S.; Xu, J.; Lee, S.; Magoulas, I.; Zhao, S.; Tang, N.; Cohen, M. K.; Paradise, O.; Kirchner, J. H.; Ovchynnikov, M.; Matos, J. O.; Shenoy, A.; Wang, M.; Nie, Y.; Sztyber-Betley, A.; Faraboschi, P.; Riblet, R.; Crozier, J.; Halasyamani, S.; Verma, S.; Joshi, P.; Meril, E.; Ma, Z.; Andréoletti, J.; Singhal, R.; Platnick, J.; Nevirkovets, V.; Basler, L.; Ivanov, A.; Khoury, S.; Gustafsson, N.; Piccardo, M.; Mostaghimi, H.; Chen, Q.; Singh, V.; Khánh, T. Q.; Rosu, P.; Szlyk, H.; Brown, Z.; Narayan, H.; Menezes, A.; Roberts, J.; Alley, W.; Sun, K.; Patel, A.; Lamparth, M.; Reuel, A.; Xin, L.; Xu, H.; Loader, J.; Martin, F.; Wang, Z.; Achilleos, A.; Preu, T.; Korbak, T.; Bosio, I.; Kazemi, F.; Chen, Z.; Bálint, B.; Lo, E. J. Y.; Wang, J.; Nunes, M. I. S.; Milbauer, J.; Bari, M. S.; Wang, Z.; Ansarinejad, B.; Sun, Y.; Durand, S.; Elgnainy, H.; Douville, G.; Tordera, D.; Balabanian, G.; Wolff, H.; Kvistad, L.; Milliron, H.; Sakor, A.; Eron, M.; O., A. F. D.; Shah, S.; Zhou, X.; Kamalov, F.; Abdoli, S.; Santens, T.; Barkan, S.; Tee, A.; Zhang, R.; Tomasiello, A.; Luca, G. B. D.; Looi, S.-Z.; Le, V.-K.; Kolt, N.; Pan, J.; Rodman, E.; Drori, J.; Fossum, C. J.; Muennighoff, N.; Jagota, M.; Pradeep, R.; Fan, H.; Eicher, J.; Chen, M.; Thaman, K.; Merrill, W.; Firsching, M.; Harris, C.; Ciobâcă, S.; Gross, J.; Pandey, R.; Gusev, I.; Jones, A.; Agnihotri, S.; Zhelnov, P.; Mofayezi, M.; Piperski, A.; Zhang, D. K.; Dobarskyi, K.; Leventov, R.; Soroko, I.; Duersch, J.; Taamazyan, V.; Ho, A.; Ma, W.; Held, W.; Xian, R.; Zebaze, A. R.; Mohamed, M.; Leser, J. N.; Yuan, M. X.; Yacar, L.; Lengler, J.; Olszewska, K.; Fratta, C. D.; Oliveira, E.; Jackson, J. W.; Zou, A.; Chidambaram, M.; Manik, T.; Haffenden, H.; Stander, D.; Dasouqi, A.; Shen, A.; Golshani, B.; Stap, D.; Kretov, E.; Uzhou, M.; Zhidkovskaya, A. B.; Winter, N.; Rodriguez, M. O.; Lauff, R.; Wehr, D.; Tang, C.; Hossain, Z.; Phillips, S.; Samuele, F.; Ekström, F.; Hammon, A.; Patel, O.; Farhidi, F.; Medley, G.; Mohammadzadeh, F.; Peñaflor, M.; Kassahun, H.; Friedrich, A.; Perez, R. H.; Pyda, D.; Sakal, T.; Dhamane, O.; Mirabadi, A. K.; Hallman, E.; Okutsu, K.; Battaglia, M.; Maghsoudimehrabani, M.; Amit, A.; Hulbert, D.; Pereira, R.; Weber, S.; Handoko; Peristyy, A.; Malina, S.; Mehkary, M.; Aly, R.; Reidegeld, F.; Dick, A.-K.; Friday, C.; Singh, M.; Shapourian, H.; Kim, W.; Costa, M.; Gurdogan, H.; Kumar, H.; Ceconello, C.; Zhuang, C.; Park, H.; Carroll, M.; Tawfeek, A. R.; Steinerberger, S.; Aggarwal, D.; Kirchhof, M.; Dai, L.; Kim, E.; Ferret, J.; Shah, J.; Wang, Y.; Yan, M.; Burdzy, K.; Zhang, L.; Franca, A.; Pham, D. T.; Loh, K. Y.; Robinson, J.; Jackson, A.; Giordano, P.; Petersen, P.; Cosma, A.; Colino, J.; White, C.; Votava, J.; Vinnikov, V.; Delaney, E.; Spelda, P.; Stritecky, V.; Shahid, S. M.; Mourrat, J.-C.; Vetoshkin, L.; Sponselee, K.; Bacho, R.; Yong, Z.-X.; de la Rosa, F.; Cho, N.; Li, X.; Malod, G.; Weller, O.; Albani, G.; Lang, L.; Laurendeau, J.; Kazakov, D.; Adesanya, F.; Portier, J.; Hollom, L.; Souza, V.; Zhou, Y. A.; Degorre, J.; Yalın, Y.; Obikoya, G. D.; Rai; Bigi, F.; Boscá, M. C.; Shumar, O.; Bacho, K.; Recchia, G.; Popescu, M.; Shulga, N.; Tanwie, N. M.; Lux, T. C. H.; Rank, B.; Ni, C.; Brooks, M.; Yakimchyk, A.; Huanxu; Liu; Cavalleri, S.; Häggström, O.; Verkama, E.; Newbould, J.; Gundlach, H.; Brito-Santana, L.; Amaro, B.; Vajipey, V.; Grover, R.; Wang, T.; Kratish, Y.; Li, W.-D.; Gopi, S.; Caciolai, A.; de Witt, C. S.; Hernández-Cámara, P.; Rodolà, E.; Robins, J.; Williamson, D.; Cheng, V.; Raynor, B.; Qi, H.; Segev, B.; Fan, J.; Martinson, S.; Wang, E. Y.; Hausknecht, K.; Brenner, M. P.; Mao, M.; Demian, C.; Kassani, P.; Zhang, X.; Avagian, D.; Scipio, E. J.; Ragoler, A.; Tan, J.; Sims, B.;

Plecnik, R.; Kirtland, A.; Bodur, O. F.; Shinde, D. P.; Labrador, Y. C. L.; Adoul, Z.; Zekry, M.; Karakoc, A.; Santos, T. C. B.; Shamseldeen, S.; Karim, L.; Liakhovitskaia, A.; Resman, N.; Farina, N.; Gonzalez, J. C.; Maayan, G.; Anderson, E.; Pena, R. D. O.; Kelley, E.; Mariji, H.; Pouriamanesh, R.; Wu, W.; Finocchio, R.; Alarab, I.; Cole, J.; Ferreira, D.; Johnson, B.; Safdari, M.; Dai, L.; Arthornthurasuk, S.; McAlister, I. C.; Moyano, A. J.; Pronin, A.; Fan, J.; Ramirez-Trinidad, A.; Malysheva, Y.; Pottmaier, D.; Taheri, O.; Stepanic, S.; Perry, S.; Askew, L.; Rodríguez, R. A. H.; Minissi, A. M. R.; Lorena, R.; Iyer, K.; Fasiludeen, A. A.; Clark, R.; Ducey, J.; Piza, M.; Somrak, M.; Vergo, E.; Qin, J.; Borbás, B.; Chu, E.; Lindsey, J.; Jallon, A.; McInnis, I. M. J.; Chen, E.; Semler, A.; Gloor, L.; Shah, T.; Carauleanu, M.; Lauer, P.; uc Huy, T.; Shahrtash, H.; Duc, E.; Lewark, L.; Brown, A.; Albanie, S.; Weber, B.; Vaz, W. S.; Clavier, P.; Fan, Y.; e Silva, G. P. R.; Long; Lian; Abramovitch, M.; Jiang, X.; Mendoza, S.; Islam, M.; Gonzalez, J.; Mavroudis, V.; Xu, J.; Kumar, P.; Goswami, L. P.; Bugas, D.; Heydari, N.; Jeanplong, F.; Jansen, T.; Pinto, A.; Apronti, A.; Galal, A.; Ze-An, N.; Singh, A.; Jiang, T.; of Arc Xavier, J.; Agarwal, K. P.; Berkani, M.; Zhang, G.; Du, Z.; de Oliveira Junior, B. A.; Malishev, D.; Remy, N.; Hartman, T. D.; Tarver, T.; Mensah, S.; Loume, G. A.; Morak, W.; Habibi, F.; Hoback, S.; Cai, W.; Gimenez, J.; Montecillo, R. G.; Łucki, J.; Campbell, R.; Sharma, A.; Meer, K.; Gul, S.; Gonzalez, D. E.; Alapont, X.; Hoover, A.; Chhablani, G.; Vargus, F.; Agarwal, A.; Jiang, Y.; Patil, D.; Outevsky, D.; Scaria, K. J.; Maheshwari, R.; Dendane, A.; Shukla, P.; Cartwright, A.; Bogdanov, S.; Mündler, N.; Möller, S.; Arnaboldi, L.; Thaman, K.; Siddiqi, M. R.; Saxena, P.; Gupta, H.; Fruhauff, T.; Sherman, G.; Vincze, M.; Usawasutsakorn, S.; Ler, D.; Radhakrishnan, A.; Enyekwe, I.; Salauddin, S. M.; Muzhen, J.; Maksapetyan, A.; Rossbach, V.; Harjadi, C.; Bahaloohoreh, M.; Sparrow, C.; Sidhu, J.; Ali, S.; Bian, S.; Lai, J.; Singer, E.; Uro, J. L.; Bateman, G.; Sayed, M.; Menshawy, A.; Duclosel, D.; Bezzi, D.; Jain, Y.; Aaron, A.; Tiryakioglu, M.; Siddh, S.; Krenek, K.; Shah, I. A.; Jin, J.; Creighton, S.; Peskoff, D.; EL-Wasif, Z.; V, R. P.; Richmond, M.; McGowan, J.; Patwardhan, T.; Sun, H.-Y.; Sun, T.; Zubić, N.; Sala, S.; Ebert, S.; Kaddour, J.; Schottdorf, M.; Wang, D.; Petruzella, G.; Meiburg, A.; Medved, T.; ElSheikh, A.; Hebbar, S. A.; Vaquero, L.; Yang, X.; Poulos, J.; Zouhar, V.; Bogdanik, S.; Zhang, M.; Sanz-Ros, J.; Anugraha, D.; Dai, Y.; Nhu, A. N.; Wang, X.; Demircali, A. A.; Jia, Z.; Zhou, Y.; Wu, J.; He, M.; Chandok, N.; Sinha, A.; Luo, G.; Le, L.; Noyé, M.; Perełkiewicz, M.; Pantidis, I.; Qi, T.; Purohit, S. S.; Parcalabescu, L.; Nguyen, T.-H.; Winata, G. I.; Ponti, E. M.; Li, H.; Dhole, K.; Park, J.; Abbondanza, D.; Wang, Y.; Nayak, A.; Caetano, D. M.; Wong, A. A. W. L.; del Rio-Chanona, M.; Kondor, D.; Francois, P.; Chalstrey, E.; Zsambok, J.; Hoyer, D.; Reddish, J.; Hauser, J.; Rodrigo-Ginés, F.-J.; Datta, S.; Shepherd, M.; Kamphuis, T.; Zhang, Q.; Kim, H.; Sun, R.; Yao, J.; Dernoncourt, F.; Krishna, S.; Rismanchian, S.; Pu, B.; Pinto, F.; Wang, Y.; Shridhar, K.; Overholt, K. J.; Briia, G.; Nguyen, H.; David; Bartomeu, S.; Pang, T. C.; Wecker, A.; Xiong, Y.; Li, F.; Huber, L. S.; Jaeger, J.; Maddalena, R. D.; Lù, X. H.; Zhang, Y.; Beger, C.; Kon, P. T. J.; Li, S.; Sanker, V.; Yin, M.; Liang, Y.; Zhang, X.; Agrawal, A.; Yifei, L. S.; Zhang, Z.; Cai, M.; Sonmez, Y.; Cozianu, C.; Li, C.; Slen, A.; Yu, S.; Park, H. K.; Sarti, G.; Briański, M.; Stolfo, A.; Nguyen, T. A.; Zhang, M.; Perlitz, Y.; Hernandez-Orallo, J.; Li, R.; Shabani, A.; Juefei-Xu, F.; Dhingra, S.; Zohar, O.; Nguyen, M. C.; Pondaven, A.; Yilmaz, A.; Zhao, X.; Jin, C.; Jiang, M.; Todoran, S.; Han, X.; Kreuer, J.; Rabern, B.; Plassart, A.; Maggetti, M.; Yap, L.; Geirhos, R.; Kean, J.; Wang, D.; Mollaei, S.; Sun, C.; Yin, Y.; Wang, S.; Li, R.; Chang, Y.; Wei, A.; Bizeul, A.; Wang, X.; Arrais, A. O.; Mukherjee, K.; Chamorro-Padial, J.; Liu, J.; Qu, X.; Guan, J.; Bouyamourn, A.; Wu, S.; Plomecka, M.; Chen, J.; Tang, M.; Deng, J.; Subramanian, S.; Xi, H.; Chen, H.; Zhang, W.; Ren, Y.; Tu, H.; Kim, S.; Chen, Y.; Marjanović, S. V.; Ha, J.; Luczyna, G.; Ma, J. J.; Shen, Z.; Song, D.; Zhang, C. E.; Wang, Z.; Gendron, G.; Xiao, Y.; Smucker, L.; Weng, E.; Lee, K. H.; Ye, Z.; Ermon, S.; Lopez-Miguel, I. D.; Knights, T.; Gitter, A.; Park, N.; Wei, B.; Chen, H.; Pai, K.; Elkhanany, A.; Lin, H.; Siedler, P. D.; Fang, J.; Mishra, R.; Zsolnai-Fehér, K.; Jiang, X.; Khan, S.; Yuan, J.; Jain, R. K.; Lin, X.; Peterson, M.; Wang, Z.; Malusare, A.; Tang, M.; Gupta, I.; Fosin, I.; Kang, T.; Dworakowska, B.; Matsumoto, K.; Zheng, G.; Sewuster, G.; Villanueva, J. P.; Rannev, I.; Chernyavsky, I.; Chen, J.; Banik, D.; Racz, B.; Dong, W.; Wang, J.; Bashmal, L.; Gonçalves, D. V.; Hu, W.; Bar, K.; Bohdal, O.; Patlan, A. S.; Dhuliawala, S.; Geirhos, C.; Wist, J.; Kansal, Y.; Chen, B.; Tire, K.; Yücel, A. T.; Christof, B.; Singla, V.; Song, Z.; Chen, S.; Ge, J.; Ponkshe, K.; Park, I.; Shi, T.; Ma, M. Q.; Mak, J.; Lai, S.; Moulin, A.; Cheng, Z.; Zhu, Z.; Zhang, Z.; Patil, V.; Jha, K.; Men, Q.; Wu, J.; Zhang, T.; Vieira, B. H.; Aji, A. F.; Chung, J.-W.; Mahfoud, M.; Hoang, H. T.; Sperzel, M.; Hao, W.; Meding, K.; Xu, S.; Kostakos, V.; Manini, D.; Liu, Y.; Toukmaji, C.; Paek, J.; Yu, E.; Demircali, A. E.; Sun, Z.; Dewerpe, I.; Qin, H.; Pflugfelder, R.; Bailey, J.; Morris, J.; Heilala, V.; Rosset, S.; Yu, Z.; Chen, P. E.; Yeo, W.; Jain, E.; Yang, R.; Chigurupati, S.; Chernyavsky, J.; Reddy, S. P.; Venugopalan, S.; Batra, H.; Park, C. F.; Tran, H.; Maximiano, G.; Zhang, G.; Liang, Y.; Shiyu, H.; Xu, R.; Pan, R.; Suresh, S.; Liu, Z.; Gulati, S.; Zhang, S.; Turchin, P.; Bartlett, C. W.; Scotese, C. R.; Cao, P. M.; Nattanmai, A.; McKellips, G.; Cheraku, A.; Suhail, A.; Luo, E.; Deng, M.; Luo, J.; Zhang, A.; Jindel, K.; Paek, J.; Halevy, K.; Baranov, A.; Liu, M.; Avadhanam, A.; Zhang, D.; Cheng, V.; Ma, B.; Fu, E.; Do, L.; Lass, J.; Yang, H.; Sunkari, S.; Bharath, V.; Ai, V.; Leung, J.; Agrawal, R.; Zhou, A.; Chen, K.; Kalpathi, T.; Xu, Z.; Wang, G.; Xiao, T.; Maung, E.; Lee, S.; Yang, R.; Yue, R.; Zhao, B.; Yoon, J.; Sun, S.; Singh, A.; Luo, E.; Peng, C.; Osbey, T.; Wang, T.; Echeazu, D.; Yang, H.; Wu, T.; Patel, S.; Kulkarni, V.; Sundarapandiyan, V.; Zhang, A.; Le, A.; Nasim, Z.; Yalam, S.; Kasamsetty, R.; Samal, S.; Yang, H.; Sun, D.; Shah, N.; Saha, A.; Zhang, A.; Nguyen, L.; Nagumalli, L.; Wang, K.; Zhou, A.; Wu, A.; Luo, J.; Telluri, A.; Yue, S.; Wang, A.; and Hendrycks, D. 2025. Humanity's Last Exam. arXiv:2501.14249.

Rein, D.; Li Hou, B.; Cooper Stickland, A.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. arXiv:1811.00937.