LLMEval-Fair: A Large-Scale Longitudinal Study on Robust and Fair Evaluation of Large Language Models

Ming Zhang[1*], **Yujiong Shen**[1*], **Jingyi Deng**[1*], **Yuhui Wang**[1*], **Huayu Sha**[1], **Kexin Tan**[1], **Qiyuan Peng**[1], **Yue Zhang**[1], **Junzhe Wang**[1], **Shichun Liu**[1], **Yueyuan Huang**[1], **Jingqi Tong**[1], **Changhao Jiang**[1], **Yilong Wu**[1], **Zhihao Zhang**[1], **Mingqi Wu**[1], **Mingxu Chai**[1], **Zhiheng Xi**[1], **Shihan Dou**[1], **Tao Gui**[1,2], **Qi Zhang**[1,2,3†], **Xuanjing Huang**[1,2,3]

[1]Institute of Trustworthy Embodied AI, Fudan University
[2]Shanghai Key Laboratory of Multimodal Embodied AI
[3]Shanghai Key Lab of Intelligent Information Processing
mingzhang23@m.fudan.edu.cn
qz@fudan.edu.cn

Abstract

Existing evaluation of Large Language Models (LLMs) on static benchmarks is vulnerable to data contamination and leaderboard overfitting, critical issues that obscure true model capabilities. To address this, we introduce LLMEval-Fair, a framework for dynamic evaluation of LLMs. LLMEval-Fair is built on a proprietary bank of 220k graduate-level questions, from which it dynamically samples unseen test sets for each evaluation run. Its automated pipeline ensures integrity via contamination-resistant data curation, a novel anti-cheating architecture, and a calibrated LLM-as-a-judge process achieving 90% agreement with human experts, complemented by a relative ranking system for fair comparison. A 30-month longitudinal study of nearly 60 leading models reveals a performance ceiling on knowledge memorization and exposes data contamination vulnerabilities undetectable by static benchmarks. The framework demonstrates exceptional robustness in ranking stability and consistency, providing strong empirical validation for the dynamic evaluation paradigm. LLMEval-Fair offers a robust and credible methodology for assessing the true capabilities of LLMs beyond leaderboard scores, promoting the development of more trustworthy evaluation standards.[1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) has led to a proliferation of benchmarks designed to assess their capabilities (Chang et al., 2023; Laskar et al., 2024; Liu et al., 2025). However, these benchmarks predominantly rely on a static evaluation paradigm where models are tested on fixed, publicly available datasets (Chen et al., 2025). This approach is fundamentally vulnerable to data contamination and test set overfitting, contributing to a growing "evaluation crisis" where benchmark scores may no longer reliably reflect a model's generalizable abilities (Banerjee et al., 2024; Deng et al., 2024a; Dekoninck et al., 2024). For example, GPT-4 achieved exact match rates of 52% and 57% when guessing the masked options in MMLU (Hendrycks et al., 2021a) test sets, far exceeding random chance (Deng et al., 2024b). Similarly, Qwen-1.8B has been shown to exactly replicate complete n-grams from both training and test splits of GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b), including 25 exact 5-gram matches in the MATH test set, indicating potential undetected data leakage and memorization (Xu et al., 2024b). Furthermore, static benchmark designs, dataset contamination, and biased evaluation protocols could create misleading perceptions of LLM capabilities, undermining the reliability of current performance assessments (Banerjee et al., 2024).

The crisis compels us to shift our focus from what capabilities to evaluate, such as knowledge and reasoning, to a more foundational question: how to evaluate in a manner that is robust, fair, and resistant to strategic manipulation. Based on our analysis of the current evaluation crisis, we identify three fundamental challenges for constructing a trustworthy evaluation framework. These challenges correspond to the core stages of any assessment: the data, the protocol, and the ranking system.

Challenge 1: How can we ensure the integrity of the evaluation data? The cornerstone of any

[*]Equal Contribution.
[†]Corresponding Authors.
[1]Our code and data are publicly available at https://github.com/llmeval/LLMEval-Fair.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21246–21273
July 2-7, 2026

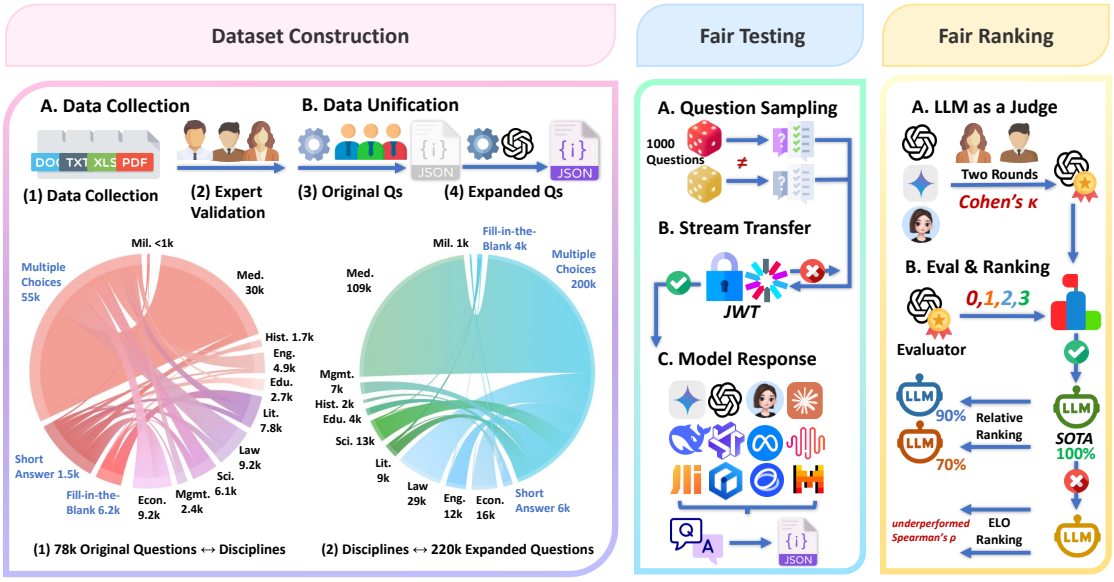

Figure 1: The LLMEval-Fair framework comprises three core stages. First, in Data Construction, diverse exam data are collected, filtered by experts, standardized into original questions in JSON format, and expanded to enhance formality and diversity. Second, Fair Testing involves sampling 1,000 unique questions for models' evaluation, transmitting data via a secure JWT stream, and collecting model responses. Third, Fair Ranking begins by selecting evaluators based on human-machine agreement, measured by Cohen's κ coefficients. Subsequently, the process entails generating relative rankings, validating robustness through ablation studies, and analyzing model errors. Discipline abbreviations: Eng. (Engineering), Econ. (Economics), Edu. (Education), Lit. (Literature), Mgmt. (Management), Sci. (Science), Hist. (History), Med. (Medicine), Mil. (Military).

benchmark is its test set. If the data are public or predictable, they become susceptible to leakage into model training corpora, leading to inflated scores that reflect memorization rather than true capability. Therefore, building a benchmark that is inherently resilient to data contamination is the primary challenge. To address this, we construct a private question bank of over 220k graduate-level questions, augmented with structural variations to mitigate memorization.

Challenge 2: How can we design an unpredictable assessment protocol? Even with private data, a static protocol with fixed test sets can be reverse-engineered or exploited over time. A truly robust evaluation requires a dynamic and unpredictable process that prevents strategic games. To achieve this, we implement a dynamic evaluation protocol where models are served unseen, randomly sampled questions in each session through a secure, two-layer anti-cheating architecture, ensuring that each evaluation is unique and unpredictable.

Challenge 3: How can we establish a fair and stable ranking system under dynamic conditions? When the evaluation content is not fixed, traditional absolute scoring becomes unreliable for cross-model comparison. A fair ranking system must remain stable and consistent, even when models are tested on different, albeit equivalent, sets of questions. To solve this, we develop a novel relative ranking system powered by a highly-calibrated LLM-as-a-Judge framework. This system ranks models by comparing their performance within each evaluation session rather than relying on absolute scores, ensuring fair rankings with negligible variance across different data samples.

Collectively, these solutions constitute LLMEval-Fair, a comprehensive framework for dynamic LLM evaluation. We leverage this framework to conduct an extensive longitudinal study on our official platform, spanning from the first half of 2023 to the second half of 2025. The study is longitudinal across three simultaneous dimensions: *temporal* (continuous tracking over 30 months rather than single snapshots), *data* (the question bank is incrementally expanded, so later

evaluation rounds sample from a substantively different pool than earlier ones), and *model* (successive generations within the same family, e.g., GPT-3.5 → GPT-4o → GPT-5, are compared under a unified protocol). Throughout this period, we have continuously evaluated nearly 60 leading proprietary and open-source models, typically assessing them within one week of their public release. Each model is pinned to a specific, fixed API snapshot (e.g., `gpt-4o-2024-11-20`) and undergoes at least three independent, randomly sampled evaluation runs to ensure stability. To date, this rigorous, ongoing effort has accumulated over 180k evaluation data points.

Leveraging this large-scale evaluation data, we systematically investigate three research questions corresponding to our core challenges and find that (a) all models converge to a performance ceiling around 90% persistent gaps in specialized domains like literature and medicine, (b) dynamic rankings diverge from static benchmarks, and static benchmarks suffer from severe data contamination, (c) our framework demonstrates exceptional ranking stability with negligible variance under multi-round resampling and varying sample sizes. More insightful findings are presented in Section 3.

Overall, our contributions are threefold:

1. We construct LLMEval-Fair, a large-scale anti-cheating evaluation platform featuring a proprietary 220k question bank, secure dynamic sampling protocols, and robust anti-manipulation mechanisms for trustworthy LLM assessment.

2. We conduct an extensive 30-month longitudinal evaluation campaign across nearly 60 leading models, accumulating over 180k evaluation data points through continuous anti-cheating assessments and comparative analysis with static benchmarks.

3. We conduct extensive empirical analysis across three research questions corresponding to our core challenges, revealing eight key findings about model performance ceilings, ranking stability, and contamination vulnerabilities in current evaluation practices.

2 Design

In this section, we detail the design and implementation of LLMEval-Fair, which addresses the three fundamental challenges through a three-stage framework. *Dataset Construction* tackles data integrity by building a contamination-resistant private question bank. *Evaluation Process* addresses unpredictable assessment through dynamic sampling and anti-cheating mechanisms. *Ranking System* implements a calibrated ranking system for fair model comparison.

2.1 Dataset Construction

To build a contamination-resistant and high-quality question bank, we sourced postgraduate and undergraduate exam questions from Chinese universities, covering 13 primary and over 50 secondary academic disciplines. Figure 1 offers a comprehensive overview of the design, highlighting three key stages: Dataset Construction, Fair Testing and Fair Ranking.

All questions and evaluation prompts are in Chinese, as the benchmark targets Chinese-language capabilities; translated examples are provided in the appendix for international readers. The construction follows a rigorous pipeline. First, we collect original exam questions from diverse formats and invite over 30 graduate-student annotators with domain expertise corresponding to the question disciplines for quality screening. Each item is independently reviewed by two annotators before finalization. This dual-review process yields 78,009 high-quality original questions after eliminating those with factual errors or irrelevant answers. Second, we employ an LLM-driven augmentation process to expand coverage and diversity. For instance, each Multiple-Choice question with n options is converted into n Fill-in-the-Blank variants, while Material Analysis questions are decomposed into multiple true/false questions based on key information. Finally, all augmented questions undergo format verification and metadata enrichment to ensure quality and traceability. A controlled comparison confirms that augmentation preserves evaluation validity: model rankings remain identical (Spearman $\rho = 1.0$) with no statistically significant score difference ($p = 0.736$, Cohen's $d = 0.16$); details are in Appendix J.

As of early 2025, this process has resulted in the **LLMEval-Fair dataset**, which comprises over 220k questions across six main categories. The full dataset covers all 13 primary disciplines; for evaluation, we select the 10 most data-sufficient disciplines to ensure statistically balanced cross-discipline comparisons, excluding Agronomy, Arts, and Philosophy due to insufficient question vol-

umes at the time of initial sampling.[2] To maintain evaluation freshness and prevent contamination, we continuously expand the question bank through the same manual collection and automated augmentation pipeline. Detailed statistical analysis of the dataset, including disciplinary breakdown and content diversity, is provided in Appendix A.

2.2 Evaluation Process

To ensure a reliable and fair evaluation, we design a dynamic process centered on a multi-layered anti-cheating architecture. This approach guarantees that each evaluation is unique, robust against manipulation, and accurately reflects a model's capabilities. The process is built upon two core strategies: dynamic question sampling and a secure delivery architecture.

2.2.1 Dynamic Question Sampling

To ensure unpredictability, each model evaluation is based on a unique set of 1,000 questions sampled from our private question bank via a three-stage stratified procedure. **Stage 1 (Discipline-level quota):** We allocate approximately 100 questions to each of the 10 evaluated disciplines, ensuring balanced coverage. **Stage 2 (Question-type allocation):** Within each discipline's quota, questions are allocated proportionally by type (e.g., multiple-choice, short-answer) according to the type distribution in the augmented dataset. **Stage 3 (Sub-discipline sampling):** Questions are randomly drawn from each secondary discipline in proportion to its share within the discipline–type stratum; both original and augmented questions are mixed in the sampling pool. After sampling, each question enters a three-state lifecycle (*allocated* $\rightarrow$ *pending* $\rightarrow$ *completed*) managed by the inner-layer process control (Section 2.2.2), preventing re-use within or across sessions. Models must answer questions in the pre-allocated order, preventing any "cherry-picking" strategies. This ensures that every evaluation is a distinct event reflecting the model's true generalization ability.

2.2.2 Secure Anti-Cheating Architecture

The evaluation process is protected by a secure, two-layer anti-cheating architecture.

- **The Outer Layer (Access Control):** This layer manages authentication and authorization. We use JSON Web Tokens (JWT)(Jones et al., 2015) to secure every API request, ensuring that only authenticated models can participate in an evaluation session. A strict Role-Based Access Control (RBAC) system prevents any cross-session or cross-user data access, isolating each evaluation.
- **The Inner Layer (Process Control):** This layer enforces the evaluation rules. A multi-level quota system tracks the number of questions allocated, pending, and completed, effectively preventing models from attempting to acquire more questions than permitted or resubmitting answers. As a final safeguard, our system automatically strips all answers and explanations from the data transmitted to the model, ensuring that only the question content is exposed and preventing answer leakage through data parsing.

2.3 Ranking System

To establish fair and stable rankings under dynamic evaluation conditions, we develop a calibrated ranking framework that combines LLM-as-a-Judge evaluation with relative scoring mechanisms. This approach ensures consistent and reliable model comparisons even when different question sets are used across evaluation sessions.

2.3.1 LLM-as-a-Judge Evaluation

To quantify answer quality, we establish a standardized scoring metric with an integer range of [0,3], ensuring consistent evaluation of response efficacy across diverse model architectures. For scoring implementation, we uniformly employ GPT-4o (OpenAI, 2023a) as our judge, which has demonstrated high human-machine agreement through rigorous validation (detailed in Section 3.4). This choice of a single, validated scoring model eliminates potential biases introduced by varying evaluative criteria.

The scoring focuses on both core correctness and explanation quality, with core correctness serving as the primary indicator for score determination. The specific evaluation prompt and criteria are provided in Appendix B.

2.3.2 Evaluation Metrics

To mitigate systematic bias introduced by random sampling questions, LLMEval-Fair employs both relative score and absolute scores as evaluation metrics.

[2] Philosophy questions were later collected in sufficient quantity but were not introduced mid-evaluation to maintain cross-model fairness.

Model	$R_{\text{SOTA}}^{\text{model}}$	S_{model}	Eng.	Econ.	Edu.	Law	Lit.	Mgmt.	Sci.	Hist.	Med.	Mil.
Open-source LLMs												
DeepSeek-R1	97.40	**91.23**	**9.47**	9.43	**9.27**	9.37	**8.83**	9.37	**9.03**	**9.53**	8.50	8.43
DeepSeek-V3	96.47	90.36	9.30	**9.57**	8.93	9.23	8.60	9.13	8.97	9.47	**8.83**	8.33
Qwen-3-235B	96.42	90.32	9.23	9.43	9.03	**9.50**	8.23	9.43	8.97	9.17	8.73	**8.60**
Qwen-3-32B	92.22	86.38	8.43	9.10	8.57	9.10	7.77	**9.47**	8.67	9.30	7.70	8.27
Closed-source LLMs												
Doubao-1.5-Thinking-Pro	**100.00**	**93.67**	**9.47**	**9.67**	**9.43**	**9.77**	**8.93**	9.53	**9.23**	**9.70**	**8.97**	**8.97**
Gemini-2.5-Pro	97.22	91.07	9.20	9.47	9.20	9.30	8.43	**9.63**	9.07	9.40	8.50	8.87
Gemini-2.5-Pro-Thinking	97.15	91.00	9.13	9.50	9.37	9.47	8.40	**9.63**	9.20	9.27	8.30	8.73
Doubao-1.5-Pro	95.68	89.62	8.83	9.03	9.13	9.43	8.57	9.27	8.83	9.10	8.60	8.83
Kimi-K2	94.27	88.30	9.23	9.17	8.80	9.00	8.40	9.17	8.77	9.13	8.53	8.10
GPT-5	93.84	87.90	8.83	9.37	8.90	8.87	8.10	9.10	8.90	9.03	8.50	8.30
Claude-Sonnet-4.5-Thinking	93.48	87.57	8.90	9.17	8.80	8.97	8.00	9.23	8.90	9.00	8.27	8.33
o1	93.36	87.45	8.90	9.30	8.67	8.77	7.73	9.27	8.90	8.97	8.17	8.77
Claude-Sonnet-4-Thinking	91.03	85.27	8.57	9.00	8.63	8.73	7.57	9.10	8.93	8.70	7.97	8.07
Claude-Sonnet-4	91.00	85.24	8.57	8.80	8.50	8.70	7.80	9.03	8.80	8.80	8.17	8.07
GPT-4o-search	89.40	83.74	8.27	8.77	8.43	8.67	7.77	8.80	8.20	8.73	8.27	7.83
GPT-4o	88.09	82.51	7.90	8.67	8.30	8.33	7.17	8.97	8.57	8.67	7.63	8.30
o3-mini	84.13	78.80	7.97	8.60	8.30	8.20	6.73	8.57	8.53	7.17	7.03	7.70

Table 1: Overall and Subject-Level Scores. $R_{\text{SOTA}}^{\text{model}}$ represents the relative score (0-100 scale) as defined in Equation (2), with Doubao-1.5-Thinking-Pro as the reference SOTA model. S_{model} represents the absolute score (0-100 scale) as defined in Equation (1). Subject-level scores use a 10-point scale.

The absolute score S_{model} represents a model's performance on $N = 1000$ questions, where each question receives a score s_i (with maximum score $s_{\max} = 3$), mapped to the [0,100] interval:

$$S_{\text{model}} = \frac{\sum_{i=1}^{N} s_i}{N \times s_{\max}} \times 100 \tag{1}$$

The relative score $R_{\text{SOTA}}^{\text{model}}$ is defined as the model's absolute score relative to the current SOTA model's absolute score on the same question set, mapped to the [0,100] interval:

$$R_{\text{SOTA}}^{\text{model}} = \frac{S_{\text{model}}}{S_{\text{SOTA}}} \times 100 \tag{2}$$

In our current evaluation, we use *Doubao-1.5-Thinking-Pro* as the reference SOTA model, as it achieved the highest absolute score (93.67) among all evaluated models within our unified protocol while exhibiting the highest multi-round stability (variance $= 0.00$ across three independent 1,000-question trials; see Appendix F). This is an operational designation: should a stronger model emerge, the anchor can be updated and all relative scores recalculated, as the formula is anchor-agnostic by design.

3 Experiment and Analysis

Based on the three core challenges identified in our introduction, we design the LLMEval-Fair evaluation framework. In this section, we systematically investigate three critical research questions that arise from these challenges:

Research Question I: What authentic capability distributions and longitudinal trends do LLMs exhibit under LLMEval-Fair?

Research Question II: How does LLMEval-Fair dynamic evaluation compare with static benchmarks regarding ranking accuracy and contamination issues?

Research Question III: How stable and reliable is LLMEval-Fair's relative ranking system under multi-round resampling and human-machine consistency validation?

Through comprehensive experiments designed to address these research questions, we aim to validate the effectiveness of our dynamic evaluation paradigm and provide empirical evidence for the superiority of contamination-resistant assessment frameworks.

3.1 Experimental Setup

3.1.1 Benchmarking LLMs on LLMEval-Fair

We tracked nearly 60 LLMs from June 2023 to December 2025, presenting full results in Appendix E and focusing here on 17 representative models (proprietary and open-source). Appendix C provides a detailed overview of these representative models. Each model is evaluated across three prompting paradigms (Zero-Shot, Few-Shot, Chain-of-Thought) and 10 academic disciplines. In addition, we sample incorrect responses from all evaluated

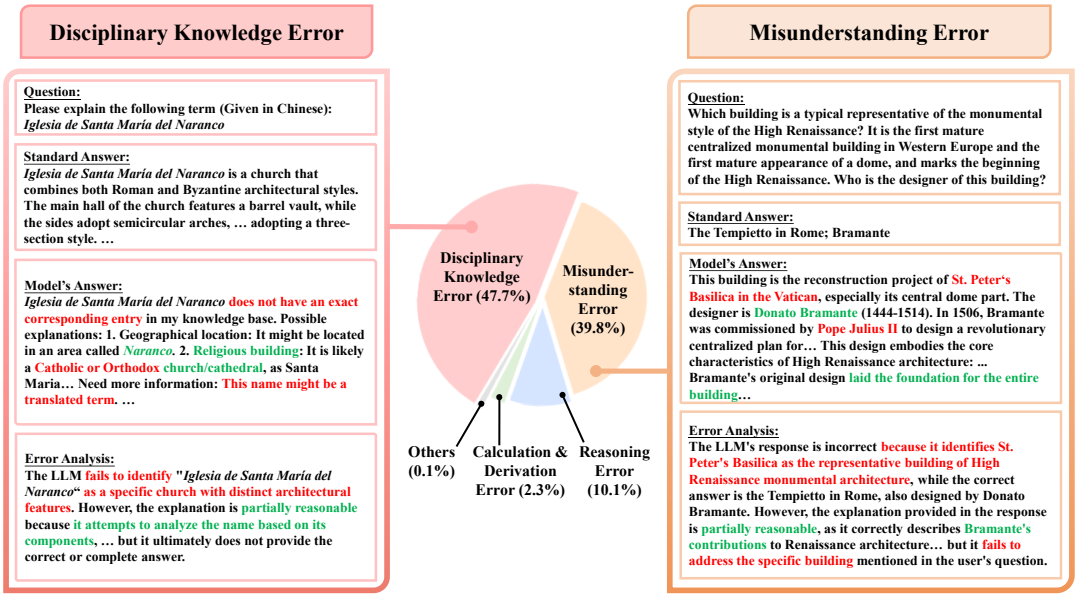

Figure 2: Distribution of model error causes and illustrative cases of the two most prevalent error types.

models and manually classify failures into five categories: disciplinary knowledge, misunderstanding, logical reasoning, factual inaccuracies, and format compliance.

3.1.2 Ablation Studies

We conduct comprehensive ablation studies across two key dimensions:

Benchmark Comparison: We measure Spearman correlation between LLMEval-Fair and static benchmarks (AGIEval (Zhong et al., 2024), C-Eval (Huang et al., 2023)) and perform fill-in-the-blank replay tests (1,000 questions, three attempts each) to assess contamination.

Ranking Validation: We conduct multi-round resampling (n=1000, 2000, 4000) to test stability. Human-machine agreement validation involves two independent rounds of human evaluation with Cohen's κ coefficients computed against three LLM-as-Judge evaluators. Second, we run an ablation study comparing our relative ranking to the traditional Elo scoring system. The introduction to the Elo scoring system can be found in Appendix D.

3.2 Research Question I

Finding 1: All models converge to a performance ceiling of around 90% over a longitudinal period, with leading open-source LLMs rivaling proprietary SOTA. Figure 3 illustrates the performance growth trajectories of different model series over time. Table 1 presents comprehensive evaluation scores and subject-level breakdowns for each model under the LLMEval-Fair framework. Nearly all models approach a performance ceiling around 90 on academic knowledge tasks. This convergence indicates fundamental limits in current model architectures for knowledge-intensive evaluation, consistent with theoretical analyses of knowledge retention upper bounds in pre-training (Jiang et al., 2025).

Model	Paradigm			Statistic	
	ZS	FS	CoT	Avg.	Var.
Claude-Sonnet-4-Thinking	85.27	85.60	86.87	85.91	0.48
DeepSeek-R1	91.23	89.33	88.43	89.67	1.36
Doubao-1.5-Thinking-Pro	93.67	90.63	91.73	92.01	1.58

Table 2: Capability under three prompting paradigms.

The top proprietary model (Doubao-1.5-Thinking-Pro, 93.67) and the top open-source model (DeepSeek-R1, 91.23) both substantially outperform established systems such as GPT-4o (82.51), demonstrating that open-source LLMs can rival proprietary SOTA. Performance within a model family is not always monotonically increasing (e.g., DeepSeek-V3.2 scores below DeepSeek-V3); we discuss these non-monotonic trends in Appendix I.

Finding 2: Models demonstrate significant domain-specific performance variations, with specialized "thinking" abilities offering only marginal gains. As shown in Table 1, all models excel in Management and Economics but consistently underperform in Literature, Medicine,

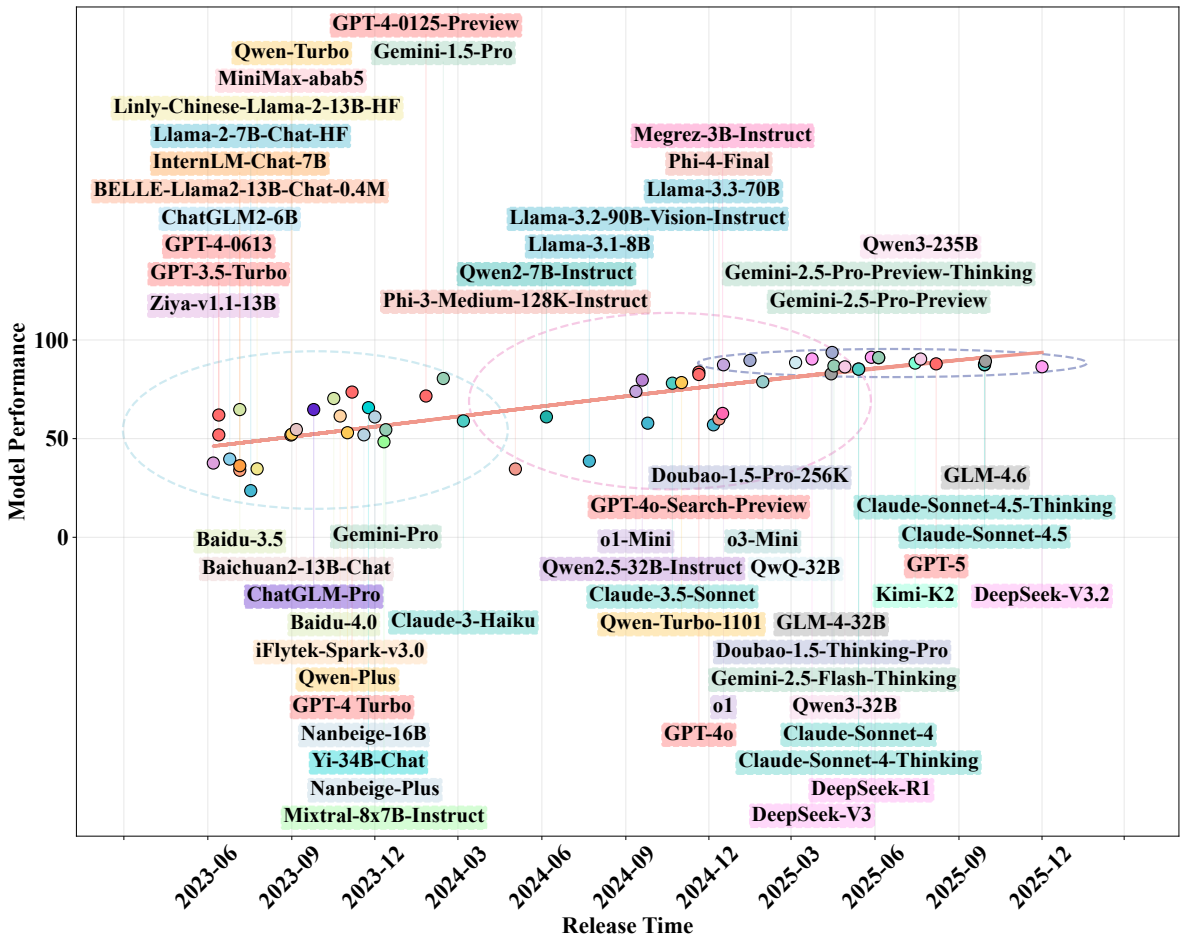

Figure 3: Trend of model series. Models of the same series are primarily illustrated in the same color for better distinction. The fitted curve highlights the overall growth of performance over the observed period.

and Military, revealing persistent domain-specific knowledge gaps. Dedicated "thinking" modes yield only modest gains (e.g., Claude-Sonnet-4-Thinking exceeds its base by ∼0.03 points), suggesting that domain coverage rather than reasoning mode remains the dominant performance driver.

Finding 3: In dynamic knowledge-intensive evaluations, prompting paradigms have minimal impact, whereas external augmentation may boost performance. As shown in Table 2, performance varies within three points across Zero-Shot, Few-Shot, and Chain-of-Thought prompting for all models (full results in Appendix E). By contrast, enabling web search boosts GPT-4o by 1.23 points (82.51 to 83.74), with the largest gains in Medicine and Literature, suggesting that external knowledge access is more impactful than prompt engineering for knowledge-intensive tasks.

Finding 4: Systematic error analysis reveals that disciplinary knowledge gaps and comprehension failures are the primary limitations of current models. Disciplinary knowledge gaps (47.7%) and misunderstanding errors (39.8%) together account for nearly 90% of all failures, indicating that knowledge coverage and contextual understanding are the primary bottlenecks. Figure 2 shows representative cases. The error taxonomy and coding methodology are detailed in Appendix K.

3.3 Research Question II

Finding 5: Dynamic rankings differ from static benchmarks, revealing contamination-driven distortions. To quantify the discrepancy between dynamic and static evaluation, we compare the model rankings from LLMEval-Fair with two representative static benchmarks, AGIEval (Zhong et al., 2024) and C-Eval (Huang et al., 2023). As shown in Table 3a, the rank correlations are moderate ($\rho \approx 0.65$–0.72), indicating that static benchmark rankings do not consistently align with those produced by our dynamic evaluation.

Benchmark	Spearman ρ	p-value
AGIEval (EN)	0.714	0.111
AGIEval (ZH)	0.657	0.156
C-Eval	0.657	0.156

(a) Rank correlation between LLMEval-Fair and static benchmarks.

Model	Ours	C-Eval		AGIEval	
		Rank	Δ	Rank	Δ
Gemini-2.5-Pro	1	2	+1	1	0
DeepSeek-V3	2	4	+2	3	+1
Doubao-1.5-Pro	3	1	−2	2	−1
Qwen3-32B	4	5	+1	5	+1
Claude-Sonnet-4	5	3	−2	4	−1
GPT-4o	6	6	0	6	0

(b) Per-model rank displacement ($\Delta = \text{Rank}_{\text{static}} - \text{Rank}_{\text{ours}}$). Negative Δ indicates over-ranking by the static benchmark.

Table 3: Comparison of LLMEval-Fair rankings with static benchmarks.

Model	AGI (EN)	AGI (ZH)	C-Eval	Ours
DeepSeek-V3	97	153	136	80
ClaudeSonnet-4	179	248	224	179
Doubao-1.5	66	105	117	76
o3-mini	58	74	45	75
GPT-4o	54	86	62	48
Qwen3-32B	55	77	72	36

Table 4: Comparison of successful fill-in completions for different models on static benchmarks and LLMEval-Fair.

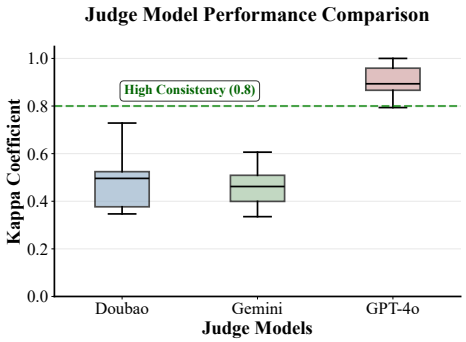

Figure 4: Cohen's κ coefficients measuring agreement between human evaluators and three LLM judges across evaluations. GPT-4o achieves almost perfect agreement with human judgments ($\kappa = 0.907$).

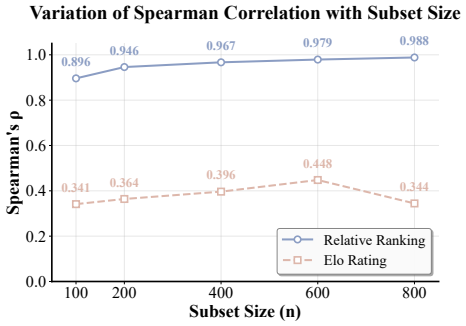

Figure 5: Relative ranking consistently outperforms Elo, reaching near-perfect correlation as the subset grows.

Table 3b further reports per-model rank displacement ($\Delta\text{rank} = \text{Rank}_{\text{static}} - \text{Rank}_{\text{ours}}$). The two models most over-ranked by C-Eval (Claude-Sonnet-4, $\Delta = -2$; Doubao-1.5-Pro, $\Delta = -2$) are precisely those with the highest fill-in-the-blank completion rates (Table 4), linking rank distortions to contamination. Overall, 26.7% of model pairs exhibit rank inversions against C-Eval versus only 13.3% against AGIEval.

Finding 6: Static Benchmarks Suffer from Severe Data Contamination. We conduct fill-in-the-blank replay tests (1,000 questions, three attempts each) on AGIEval and C-Eval. As shown in Table 4, public benchmarks yield substantially higher completion counts than our private dataset, confirming that static benchmarks suffer from significant leakage while our question bank remains contamination-resistant.

3.4 Research Question III

Finding 7: The relative ranking system demonstrates exceptional stability, with negligible variance across multi-round resampling and varying sample sizes. Multi-round resampling (n=1000, 2000, 4000) confirms that model ranking order remains identical across all runs, with relative scoring exhibiting negligible variance ($\sigma^2 \leq 1.68$). Full results are in Appendix F.

Finding 8: The relative ranking system via LLM-as-Judge achieves high human-machine agreement, is robust to judge selection, and outperforms alternative ranking systems. We validate our ranking system on three dimensions. First, Cohen's κ between LLM-as-a-Judge and human experts reaches 0.907 (Figure 4), indicating near-perfect agreement. Second, a cross-judge experiment with GPT-4o, Doubao-1.5-Thinking-Pro, and Gemini-2.5-Pro scoring the same 1,300 samples shows high ranking consistency (Spearman $\rho \geq 0.847$, all $p < 0.001$) and no family bias ($|\Delta| \leq 0.12$ on a 0–3 scale); full results are in Appendix L. Third, our relative ranking consistently outperforms an Elo-style baseline in Spearman correlation and stability across sample sizes (Figure 5).

4 Related Work

LLM Benchmarks. Early benchmarks evaluate LLMs on fixed question sets spanning factual knowledge and reasoning (Hendrycks et al., 2021a; Huang et al., 2023; Zhong et al., 2024), and the LLMEval series (Zhang et al., 2023, 2024, 2025b) extends this paradigm to Chinese academic domains. To go beyond static tests, human-preference platforms leverage pairwise comparisons to rank models on conversational quality (Lab, 2023; Zheng et al., 2024; Chiang et al., 2024; Zhang et al., 2025a), though they often lack depth in domain-specific reasoning and rely on non-expert annotators prone to stylistic bias (Raju et al., 2024). More recently, a growing body of work probes specific capabilities, including self-knowledge (Yin et al., 2023), hallucination detection (Sun et al., 2024), learning efficiency (Dou et al., 2025), instruction following (Ye et al., 2025b), tool learning (Ye et al., 2024a,b, 2025a,c, 2026), test-time scaling (Yin et al., 2025), adaptive evaluation (Ding et al., 2025), embodied interaction (Yang et al., 2025), context learning (Dou et al., 2026), taxonomy-guided research (Zhang et al., 2026), agentic coding (Ding et al., 2026), scientific workflows (Shen et al., 2026), and long-horizon active interaction (Xu et al., 2026). Despite this breadth, most benchmarks remain static and thus vulnerable to contamination.

Benchmark Contamination. The static nature of fixed benchmarks exposes them to data leakage, inflating scores and misrepresenting true capabilities (Banerjee et al., 2024; Xu et al., 2024a; Deng et al., 2024a), and encourages overfitting where models memorize answers rather than generalize (Deng et al., 2024b). These reliability concerns have prompted growing interest in dynamic evaluation protocols that mitigate contamination through private question banks and resampling strategies, which is the central motivation of our work.

LLM-as-a-Judge. Employing LLMs as scalable evaluators offers high throughput and strong human-preference alignment (Liu et al., 2023; Bai et al., 2024), yet systematic judge calibration and bias mitigation remain underexplored (Zheng et al., 2023). Our work addresses this gap through comprehensive cross-judge validation and human-machine agreement studies, demonstrating that a single well-calibrated judge can achieve near-perfect ranking consistency with human experts.

5 Conclusion

We introduced LLMEval-Fair, a dynamic, contamination-resistant evaluation framework built on a private 220k-question bank, a two-layer anti-cheating architecture, and an LLM-as-Judge relative ranking pipeline. A 30-month longitudinal study of nearly 60 open-source and proprietary models revealed a consistent performance ceiling near 90%, systematic gaps in literature, medicine, and military knowledge, and widespread data leakage in static benchmarks. Our relative ranking method demonstrated negligible variance under multi-round resampling with varying sample sizes and achieved near-perfect agreement with human experts. We further confirmed that prompting format has minimal impact on performance in knowledge-intensive tasks, underscoring the superiority of dynamic, contamination-resistant evaluation over static benchmarks and the need for more trustworthy benchmarking practices.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62521004, 62476061, 62576106, 62376061).

Limitations

The current benchmark is limited to Chinese-language academic questions, although the evaluation framework itself (dynamic sampling, anti-cheating architecture, relative ranking) is language-agnostic and applicable to any language given an appropriate question bank. Additionally, the sheer size of the question bank (over 220,000 items) makes comprehensive evaluation resource-intensive, requiring substantial compute, time, and human effort for large-scale inference, result validation, and ongoing dataset maintenance.

To facilitate reproducibility and community adoption, we have open-sourced the full 220k question bank, all evaluation results, a detailed data schema, recommended sampling strategies, and ethical usage guidelines at `https://github.com/llmeval/LLMEval-Fair`.

References

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael

Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *CoRR*, abs/2412.08905.

Alibaba Cloud. 2024. qwen-turbo-2024-11-01 (qwen-turbo-1101). https://www.alibabacloud.com/help/en/model-studio/models. Official model list. Accessed 2026-04-10.

Alibaba Cloud. 2026a. Qwen-plus. https://www.alibabacloud.com/help/en/model-studio/models. Official model list / documentation. Accessed 2026-04-10.

Alibaba Cloud. 2026b. Qwen-turbo. https://www.alibabacloud.com/help/en/model-studio/models. Official model list / documentation. Accessed 2026-04-10.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Anthropic. 2024b. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. Technical report, Anthropic.

Anthropic. 2025a. Claude sonnet 4.5 system card. System card, Anthropic.

Anthropic. 2025b. System card: Claude opus 4 & claude sonnet 4. Technical Report / System Card.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. 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)*, page 7421–7454. Association for Computational Linguistics.

Baidu. 2023a. Baidu announces second quarter 2023 results. Investor relations press release.

Baidu. 2023b. Baidu announces third quarter 2023 results. Investor relations press release.

Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The vulnerability of language model benchmarks: Do they accurately reflect true LLM performance? *CoRR*, abs/2412.03597.

BELLE Group. 2023. Belle-llama2-13b-chat-0.4m. https://huggingface.co/BELLE-2/BELLE-Llama2-13B-chat-0.4M. Official model card. Accessed 2026-04-10.

ByteDance. 2025. Bytedance ai introduces doubao-1.5-pro language model with a deep thinking mode. MarkTechPost article.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Simin Chen, Pranav Pusarla, and Baishakhi Ray. 2025. Dynamic benchmarking of reasoning capabilities in code large language models under data contamination. *CoRR*, abs/2503.04149.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Jasper Dekoninck, Mark Niklas Müller, Maximilian Baader, Marc Fischer, and Martin T. Vechev. 2024. Evading data contamination detection for language models is (too) easy. *CoRR*, abs/2402.02823.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024a. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8706–8719. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024b. Investigating data

contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Deming Ding, Shichun Liu, Enhui Yang, Jiahang Lin, Ziying Chen, Shihan Dou, Honglin Guo, Weiyu Cheng, Pengyu Zhao, Chengjun Xiao, Qunhong Zeng, Qi Zhang, Xuanjing Huang, Qidi Xu, and Tao Gui. 2026. Octobench: Benchmarking scaffold-aware instruction following in repository-grounded agentic coding. *CoRR*, abs/2601.10343.

Xuanwen Ding, Chengjun Pan, Zejun Li, Jiwen Zhang, Siyuan Wang, and Zhongyu Wei. 2025. Autojudger: An agent-driven framework for efficient benchmarking of mllms. *arXiv preprint arXiv:2505.21389*.

Shihan Dou, Ming Zhang, Chenhao Huang, Jiayi Chen, Feng Chen, Shichun Liu, Yan Liu, Chenxiao Liu, Cheng Zhong, Zongzhang Zhang, Tao Gui, Chao Xin, Wei Chengzhi, Lin Yan, Qi Zhang, Yonghui Wu, and Xuanjing Huang. 2025. Evalearn: Quantifying the learning capability and efficiency of llms via sequential problem solving. *CoRR*, abs/2506.02672.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, Huaibing Xie, Jianglu Hu, Shaolei Wang, Weichao Wang, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Pluto Zhou, and 8 others. 2026. Cl-bench: A benchmark for context learning. *CoRR*, abs/2602.03587.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in neural information processing systems*, 36:62991–63010.

iFLYTEK. 2023. iflytek highlights progress towards a new ai ecosystem at the 1024 global developer festival. Official news release.

InternLM Team. 2023. Internlm-chat-7b. `https://huggingface.co/internlm/internlm-chat-7b`. Official model card. Accessed 2026-04-10.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Changhao Jiang, Ming Zhang, Yifei Cao, Junjie Ye, Xiaoran Fan, Shihan Dou, Zhiheng Xi, Jiajun Sun, Yi Dong, Yujiong Shen, and 1 others. 2025. Beyond scaling: Measuring and predicting the upper bound of knowledge retention in language model pre-training. *arXiv preprint arXiv:2502.04066*.

Michael B. Jones, John Bradley, and Nat Sakimura. 2015. JSON Web Token (JWT). RFC 7519.

Tatsu Lab. 2023. Alpacaeval: An automatic evaluator for instruction-following language models. `https://github.com/tatsu-lab/alpaca_eval`. Accessed: 2025-07-31.

Md. Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee-Wei Tan, Md. Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 13785–13816. Association for Computational Linguistics.

Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuyu Yao, Tao Yuan, Dong Zhou, Yueqing Zhuang, Shengen Yan, Guohao Dai, and Yu Wang. 2025. Megrez-omni technical report. *CoRR*, abs/2502.15803.

Linly-AI. 2023. Chinese-llama-2-13b. `https://huggingface.co/Linly-AI/Chinese-LLaMA-2-13B-hf`. Official model card. Accessed 2026-04-10.

Songyang Liu, Chaozhuo Li, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S. Yu. 2025. The scales of justitia: A comprehensive survey on safety evaluation of llms. *CoRR*, abs/2506.11094.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval:

NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Meta. 2024a. Llama-3.1-8b. https://huggingface.co/meta-llama/Llama-3.1-8B. Official model card. Accessed 2026-04-10.

Meta. 2024b. Llama-3.2-90b-vision-instruct. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct. Official model card. Accessed 2026-04-10.

Meta. 2024c. Llama-3.3-70b-instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Official model card. Accessed 2026-04-10.

MiniMax. 2024. Tongyong da moxing abab6.5s xilie. Official release note.

Moonshot AI. 2025. Kimi k2 technical report. Technical report, Moonshot AI.

Nanbeige LLM Lab. 2023. Nanbeige-16b. https://github.com/Nanbeige/Nanbeige. Official repository for the Nanbeige-16B family. Accessed 2026-04-10.

Nanbeige LLM Lab. 2024. Nanbeige-plus-chat-v0.1. https://huggingface.co/Nanbeige/Nanbeige-Plus-Chat-v0.1. Official Hugging Face model page path indexed in public benchmark registry; verify accessibility before final submission. Accessed 2026-04-10.

OpenAI. 2023a. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023b. Introducing chatgpt and whisper apis. OpenAI blog.

OpenAI. 2024. o1 system card. Technical report, OpenAI.

OpenAI. 2025a. Gpt-5 system card. System card, OpenAI.

OpenAI. 2025b. o3-mini system card. Technical report, OpenAI.

Qwen Team. 2024. Qwq-32b. https://huggingface.co/Qwen/QwQ-32B. Official model card. Accessed 2026-04-10.

Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for llm-as-a-judge. *Preprint*, arXiv:2408.08808.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Yujiong Shen, Yajie Yang, Zhiheng Xi, Binze Hu, Huayu Sha, Jiazheng Zhang, Qiyuan Peng, Junlin Shang, Jixuan Huang, Yutao Fan, Jingqi Tong, Shihan Dou, Ming Zhang, Lei Bai, Zhenfei Yin, Tao Gui, Xingjun Ma, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. 2026. Sciagentgym: Benchmarking multi-step scientific tool-use in LLM agents. *CoRR*, abs/2602.12984.

YuHong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. 2024. Benchmarking hallucination in large language models based on unanswerable math word problem. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2178–2188, Torino, Italia. ELRA and ICCL.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024a. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Fangzhi Xu, Hang Yan, Qiushi Sun, Jinyang Wu, Zixian Huang, Muye Huang, Jingyang Gong, Zichen Ding, Kanzhi Cheng, Yian Wang, and 1 others. 2026. Odysseyarena: Benchmarking large language models for long-horizon, active and inductive interactions. *arXiv preprint arXiv:2602.05843*.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024b. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, and 36 others. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jie Yang, Jiajun Chen, Zhangyue Yin, Shuo Chen, Yuxin Wang, Yiran Guo, Yuan Li, Yining Zheng, Xuanjing Huang, and Xipeng Qiu. 2025. Vehicleworld: A highly integrated multi-device environment for intelligent vehicle interaction. *arXiv preprint arXiv:2509.06736*.

Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiecao Chen. 2025a. ToolHop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2995–3021, Vienna, Austria. Association for Computational Linguistics.

Junjie Ye, Caishuang Huang, Zhuohan Chen, Wenjie Fu, Chenyuan Yang, Leyi Yang, Yilong Wu, Peng Wang, Meng Zhou, Xiaolong Yang, and 1 others. 2025b. A multi-dimensional constraint framework for evaluating and improving instruction following in large language models. *arXiv preprint arXiv:2505.07591*.

Junjie Ye, Guanyu Li, SongYang Gao, Caishuang Huang, Yilong Wu, Sixian Li, Xiaoran Fan, Shihan Dou, Tao Ji, Qi Zhang, Tao Gui, and Xuanjing Huang. 2025c. ToolEyes: Fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 156–187, Abu Dhabi, UAE. Association for Computational Linguistics.

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. ToolSword: Unveiling safety issues of large language models in tool learning across three stages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2181–2211, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. RoTBench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 313–333, Miami, Florida, USA. Association for Computational Linguistics.

Junjie Ye, Guoqiang Zhang, Wenjie Fu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. Cctu: A benchmark for tool use under complex constraints. *arXiv preprint arXiv:2603.15309*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Zhiyuan Yu, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2025. Arise: An adaptive resolution-aware metric for test-time scaling evaluation in large reasoning models. *arXiv preprint arXiv:2510.06014*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, and 6 others. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Jiazheng Zhang, Wenqing Jing, Zizhuo Zhang, Zhiheng Xi, Shihan Dou, Rongxiang Weng, Jiahuan Li, Jingang Wang, Mingxu Chai, Shibo Hong, and 1 others. 2025a. Two minds better than one: Collaborative reward modeling for llm alignment. *arXiv preprint arXiv:2505.10597*.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025b. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4888–4914, Suzhou, China. Association for Computational Linguistics.

Ming Zhang, Yue Zhang, Shichun Liu, Haipeng Yuan, Junzhe Wang, Yurui Dong, Jingyi Deng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Llmeval-2.

Ming Zhang, Jiabao Zhuang, Wenqing Jing, Kexin Tan, Ziyu Kong, Jingyi Deng, Yujiong Shen, Yuhang Zhao, Ning Luo, Renzhe Zheng, Jiahui Lin, Mingqi Wu, Long Ma, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. Can deep research agents retrieve and organize? evaluating the synthesis gap with expert taxonomies. *CoRR*, abs/2601.12369.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Llmeval: A preliminary study on how to evaluate large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19615–19622.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhipu AI. 2025. Glm-4.6. Official documentation.

Zhipu AI and THUDM. 2026. Glm-4-32b-0414. `https://huggingface.co/zai-org/GLM-4-32B-0414`. Official model card. Accessed 2026-04-10.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2299–2314. Association for Computational Linguistics.

A Dataset

This section provides supplementary information on our LLMEval-Fair dataset employed in the study, to clarify the academic disciplines and question types covered, and to elaborate methodologies for scaling the questions.

A.1 Categories of Academic Displines

A.1.1 Categories of Academic Disciplines

As illustrated in Figure 6, we collected graduate-level examination questions spanning 13 primary (Philosophical Sciences, Economic Sciences, Law, Education, Literature, History, Engineering, Agronomy, Medicine, Military Science, Management Sciences, Arts, and Sciences) and more than 50 secondary academic disciplines recognized by China's Ministry of Education. Two-thirds of the questions are derived from Chinese universities' Postgraduate Entrance Exams, and one-third are from Undergraduate Final Exams of comparable difficulty. Detailed distribution of question sources is listed in Table 5.

A.1.2 Categories of Question Types

The raw dataset encompasses a diverse range of original question formats, including multiple-choices, fill-in-the-blank, true-or-false, short-answer, term explanation, and material analysis questions. Following question expansion and formatting, we have unified all questions and their answers into a fill-in-the-blank-like question-answer format, abandoning the complex answer structures of various question types, such as options A to E in multiple-choice questions, and "true" or "false" in true-or-false questions. This natural question-answer format enables the dataset to more thoroughly showcase the model's capabilities.

A.2 Details of Expanding the Dataset

As shown in Table 6, we have amassed a substantially large quantity of original questions, with a marked surge in numbers following the expansion of our dataset.

A.2.1 Original Data Construction Pipeline

Original questions, structured simply and organized by subject for initial collation, follow this construction pipeline: first, converting Excel, Word, and PDF test papers to TXT; then batch splitting into JSON-formatted questions via scripts; and finally conducting data screening.

Type	Number of Topics	Proportion (%)
Undergraduate Final Exams	26633	34.1
Postgraduate Entrance Exams	51376	65.9
Total	**78009**	100.0
Undergraduate Final Exams	71038	31.1
Postgraduate Entrance Exams	157566	68.9
Total	**228604**	100.0

Table 5: Distribution of question number and proportions for Undergraduate Final Exams and Postgraduate Entrance Exams.

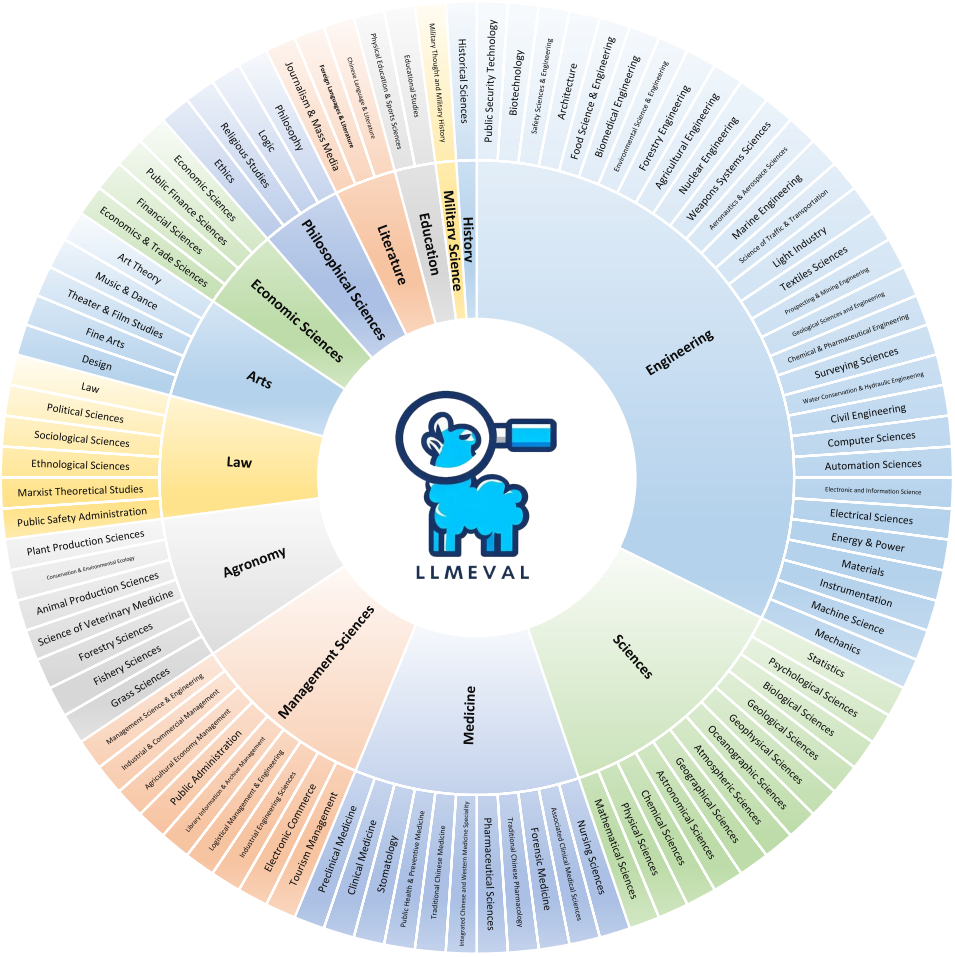

Figure 6: Categories of Primary and Secondary Academic Disciplines.

Subject	Original Count	Rewrited Count	Increase Count	Increase Percentage (%)
Philosophical Sciences	2194	10969	8775	399.95
Medicine	30772	109974	79202	257.38
Law	9262	30116	20854	225.16
Management Sciences	2448	7945	5497	224.55
Engineering	4926	13263	8337	169.24
Sciences	6182	15669	9487	153.46
Economic Sciences	9245	18124	8879	96.04
Military Science	611	1187	576	94.27
Education	2781	5094	2313	83.17
History	1749	3178	1429	81.70
Literature	7839	13085	5246	66.92
Total	**78009**	**228604**	**150595**	**193.05**

Table 6: Distribution of Original and Rewritten Counts Across Disciplines.

The latter involves three steps: (1) Expert review removes factually erroneous or irrelevant questions. (2) Batch splitting classifies and isolates questions, addressing errors such as content overlap, missing questions, or misclassifications. (3) Format cleaning resolves encoding conflicts, special characters, symbol consistency, redundancy, and typos.

A.2.2 Data Expanding Pipeline

To accommodate diverse application scenarios, this study proposes an augmented data format that complements the original question structure, The augmented dataset incorporates comprehensive metadata, including primary disciplinary categories, secondary disciplinary categories, question descriptions, answer content, and unique identifiers (UUIDs), which facilitates categorized data management. An example of expanding the Multiple-Choices question is shown in Figure 7.

Two additional key verification procedures are implemented upon the core augmentation strategy: (1) Format validation, which entails checking the consistency of option counts for multiple-choice questions and the alignment of answer spaces for Fill-in-the-Blank questions; and (2) Redundancy checks, which involve detecting duplicates among split questions and ensuring the uniqueness of question UUIDs. The data formats before and after expanding is illustrated in Figure 8.

B Prompts

This section presents the complete set of prompts used in LLMEval-Fair for different evaluation paradigms. We provide the specific prompt templates for few-shot learning, chain-of-thought reasoning, and LLM-based automated evaluation to ensure reproducibility of our experimental results.

The prompts used for testing few-shot and chain-of-thought methods are shown in Figures 9 and 10, respectively. The prompt used for LLM-based evaluation is shown in Figure 11.

Original Multiple-Choices questions:

Title: It is known that the sixth level of a complete binary tree (let the root be the first level) has eight leaves, then the number of nodes of the complete binary tree is at most ( ).
A. 39
B. 52
C. 111
D. 119

Answer: C

Expanded questions:

Question 1: It is known that the sixth level of a complete binary tree (let the root be the first level) has eight leaves, then the number of nodes of the complete binary tree is at most: ( )
Is it correct to place the answer “39" in the provided space?
Answer: False

Question 2: It is known that the sixth level of a complete binary tree (let the root be the first level) has eight leaves, then the number of nodes of the complete binary tree is at most: ( )
Is it correct to place the answer “52" in the provided space?
Answer: False

Question 3: It is known that the sixth level of a complete binary tree (let the root be the first level) has eight leaves, then the number of nodes of the complete binary tree is at most: ( )
Is it correct to place the answer “111" in the provided space?
Answer: True

Question 4: It is known that the sixth level of a complete binary tree (let the root be the first level) has eight leaves, then the number of nodes of the complete binary tree is at most: ( )
Is it correct to place the answer “119" in the provided space?
Answer: False

Figure 7: Example of how to expand a Multiple Choice question.

category: Engineering,
sub_category: Computer Science,
question_uuid: a4b92619-f0ac-xxxx-xxxx-9f11ce3b4704,
dialog:
role: user,
content: Judge the correctness of the following statements, answer true/false, and give your reasons. The continue statement in a loop breaks out of the current loop.
hint: false,

gpt4res: Answer: false. The reason is that the continue statement does not break out of the entire loop.It skips the rest of the current loop and immediately moves on to the next iteration of the loop instead of ending the current loop completely.
gpt4judge:
judge: Overall Rating: 3 The reason why you gave this rating: The LLM provided the correct answer (false) and gave a reasonable explanation. It accurately explained that the ‘continue‘ statement does not terminate the entire loop but instead skips the remaining part of the current iteration and proceeds to the next iteration of the loop. This explanation aligns with the correct behavior of the ‘continue‘ statement in programming.,
stars: 3,
reason: The LLM provided the correct answer (false) and gave a reasonable explanation. It accurately explained that the ‘continue‘ statement does not terminate the entire loop but instead skips the remaining part of the current iteration and proceeds to the next iteration of the loop. This explanation aligns with the correct behavior of the ‘continue‘ statement in programming.

Figure 8: Detailed entries of a single question after expanding.

Input:

Here are several examples:

Question:
Please determine the correctness of the following statement. Answer with true/false and provide a reason. If all elements below the diagonal in the adjacency matrix of a directed graph are zero, then the graph must have a topological ordering.
Answer:
true. Because if all elements below the diagonal in the adjacency matrix of a directed graph are zero, the graph is a Directed Acyclic Graph (DAG), so it must have a topological ordering.

Question:
Please explain the following term: “Double Hundred Policy".
Answer:
It refers to “let a hundred flowers bloom, let a hundred schools of thought contend." This was a policy on science and culture officially proposed by Mao Zedong in 1956 and confirmed by the Central Committee of the Communist Party of China. The policy was severely undermined after 1957, but was reestablished and implemented following the Third Plenary Session of the 11th Central Committee.

Question:
Please determine the correctness of the following statement. Answer with true/false and provide a reason. According to the convertibility theory, the scope of commercial banks’ assets expanded from short-term turnover loans to consumer loans.
Answer:
false. Convertibility Theory: also known as the asset conversion theory. This theory suggests that to maintain liquidity for withdrawals, commercial banks can invest part of their funds in transferable securities. Since these profitable assets can be sold at any time and converted into cash, loans are not necessarily limited to short-term and self-liquidating types. Clearly, this theory emerged in the context of developing financial instruments and markets. Significance: it expanded the scope of bank asset operations. Drawbacks: it does not guarantee that assets can be liquidated without capital loss (which requires high asset quality and stable market conditions); it is also constrained by central bank monetary policy (e.g., the risk of a rise in discount rates). This theory provides a theoretical basis for Chinese commercial banks to engage in securities business (investment operations). However, in China, commercial banks’ investment activities are restricted due to: 1. limited investment instruments and underdeveloped credit mechanisms; 2. management systems narrowing investment scopes, and separation of operations preventing commercial banks from investing in stocks; 3. the nature of state-owned commercial banks limits their willingness for autonomous investment.

Question:
Why can the results of animal experiments not be fully applied to clinical practice?
Answer:
Because there are differences between humans and animals not only in cellular morphology and metabolism, but also due to the highly developed human nervous system, which is associated with language and thought (the second signaling system). Although there are similarities, the essential differences mean that human diseases cannot all be replicated in animals. Even if they can be replicated, animal responses are simpler than human responses. Therefore, results from animal experiments cannot be mechanically and fully applied to clinical practice without analysis. Only by comparing, analyzing, and synthesizing animal experiment results with clinical data can they be used as references in clinical medicine and provide a basis for studying the causes, mechanisms, prevention, and treatment of clinical diseases.

The following is the question to be answered:
Question:`{question}`
Answer:

Figure 9: The Few-shot Prompt Template.

Input:

`{question}`
Please think step by step and provide the final answer.

Figure 10: The Chain-of-Thought Prompt Template.

Input:

Please evaluate the following response from the
LLM regarding a discipline-specific question based
on the following criteria. You must score it on a scale of 0, 1, 2 or 3
stars:

Overall Rating:
0 star indicates wrong answer with a wrong explanation
1 stars indicate wrong answer but a partially reasonable explanation
2 stars indicate a correct answer with a partially reasonable explanation
3 stars indicate an correct answer with a reasonable explanation

User: `{question}`

LLM: `{LLM response}`

The correct answer to user's question is: `{correct answer}`

You must provide your feedback in the following format:
"Overall Rating":numbers of its stars(int)
The reason why you gave this rating: <Your Reason>(str)'

Figure 11: The Prompt Template for LLM Judgement.

C Evaluation Model Selection

To ensure a comprehensive and rigorous assessment, we conducted a preliminary evaluation on a broad set of 59 large language models. From this extensive pool, we selected a representative subset of 17 models for the detailed analysis presented in the main text. The detailed release dates for these selected models are provided in Table 7.

The proprietary frontier models include the **OpenAI** lineup, featuring the widely deployed **GPT-4o** and **GPT-4o-search** (OpenAI, 2023a), alongside the next-generation flagship **GPT-5** (OpenAI, 2025a). Furthermore, we assess the **o1** series (OpenAI, 2024) and the efficient **o3-mini** (OpenAI, 2025b), which represent specialized reasoning models trained with large-scale reinforcement learning to solve complex scientific and mathematical problems.

We include models capable of extended reasoning from other major providers. **Anthropic** is represented by the standard **Claude-Sonnet-4** (Anthropic, 2025b), as well as the **Claude-Sonnet-4-Thinking** and the advanced **4.5-Thinking** variants (Anthropic, 2025a), which operate in a specialized mode to perform self-reflection before generating responses. Similarly, **Google**'s contribution consists of **Gemini-2.5-Pro** (Gemini Team, Google DeepMind, 2025) and its reasoning-enhanced counterpart, **Gemini-2.5-Pro-Thinking** (Gemini Team, Google DeepMind, 2025), which incorporates internal chain-of-thought processes.

Regarding open-weights architectures and diverse scaling strategies, the study includes the **DeepSeek** family: **DeepSeek-V3** (DeepSeek-AI et al., 2024), a strong Mixture-of-Experts (MoE) model, and **DeepSeek-R1** (DeepSeek-AI et al., 2025), a model specifically optimized for reasoning tasks through post-training. The **Qwen-3** series is also evaluated to represent distinct points on the parameter spectrum, featuring both the massive **235B** model and the compact **32B** variant (Team, 2025).

Finally, we evaluate other high-performing systems to ensure a broad representation of the current landscape. This includes **Moonshot**'s **Kimi-K2** (Moonshot AI, 2025) and the **Doubao-1.5** series. Notably, **Doubao-1.5-Thinking-Pro** (ByteDance, 2025) serves as a key reference in our study, having demonstrated state-of-the-art capabilities in preliminary screenings.

D The Elo Rating System

The Elo rating system is a widely recognized method for calculating relative skill levels in zero-sum games, originally developed for chess and recently popularized for evaluating Large Language Models (e.g., Chatbot Arena). This system derives ratings from the outcomes of pairwise comparisons, providing a probabilistic framework to predict the likelihood of one entity outperforming another. Given two entities A and B with current ratings R_A and R_B, the expected score E_A (representing the probability of A winning) is calculated using a logistic curve:

Table 7: The representative list of models selected from a total of 59 evaluated LLMs. Dates for unreleased models are estimated based on technical previews.

Model	Released Date
Doubao-1.5-Thinking-Pro (2025)	April 15, 2025
Doubao-1.5-Pro (2025)	January 15, 2025
Gemini-2.5-Pro-Thinking (2025)	June 5, 2025
Gemini-2.5-Pro (2025)	June 5, 2025
DeepSeek-R1 (2025)	May 28, 2025
DeepSeek-V3 (2024)	March 24, 2024
Qwen-3-235B (2025)	April 29, 2025
Qwen-3-32B (2025)	April 29, 2025
GPT-5 (2025a)	June 7, 2025
Claude-Sonnet-4.5-Thinking (2025a)	September 29, 2025
Claude-Sonnet-4-Thinking (2025b)	May 14, 2025
Claude-Sonnet-4 (2025b)	May 14, 2025
o1 (2024)	December 17, 2024
o3-mini (2025b)	January 29, 2025
GPT-4o-search (2023a)	November 20, 2024
GPT-4o (2023a)	November 20, 2024
Kimi-K2 (2025)	September 5, 2025

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{3}$$

Following a match, the ratings are updated based on the discrepancy between the actual outcome S_A (where 1 represents a win, 0 a loss, and 0.5 a tie) and the expected probability. The update rule is given by $R'_A = R_A + K(S_A - E_A)$, where K is a constant factor that determines the sensitivity of the rating adjustment.

E LLMEval-Fair Leaderboard

This section presents comprehensive evaluation results from our longitudinal study tracking nearly 60 LLMs from late 2023 to mid-2025. We provide complete performance rankings and analyze the consistency of model capabilities across different prompting paradigms.

We tracked nearly 60 LLMs from late 2023 to mid-2025. Here, we present the complete evaluation results of our model assessments. The comprehensive results, including scores for all models across 10 academic disciplines, are presented in Table 8.

The models we selected in main paper was evaluated across three prompting paradigms: Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT). As shown in Table 2, the performance variance across these paradigms remains below 1.6 points for all evaluated models, indicating that core capabilities are not significantly influenced by the prompting format.

F Details in Experiment

This section provides additional experimental details demonstrating the robustness and stability of our ranking system. We present validation experiments across different sample sizes and show the consistency of our relative scoring methodology.

F.1 Sampling Validation

To verify the stability of our ranking system, we conducted evaluations across multiple sample sizes of n=1000 (in three rounds), 2000, and 4000 questions. The results in Table 9 demonstrate remarkable consistency in ranking order across all sample sizes.

Our relative scoring methodology produces smaller variance compared with absolute scoring approaches. The variance analysis reveals that top tier models show exceptional stability. DeepSeek-V3 exhibits a variance of 0.51 and O3-mini exhibits a variance of 0.95 while GPT-4o exhibits the highest variance of 1.63. Even the maximum variance represents less than two percent fluctuation indicating robust measurement precision.

The three independent one-thousand-sample runs demonstrate high reproducibility with models maintaining consistent relative positions across all test conditions. These findings validate that our ranking methodology captures stable model capabilities rather than random fluctuations.

F.2 LLM-as-Judge Validation

We calculated Cohen's κ coefficients between human evaluations and three LLM judges Doubao Gemini and GPT-4o across two evaluation rounds for thirteen models. As shown in Table 10 GPT-4o demonstrates superior performance with κ values consistently above 0.90 with an average of 0.901 in Round 1 and 0.892 in Round 2 indicating almost perfect agreement.

The data reveal notable stability differences among judges. GPT-4o maintains consistently high agreement across both rounds with minimal variation while Doubao and Gemini exhibit more fluctuation between rounds. Specifically Doubao's performance ranges from 0.232 to 0.745 across different models and Gemini exhibits even greater instability with some models showing dramatic drops

between rounds—for example DeepSeek-V3 declines from 0.627 to 0.147.

In contrast Doubao and Gemini show lower overall agreement with average κ values of 0.493 and 0.446 for Round 1 and Round 2 for Doubao and 0.494 and 0.400 for Gemini. Based on GPT-4o's consistently high correlation with human evaluations and superior stability we selected it as our primary judge for reliable assessment.

Model	$R_{\text{SOTA}}^{\text{model}}$	S_{model}	Eng.	Econ.	Edu.	Law	Lit.	Mgmt.	Sci.	Hist.	Med.	Mil.
Open-source LLMs												
DeepSeek-R1 (2025)	**97.40**	**91.23**	**9.47**	9.43	**9.27**	9.37	**8.83**	9.37	**9.03**	**9.53**	8.50	8.43
DeepSeek-V3 (2024)	96.47	90.36	9.30	**9.57**	8.93	9.23	8.60	9.13	8.97	9.47	**8.83**	8.33
Qwen3-235B (2025)	96.42	90.32	9.23	9.43	9.03	**9.50**	8.23	9.43	8.97	9.17	8.73	**8.60**
QwQ-32B (2024)	94.51	88.53	8.30	9.46	9.23	9.33	7.83	9.46	8.65	9.27	8.57	8.43
DeepSeek-V3.2 (2025)	92.27	86.43	8.73	9.13	8.53	8.70	7.40	9.33	8.87	9.37	8.53	7.83
Qwen3-32B (2025)	92.22	86.38	8.43	9.10	8.57	9.10	7.77	**9.47**	8.67	9.30	7.70	8.27
GLM-4-32B (2026)	88.43	82.83	7.77	8.97	8.33	8.33	7.03	9.13	8.27	8.77	8.23	8.00
Qwen2.5-32B-Instruct (2024b)	85.06	79.68	7.70	8.57	8.33	8.33	6.70	8.50	8.17	7.70	7.60	8.08
Qwen-Turbo-1101 (2024)	83.72	78.42	7.97	8.37	8.03	8.23	6.40	8.50	8.10	7.50	7.27	8.05
Yi-34B-Chat (2024)	70.15	65.71	5.77	6.63	7.37	7.53	5.47	5.77	5.47	7.47	6.30	7.93
Megrez-3B-Instruct (2025)	67.01	62.77	5.80	6.77	6.80	7.13	5.40	6.87	5.70	6.53	5.70	6.07
Qwen2-7B-Instruct (2024a)	65.15	61.03	5.47	6.73	6.33	7.60	5.13	6.17	6.17	5.73	5.33	6.37
Nanbeige-Plus (2024)	65.10	60.98	5.78	5.57	6.77	7.37	5.37	5.93	5.45	6.30	5.67	6.77
Phi-4-Final (2024)	63.98	59.93	5.80	6.47	6.23	6.53	5.53	6.30	6.27	5.50	5.43	5.87
Llama-3.2-90B-Vision (2024b)	61.74	57.83	5.63	6.33	6.20	5.80	4.73	6.10	6.57	5.03	5.27	6.17
Llama-3.3-70B (2024c)	60.85	57.00	5.80	6.90	5.63	5.70	5.47	5.70	6.30	4.70	4.87	5.93
Baichuan2-13B-Chat (2023)	58.28	54.59	4.47	5.53	7.40	6.90	4.63	4.80	4.33	6.23	4.60	5.70
Qwen-plus (2026a)	56.58	53.00	4.40	5.10	6.53	6.53	5.00	4.77	4.87	5.17	5.13	5.50
Qwen-turbo (2026b)	55.76	52.23	4.10	6.07	6.63	6.43	4.43	4.53	4.97	5.27	4.37	5.43
Nanbeige-16B (2023)	55.45	51.94	4.37	5.30	6.50	6.30	3.97	4.70	4.07	5.90	4.73	6.10
Mixtral-8x7B-Instruct (2024)	51.69	48.42	4.27	5.47	6.47	6.40	3.13	4.50	5.07	3.57	4.37	5.17
ChatGLM-2-6B (2024)	42.31	39.63	2.33	3.77	5.97	6.13	2.83	3.83	2.60	3.80	4.00	4.37
Llama-3.1-8B (2024a)	41.25	38.64	3.87	4.20	4.27	4.17	3.50	3.83	4.30	3.17	3.20	4.13
Ziya-13B-v1.1 (2022)	40.18	37.64	2.77	3.97	5.17	5.33	2.80	3.77	2.53	3.70	3.03	4.57
InternLM-7B-Chat (2023)	38.71	36.26	2.63	3.67	4.87	5.57	3.17	3.33	2.33	4.03	3.13	3.53
Linly-LLaMA2-13B (2023)	37.03	34.69	2.20	3.77	4.50	5.00	2.43	3.33	2.53	3.90	2.50	4.53
Phi-3-Medium-128K (2024)	36.95	34.61	2.27	4.17	3.70	4.23	2.87	4.50	3.57	3.20	2.27	3.83
BELLE-Llama2-13B-Chat (2023)	36.25	33.96	2.57	3.07	4.93	4.73	2.83	3.80	2.43	3.33	2.40	3.87
Llama-2-7B-Chat (2023)	25.22	23.62	1.53	3.43	3.00	3.73	1.73	2.43	1.97	2.17	0.80	2.83
Closed-source LLMs												
Doubao-1.5-Thinking-Pro (2025)	**100.00**	**93.67**	**9.47**	**9.67**	**9.43**	**9.77**	**8.93**	9.53	**9.23**	**9.70**	**8.97**	**8.97**
Gemini-2.5-Pro (2025)	97.22	91.07	9.20	9.47	9.20	9.30	8.43	9.63	9.07	9.40	8.50	8.87
Gemini-2.5-Pro-Thinking (2025)	97.15	91.00	9.13	9.50	9.37	9.47	8.40	9.63	9.20	9.27	8.30	8.73
Doubao-1.5-Pro (2025)	95.68	89.62	8.83	9.03	9.13	9.43	8.57	9.27	8.83	9.10	8.60	8.83
GLM-4.6 (2025)	95.26	89.23	8.80	9.27	8.70	9.23	8.40	**9.63**	8.90	9.30	8.43	8.57
Kimi-K2 (2025)	94.27	88.30	9.23	9.17	8.80	9.00	8.40	9.17	8.77	9.13	8.53	8.10
GPT-5 (2025a)	93.84	87.90	8.83	9.37	8.90	8.87	8.10	9.10	8.90	9.03	8.50	8.30
Claude-Sonnet-4.5-Thinking (2025a)	93.48	87.57	8.90	9.17	8.80	8.97	8.00	9.23	8.90	9.00	8.27	8.33
o1 (2024)	93.36	87.45	8.90	9.30	8.67	8.77	7.73	9.27	8.90	8.97	8.17	8.77
Claude-Sonnet-4.5 (2025a)	93.31	87.40	8.80	8.97	8.93	8.73	8.37	9.10	8.97	8.93	8.13	8.47
Gemini-2.5-Flash-Thinking (2025)	92.74	86.87	8.67	9.27	8.70	9.00	7.80	8.93	8.90	9.00	8.03	8.57
Claude-Sonnet-4-Thinking (2025b)	91.03	85.27	8.57	9.00	8.63	8.73	7.57	9.10	8.93	8.70	7.97	8.07
Claude-Sonnet-4 (2025b)	91.00	85.24	8.57	8.80	8.50	8.70	7.80	9.03	8.80	8.80	8.17	8.07
GPT-4o-search (2023a)	89.40	83.74	8.27	8.77	8.43	8.67	7.77	8.80	8.20	8.73	8.27	7.83
GPT-4o (2023a)	88.09	82.51	7.90	8.67	8.30	8.33	7.17	8.97	8.57	8.67	7.63	8.30
Gemini-1.5-Pro (2024)	85.91	80.47	8.13	8.45	8.30	8.37	7.04	8.17	8.43	8.50	7.48	7.60
o3-mini (2025b)	84.13	78.80	7.97	8.60	8.30	8.20	6.73	8.57	8.53	7.17	7.03	7.70
Claude-3.5-Sonnet (2024b)	83.38	78.10	7.97	8.53	8.27	7.93	7.03	8.50	8.00	7.57	6.70	7.60
o1-mini (2024)	78.93	73.93	7.27	8.43	7.90	7.53	6.27	8.27	8.17	6.43	6.63	7.03
GPT-4-Turbo (2023a)	78.57	73.60	6.97	8.17	8.33	7.80	6.00	7.57	8.13	7.00	6.43	7.20
GPT-4-Preview (2023a)	76.44	71.60	6.90	7.40	8.03	7.30	6.00	7.47	7.63	6.87	6.33	7.67
Baidu-4.0 (2023b)	75.08	70.33	7.27	7.23	7.67	7.43	5.63	6.47	6.80	7.63	7.80	6.40
Baidu-3.5 (2023a)	69.10	64.73	6.20	6.70	7.80	6.83	5.20	5.50	6.00	7.23	6.57	6.70
ChatGLM-Pro (2024)	69.10	64.73	5.90	7.07	7.03	7.90	5.43	6.33	5.00	6.67	5.97	7.43
GPT-4-Legacy (2023a)	66.15	61.96	6.50	6.73	6.60	6.73	5.43	6.10	6.47	5.30	5.20	6.90
Spark-3.0 (2023)	65.62	61.47	5.77	6.50	7.27	7.30	5.70	5.90	5.03	6.50	5.23	6.27
Claude-3-Haiku (2024a)	62.93	58.95	5.80	6.60	6.97	6.63	4.83	5.93	6.33	4.80	5.23	5.83
Gemini-Pro (2023)	58.18	54.50	4.87	5.43	7.07	6.43	5.10	4.50	4.65	6.33	4.42	5.70
GPT-3.5-turbo (2023b)	55.42	51.91	4.97	5.37	6.40	6.47	4.43	4.67	5.43	4.20	4.37	5.60
MiniMax-ABAB5 (2024)	55.33	51.83	3.87	5.63	6.87	6.97	4.33	4.40	2.93	6.13	4.27	6.43

Table 8: Overall and Subject-Level Scores. $R_{\text{SOTA}}^{\text{model}}$ represents the relative score (0-100 scale) as defined in Equation (2), with Doubao-1.5-Thinking-Pro as the reference SOTA model. S_{model} represents the absolute score (0-100 scale) as defined in Equation (1). Subject-level scores use a 10-point scale.

G Implementation Details

This section provides detailed information about the annotation processes and evaluation procedures underlying our LLMEval-Fair platform. We describe the expert involvement in data curation, validation processes, and the associated costs to ensure transparency and reproducibility.

G.1 Data Annotation Process

A total of 38 experts were engaged in data annotation and cleaning processes, with an average of more than 3 relevant specialists assigned to each discipline. For the annotation of original data, to mitigate fatigue-induced errors, annotation tasks for each expert were distributed across a 30–60 day period.

The cumulative remuneration disbursed to experts involved in data annotation and cleaning amounted to $48,700. Ongoing investments are being allocated to further hire experts to expand the dataset.

G.2 Manual Evaluation Process

To validate our LLM-as-Judge approach, we conducted comprehensive human evaluation studies. A total of 18 experts participated in manual evaluation processes, with an average of about 2 relevant specialists assigned to each discipline. This expert-based validation ensures that our automated evaluation system maintains high agreement with human judgment standards.

The evaluation process involved multiple rounds of assessment across 13 representative models, with experts providing independent judgments that were subsequently compared against our LLM-based evaluation system using Cohen's κ coefficient.

G.3 Cost Analysis

The development and validation of LLMEval-Fair required substantial investment in both human expertise and computational resources. We spent more than $5,000 on using latest APIs of LLMs and deploying models for evaluation purposes. Additionally, $10,000 was allocated for hiring qualified volunteers to conduct manual evaluations, ensuring rigorous validation of our automated assessment framework.

H JWT Authentication Process

This section describes the detailed implementation of our JSON Web Token (JWT) authentication system, which forms the outer layer of our two-tier anti-cheating architecture. The JWT process ensures secure and authenticated access to our evaluation platform while preventing unauthorized access and session manipulation.

Our JWT implementation follows a standard three-phase protocol: token generation, transmission, and verification. Algorithm 1 outlines the complete JWT authentication workflow used in LLMEval-Fair.

I Non-Monotonic Performance Trends

Performance trajectories within a model family are not always monotonically increasing across versions. For example, DeepSeek-V3.2 scores lower than DeepSeek-V3 despite being released later. We identify three factors that explain such non-monotonic trends:

Training reward changes. DeepSeek-V3.2 removed format rewards and introduced generative reward models compared to V3, leading to outputs that favor comprehensive discussion over the precision required by exam-style questions. In short-answer questions, V3 produces concise, targeted responses while V3.2 tends to embed correct answers within verbose explanations, resulting in lower scores under our scoring rubric.

Subject-specific training data bias. The distribution of disciplinary data varies across model training corpora. Later versions may improve on some subjects while regressing on others due to shifts in training data composition.

Model specification trade-offs. Different versions within the same family (e.g., lightweight vs. full, reasoning-optimized vs. general-purpose) involve trade-offs in generation quality. Performance should be compared at equivalent scale and optimization objectives rather than purely by release chronology.

These observations underscore the value of domain-specific dynamic benchmarks for revealing capability trade-offs that aggregate leaderboard scores may obscure.

J Augmentation Validation

To verify that our question augmentation process preserves evaluation validity, we conduct a

	1000 Questions			Larger Samples		Statistics	
Model	**Trial 1**	**Trial 2**	**Trial 3**	**2000**	**4000**	**Mean**	**Variance**
Doubao-1.5-Thinking-Pro	100.0	100.0	100.0	100.0	100.0	100.0	0.00
DeepSeek-V3	96.48	96.87	98.10	98.02	97.53	97.40	0.51
Qwen-3-32B	92.21	92.58	93.93	94.15	93.45	93.26	0.71
GPT-4o	88.08	90.21	91.50	90.69	90.48	90.19	1.63
o3-mini	84.13	85.31	86.69	85.56	86.18	85.57	0.95

Table 9: Ranking stability across different sample sizes. All scores are relative scores with Doubao-1.5-Thinking-Pro as reference (100.0). The three 1000-question trials demonstrate high reproducibility, with low variance indicating robust measurement precision.

	Round 1			Round 2		
Model Name	**Doubao**	**Gemini**	**GPT-4o**	**Doubao**	**Gemini**	**GPT-4o**
Open-source LLMs						
DeepSeek-R1	0.527	0.662	0.960	0.451	0.251	0.977
DeepSeek-V3	0.326	0.627	0.949	0.366	0.147	0.800
Qwen-3-235B	0.486	0.468	0.818	0.232	0.496	0.931
Qwen-3-32B	0.560	0.390	0.886	0.414	0.271	0.872
Closed-source LLMs						
Claude-Sonnet-4	0.309	0.186	0.826	0.402	0.677	0.909
Claude-Sonnet-4-Thinking	0.451	0.399	0.830	0.443	0.373	0.684
Doubao-1.5-Pro	0.629	0.388	0.950	0.383	0.451	0.915
Doubao-1.5-Thinking-Pro	0.707	0.786	0.993	0.745	0.324	0.920
Gemini-2.5-Pro	0.451	0.539	0.831	0.376	0.682	0.858
Gemini-2.5-Pro-Thinking	0.408	0.547	0.843	0.344	0.185	0.859
GPT-4o	0.447	0.507	0.925	0.553	0.417	0.962
o1	0.599	0.535	0.933	0.571	0.555	0.957
o3-mini	0.517	0.397	0.975	0.531	0.381	0.963
Mean	**0.493**	**0.495**	**0.902**	**0.447**	**0.401**	**0.893**

Table 10: Cohen's κ correlation coefficient between human evaluation and three large language model evaluations across two rounds.

controlled comparison between original and augmented questions. We trace augmented questions back to their originals and construct paired subsets covering the same knowledge points. Five representative models spanning different capability tiers are evaluated on both subsets under identical conditions, with results reported in Table 11.

Model	**Original**	**Augmented**	Δ
Gemini-2.5-Pro	92.59	91.85	-0.74
DeepSeek-V3	92.48	91.04	-1.44
Claude-Sonnet-4	89.65	86.07	-3.59
GPT-4o	84.62	81.40	-3.22
o3-Mini	74.86	80.75	$+5.89$

Table 11: Performance comparison between original and augmented questions.

Statistical analysis confirms no significant difference: the paired t-test yields $p = 0.736$ with a 95% confidence interval of $[-5.37, +4.13]$, and Cohen's $d = 0.16$ (negligible effect). The model ranking under original questions is identical to that under augmented questions (Spearman $\rho = 1.0$). Four of five models score slightly lower on augmented questions, consistent with the removal of answer-choice scaffolding that slightly increases difficulty. The exception is o3-Mini ($+5.89$), which benefits from the fill-in-the-blank format that eliminates distractors. Top-tier models show the smallest sensitivity (Gemini: -0.74, DeepSeek: -1.44), suggesting they rely on genuine knowledge rather than format-specific cues.

K Error Categorization Methodology

Our five error categories were developed through a systematic qualitative coding process. First, we sampled error responses across multiple models

Algorithm 1: JWT Authentication Process in LLMEval-Fair

Server-side Token Generation:
Generate a unique user identity (`user_id`)
Generate current timestamp and expiration time (`exp`)
Construct payload ← {`user_id`, `timestamp`, `exp`, `session_id`, `permissions`}
Sign payload with server Secret using HMAC-SHA256 to generate JWT
return JWT to the authenticated user

Client-side Request:
Include JWT in Authorization header: `Bearer <token>`
Send request to evaluation endpoint

Server-side Verification:
Extract JWT from Authorization header
Verify JWT signature using server Secret
Parse payload and extract claims
if JWT signature is invalid **then**
    return HTTP 401 Unauthorized
end if
if current_time $>$ `exp` **then**
    return HTTP 401 Token Expired
end if
if `user_id` not found **or** permissions insufficient **then**
    return HTTP 403 Forbidden
end if
if session validation fails (e.g., concurrent sessions detected) **then**
    return HTTP 403 Session Invalid
end if
Allow evaluation operation to proceed
Log access attempt with `user_id`, `timestamp`, and `session_id`

and conducted open coding to identify recurring failure patterns (**pilot coding**). Through iterative refinement, we consolidated the patterns into five categories (**codebook development**): (1) disciplinary knowledge gaps, (2) misunderstanding, (3) logical reasoning errors, (4) factual inaccuracies, and (5) format compliance failures, each with an operational definition and decision rules. During **annotation**, each erroneous response was assigned exactly one primary label. When multiple failure modes co-occurred, a fixed priority ordering was applied (disciplinary knowledge $>$ misunderstanding $>$ logical reasoning $>$ factual inaccuracies $>$ format compliance) to ensure consistency.

- **Disciplinary knowledge gaps**: The model lacks domain-specific knowledge required to answer the question (e.g., missing a key medical term or legal principle).
- **Misunderstanding**: The model misinterprets the question's intent or misses decisive contextual cues (e.g., overlooking the keyword "first" in a question asking for the earliest example).
- **Logical reasoning errors**: The model possesses the relevant knowledge but applies incorrect reasoning chains or draws invalid conclusions.
- **Factual inaccuracies**: The model generates statements that contradict established facts, despite understanding the question correctly.
- **Format compliance failures**: The model provides a substantively correct answer but fails to conform to the required response format (e.g., providing a narrative when a list is expected).

L Cross-Judge Experiment

To validate that GPT-4o introduces no systematic bias as judge, three judges from different model families—GPT-4o (OpenAI), Doubao-1.5-Thinking-Pro (ByteDance), and Gemini-2.5-Pro (Google)—independently scored the same 1,300 samples (100 questions $\times$ 13 models) on the 0–3 scale. Table 12 reports inter-judge ranking consistency, Table 13 reports item-level agreement, and Table 14 reports the family bias analysis.

Judge Pair	ρ	p	τ	p
GPT-4o vs. Doubao	0.894	$<$0.001	0.763	$<$0.001
GPT-4o vs. Gemini	0.847	$<$0.001	0.693	0.001
Doubao vs. Gemini	0.900	$<$0.001	0.782	$<$0.001

Table 12: Inter-judge ranking consistency (Spearman ρ and Kendall τ).

M Evaluation Timeline

Each model entry corresponds to a specific, fixed API snapshot. Table 15 provides the OpenAI family timeline as an example; equivalent tables for other families are available in our released materials.

Judge Pair	Exact	±1 Tol.	Wt. κ
GPT-4o vs. Doubao	78.4%	96.9%	0.649
GPT-4o vs. Gemini	81.5%	94.6%	0.603
Doubao vs. Gemini	74.5%	94.5%	0.543

Table 13: Item-level agreement between judge pairs.

Judge	Own	Own Avg	3-Judge	Δ
GPT-4o	OpenAI	2.47	2.39	$+0.08$
Doubao	ByteDance	2.68	2.75	-0.07
Gemini	Google	2.79	2.67	$+0.12$

Table 14: Family bias test: each judge's average score for its own family vs. the three-judge mean. All differences are negligible ($|\Delta| \leq 0.12$ on a 0–3 scale).

Model	API Snapshot	S_{model}	Eval.
GPT-3.5-Turbo	gpt-3.5-turbo-0613	51.90	2023 H1
GPT-4	gpt-4-0613	61.97	2023 H1
GPT-4-Turbo	gpt-4-1106-preview	73.60	2023 H2
GPT-4-Preview	gpt-4-0125-preview	71.60	2024 H1
o1-Mini	o1-mini-2024-09-12	73.93	2024 H2
GPT-4o	gpt-4o-2024-11-20	82.50	2024 H2
GPT-4o-Search	gpt-4o-search-2024-11-20	83.73	2024 H2
o1	o1-2024-12-17	87.43	2024 H2
o3-Mini	o3-mini-2025-01-31	78.80	2025 H1
GPT-5	gpt-5-2025-08-07	87.90	2025 H1

Table 15: Evaluation timeline for the OpenAI model family. S_{model} is the absolute score. Eval. denotes the evaluation window.