

LPFQA: A Long-Tail Professional Forum-based Benchmark for LLM Evaluation

Liya Zhu^{1*}, Peizhuang Cong^{2*}, Jingzhe Ding^{1*}, Aowei Ji^{1*}, Wenya Wu¹,
Jiani Hou¹, Chunjie Wu¹, Xiang Gao¹, Jingkai Liu¹, Huan Zhou¹, Xuele Sun¹
Yang Yang¹, Jianpen Jiao¹, Liang Hu¹, Xinjie Chen¹, Jiashuo Liu¹,
Tong Yang^{2†}, Zaiyuan Wang^{1†}, Ge Zhang^{1†}, Wenhao Huang¹
¹ByteDance Seed ²Peking University
yangtong@pku.edu.cn, zaiyuanwang@humanlaya.com, gezhang@umich.edu

Abstract

Large Language Models (LLMs) perform strongly on existing reasoning and question-answering benchmarks, yet such evaluations often fail to capture their ability to handle low-exposure professional knowledge in realistic long-tail scenarios. We introduce LPFQA, a benchmark constructed from authentic discussions collected from specialized professional forums. Unlike standardized educational questions or easily searchable factual queries, forum discussions naturally involve environment-specific workflows, implicit constraints, specialized terminology, and context-dependent troubleshooting processes, resulting in problem formulations sparsely represented in large-scale training corpora. LPFQA contains 430 expert-verified tasks spanning 7 academic and industrial domains. Experiments across mainstream LLMs reveal substantial performance degradation on LPFQA, even for frontier models that perform strongly on existing benchmarks. We further observe notable ranking inconsistencies between LPFQA and prior evaluations, suggesting that long-tail professional question answering requires complementary capabilities insufficiently captured by existing benchmarks. In addition, external tool augmentation consistently introduces negative gains on LPFQA. Overall, LPFQA provides a realistic benchmark for evaluating LLMs under low-exposure professional settings. Our data is available at <https://huggingface.co/datasets/m-a-p/LPFQA>.

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress across a broad range of tasks, including question answering (Zhuang et al., 2023; Li et al., 2024c), logical reasoning (Havrilla et al., 2024; Wang et al., 2023), and code optimization (Nam et al., 2024; Gu, 2023), enabling their

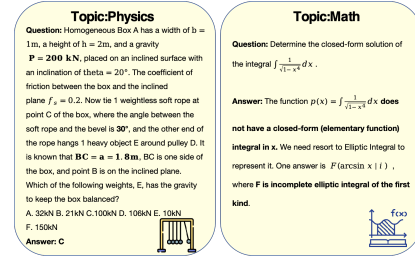


Figure 1: LPFQA task demos: multiple-choice and short-answer formats across professional domains.

widespread deployment in both professional and everyday applications (Yang et al., 2024; Zheng et al., 2025). As LLM capabilities continue to advance, reliable evaluation becomes increasingly important. Benchmark-based evaluation remains the dominant paradigm for assessing models’ ability to retrieve, reason over, and apply knowledge across diverse tasks (Chang et al., 2024).

A critical factor influencing LLM performance and the validity of their evaluation is the frequency with which knowledge appears during pre-training (Kandpal et al., 2023). Prior studies have shown that real-world knowledge distributions are highly long-tailed: while a small subset of knowledge patterns occurs frequently in large-scale corpora, the majority appear only sparsely (Li et al., 2024a; Godbole and Jia, 2023). Existing work further suggests that as domain specialization increases, knowledge tends to become increasingly context-dependent and underrepresented in mainstream web corpora and standardized educational resources (Zhang et al., 2023; Yang et al., 2022; Moslem et al., 2022; Touvron et al., 2023). Consequently, evaluating model behavior on such low-exposure knowledge is important for assessing robust generalization and reasoning beyond high-frequency training distributions (Naik et al., 2022).

However, existing benchmarks remain limited in evaluating long-tail professional capabilities. Most

*Equal contribution.

†Corresponding authors.

widely used evaluations are still dominated by high-frequency knowledge distributions (Hendrycks et al., 2021; Wang et al., 2024), while benchmarks targeting specialized domains often rely on synthetic generation, crowdsourced annotations, or academic-style problem construction (Mallen et al., 2023; Kumar et al., 2024; Huang et al., 2025; Li et al., 2024b; Phan et al., 2025), limiting the authenticity of real-world professional discourse. Moreover, many existing evaluations primarily emphasize retrieval-oriented or popularity-driven settings (Zhang et al., 2024; Sciavolino et al., 2021), while overlooking realistic professional workflows, evolving domain practices, and expert-grounded verification.

In particular, long-tail professional knowledge often emerges from realistic expert-level scenarios involving environment-specific workflows, specialized terminology, implicit contextual assumptions, and evolving domain practices. Although broad domain concepts may exist in pre-training corpora, their realistic professional formulations and problem-solving patterns appear substantially less frequently as specialization increases (Brookes and McEnery, 2019). Motivated by these observations, we focus on authentic discussions collected from **specialized professional forums**, where users seek assistance for practical real-world problems. Compared with synthetic academic-style questions or standardized educational resources, such discussions more naturally reflect highly specialized and low-exposure professional knowledge distributions encountered in real-world applications.

To systematically evaluate LLMs under realistic long-tail professional settings, we construct a Long-tail Professional Forum QA benchmark (LPFQA) grounded in authentic discussions collected from specialized professional forums. Unlike standardized educational questions or easily searchable factual queries, forum discussions typically involve environment-specific workflows, implicit operational constraints, specialized terminology, and context-dependent troubleshooting processes, resulting in professional problem formulations that appear only sparsely in large-scale training corpora. Comparison between LPFQA and existing benchmarks is shown at Table 1. To preserve such realism while ensuring evaluation reliability, we adopt a three-stage construction pipeline consisting of data collection and pre-processing, automated question generation with quality control, and expert verification with difficulty calibration. The

resulting benchmark contains 430 question items derived from multiple vetted public professional forums, spanning 7 primary domains and 21 secondary domains. Examples of LPFQA tasks are shown in Figure 1.

Based on systematic evaluations of a broad range of mainstream LLMs on LPFQA, we observe that, despite their strong performance on existing knowledge-oriented benchmarks, current models exhibit substantial performance degradation when confronted with long-tail professional questions. These failures are not isolated, but instead reveal systematic error patterns that expose a clear gap between general knowledge proficiency and the core capabilities required for reliable long-tail professional question answering, including domain-specific knowledge coverage and fine-grained question understanding.

The main contributions of our work are summarized as follows:

- We introduce LPFQA, a benchmark for long-tail professional question answering constructed from real-world discussions in highly specialized online forums, covering low-exposure domain knowledge sparsely represented in standard training corpora.
- We systematically evaluate mainstream LLMs on LPFQA and show that current models exhibit substantial performance degradation on long-tail professional queries despite strong performance on existing benchmarks.
- We provide a capability-oriented analysis of model failures, identifying recurring error patterns and key abilities required for long-tail professional question answering.

2 Related Work

2.1 Long-Tail Knowledge Benchmarks

LLM evaluation has rapidly evolved from early benchmarks (Wang et al., 2018, 2019), which focus on general language understanding, to more comprehensive benchmarks like MMLU (Wang et al., 2024), BIG-bench (Srivastava et al., 2023), and HELM (Liang et al., 2022), which extend evaluation to multi-disciplinary knowledge, reasoning, and holistic dimensions. Despite these advances, existing benchmarks remain limited in their ability to capture long-tail knowledge and authentic, user-centric information needs encountered in real-world scenarios, motivating the development of

Benchmark	Data Source	Construction	Knowledge Type	Real-World Context
BIG-Bench	Community tasks	Research-designed	Emergent reasoning	Limited
MMLU	Exams & textbooks	Academic curation	Canonical knowledge	No
HLE	Expert-authored problems	Expert curation	Formalized expertise	No
LPFQA(Ours)	Professional forums	Natural discussions	Long-tail practical knowledge	Yes

Table 1: Comparison between LPFQA and existing knowledge benchmarks.

new evaluation paradigms that go beyond curated or high-frequency knowledge distributions.

Given the significance of long-tail knowledge in assessing the authentic performance of LLMs, researchers have developed various specialized benchmarks through two primary construction paths. The first involves natural data collection from real-world distributions, such as the iNaturalist biodiversity dataset (Van Horn et al., 2018). The second method utilizes synthetic construction by creating imbalanced sampling from existing datasets, exemplified by ImageNet-LT (Liu et al., 2019). Despite these efforts, current long-tail knowledge benchmarks exhibit several pervasive deficiencies in practical applications. First, there is a notable lack of authenticity and naturalness in task design. Benchmarks like PopQA (Mallen et al., 2023) and Automatic QA Generation (Kumar et al., 2024), rely on templates or static knowledge graphs like LTGen (Huang et al., 2025). This leads to mechanical language patterns and fails to capture the dynamic, evolving nature of real-world information, a limitation that persists even in studies like RetrievalQA (Zhang et al., 2024) due to their narrow coverage. Second, the task complexity is frequently insufficient. Most datasets, including Entity-Questions (Sciavolino et al., 2021), focus primarily on single-hop factual recall while neglecting complex reasoning and cross-domain inference. The reliance on automated pipelines without systematic human verification may result in compromised data quality and factual inaccuracies.

2.2 User-Centric and Challenging Benchmarks

Beyond static long-tail knowledge evaluation, another line of work studies user-centric and dynamic assessment of LLM capabilities. Chatbot Arena (Chiang et al., 2024) evaluates models through crowdsourced blind comparisons in open-ended interactions, but offers limited control over task difficulty and domain expertise. ArenaHard (Li et al., 2024b) improves discriminative difficulty via adversarial questioning, though its distribution remains relatively concentrated. Hu-

manity’s Last Exam (HLE) (Phan et al., 2025) targets frontier-level reasoning through extremely difficult cross-disciplinary questions, while recent work such as DoLoMiTes (Malaviya et al., 2025) explores realistic domain-specific long-form procedural generation tasks. However, these benchmarks remain limited in systematically evaluating realistic long-tail professional question answering.

3 LPFQA Benchmark

LPFQA is constructed from authentic discussions collected from specialized professional forums, where users seek solutions to practical, environment-specific problems. Compared with standardized educational resources or curated benchmark questions, such discussions more naturally reflect low-exposure and context-dependent professional knowledge distributions. To evaluate LLMs under realistic professional settings, each discussion is transformed into either a multiple-choice or short-answer task paired with an expert-verified reference answer, ensuring both realism and evaluation reliability.

We adopt a three-stage construction pipeline shown in Figure 2, consisting of data collection and pre-processing, automated question generation and quality control, and expert verification with difficulty categorization.

3.1 Construction of LPFQA

3.1.1 Data Collection and Pre-processing

At first, we target collecting domain-specific discussions that reflect realistic long-tail QA scenarios. To this end, we recruit 209 human experts spanning 35 secondary disciplines, primarily affiliated with or educated at top-tier academic institutions, over 90% of whom either hold or are currently pursuing a master’s or doctoral degree, to curate professional online forums across specialized domains. For each candidate forum, experts provide its domain classification, access link, and representative discussion samples. To ensure domain relevance and data authenticity, all submissions undergo cross-review within the expert team. This process verifies that

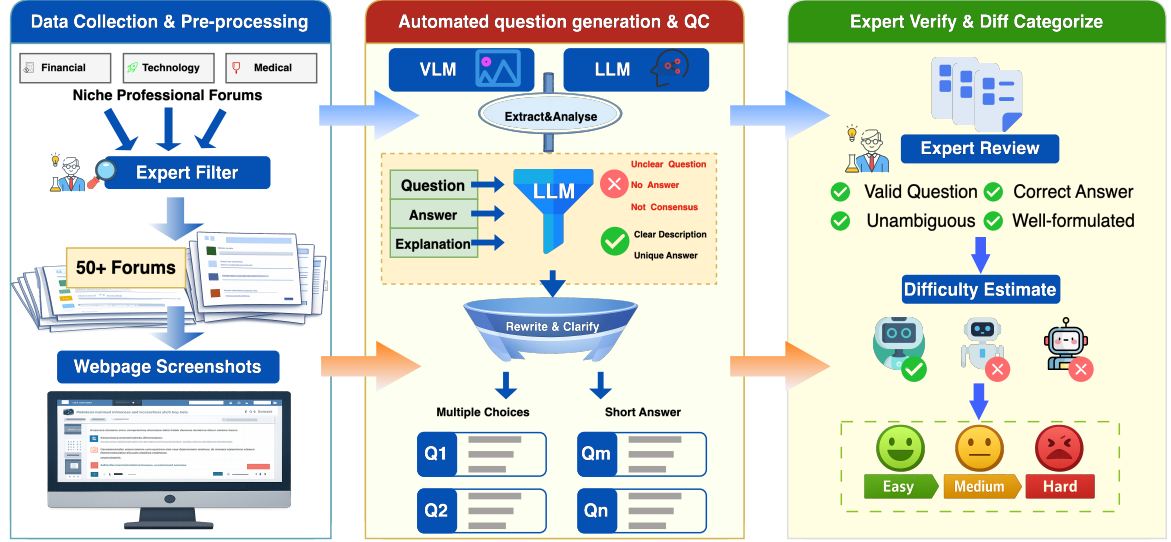


Figure 2: Construction pipeline of LPFQA.

the selected forums correspond to genuine professional fields and that the sampled questions represent realistic long-tail inquiries from real-world usage. Detailed expert qualification criteria and curation protocols are provided in Appendix E.

After filtering, we use automated tools to collect raw discussion data from over 50 vetted forums, yielding **more than 2,000** discussion threads from the past 7 months¹. Given the substantial heterogeneity of webpage structures across platforms, and to preserve contextual information, including layout, formatting, and question-answer relationships for following construction process, all discussions are stored as screenshots rather than plain text. The complete forum list is provided in Appendix D.

3.1.2 Automated Question Generation and Quality Control

Building upon the collected visual data, we further leverage models for preliminary filtering and structured annotation. Specifically, each screenshot is first processed by a vision-language model (VLM) to transfer its content into a structured format with predefined fields, including the question, candidate answer(s), and corresponding explanations. Based on extracted information, an LLM is then employed to assess whether the discussion contains a clearly defined question paired with a valid and unambiguous answer. Cases involving vague statements, missing answers, or controversial or incorrect responses are filtered out to ensure

answer reliability and consistency.

For the remaining tasks, we categorize their domains according to their nature and assign appropriate question formats based on task characteristics. Knowledge-oriented queries are predominantly transformed into multiple-choice questions, where the LLM constructs one correct option along with 4–6 plausible distractors grounded in the original discussion content. In contrast, reasoning- or computation-intensive problems are formulated as short-answer questions with concise, coherent, and well-justified reference answers. Regardless of question type, the model further refines and restructures problem descriptions to improve clarity, conciseness, and formal rigor for standardized evaluation, while preserving the original problem intent, reasoning structure, and answer semantics of the source discussion. To promote diversity in question formulation, we randomly select an LLM used for each task from a pool of multiple models during this stage. Importantly, the original semantics and reasoning intent of the source discussions are largely preserved: approximately 55% of retained tasks preserve the original forum question without rewriting, except for distractor addition in multiple-choice settings. Prompts to models used during this stage are provided in Appendix C. We retain **approximately 1,000** candidate tasks after this stage.

¹This work was conducted in July 2025; all collected threads were posted after January 2025.

Domain	Multiple choice	Short answer	Total
Natural Sci	73	24	97
Life Sci	57	14	71
Maths	30	31	61
CS	9	49	58
EE	43	19	62
Applied Tech	18	20	38
Social Sci	19	24	43
Overall	249	181	430

Table 2: Domain distribution of LPFQA benchmark.

Difficulty Level	Pass Rate Criterion	Nums
L1 (Easy)	> 50%	137
L2 (Medium)	30% – 50%	133
L3 (Hard)	< 30%	160

Table 3: Difficulty level distribution of LPFQA. The level of a task is calculated by the average pass rate of all models evaluated during the preliminary experiment.

3.1.3 Expert Verification and Difficulty Categorization

The remaining tasks are subsequently subjected to expert-driven refinement and difficulty calibration.

Expert manual verification: Domain specialists carefully review each item to verify that it satisfies the long-tail criterion, preserves the original forum semantics and reasoning intent after transformation, and that its question–answer pair is unambiguous, well-formulated, and factually correct. Any task failing to meet either requirement is discarded to ensure both realism and answer reliability.

Difficulty categorization by preliminary experiment: For the expert-approved tasks, we conduct difficulty stratification via a pilot evaluation protocol. We randomly select 6-8 SOTA LLMs to evaluate whether it could be solved correctly and determine its difficulty level by the proportion of models solving it. Following this procedure, tasks are categorized into three difficulty tiers (easy, medium, and hard) and overly trivial items are removed from the dataset. The number of different level tasks and corresponding criterion is shown in Table 3.

3.2 Statistics of LPFQA

We construct **LPFQA**, a benchmark comprising **430** professional questions, including 249 multiple-

choice items and 181 short-answer questions, spanning 7 primary domains. Each domain comprises authentic, highly specialized, and sparsely exposed long-tail questions collected from highly specialized professional forums and grounded in real-world scenarios, paired with expert-verified reference answers. The distribution of primary domains is summarized in Table 2. The detailed 21 secondary domains and typical examples of tasks from different domains can be found in Appendix A. As described in §3.1.3, every task is also annotated with a difficulty level. LPFQA is designed to rigorously evaluate LLMs not only in terms of their domain-specific factual knowledge in advanced and niche areas, but also their comprehensive analytical capabilities when addressing complex, professional-level queries. LPFQA contains tasks in two languages, with 239 questions in Chinese and 191 in English.

4 Experiments

Based on LPFQA, we evaluate a diverse set of 13 representative LLMs, covering both closed-source and open-source models. To improve reliability, we set $temperature = 1$ and $top_p = 1$ for all models in the experiment. The list of models is shown in Appendix F. For automatic evaluation, we use an LLM-based judge that compares the generated answer against the expert-verified reference answer for each task. The judge determines whether the two answers are semantically equivalent under a binary scoring scheme (1 for correct, 0 otherwise), and the final score is reported as average accuracy over all tasks. The LLM judge shows high agreement with human reviews. Prompt and human agreement statistics of the LLM judge are provided in Appendix C and Appendix G, respectively.

4.1 Main Results

The performance of different models across domains (Table 4) reveals several important insights.

Current models remain far from adequately solving real-world long-tail professional tasks. The average scores of all evaluated models fall within a narrow range of 35%–55%. Even the best-performing model, GPT-5, achieves only 54.42%, which still fails to reach the commonly accepted passing threshold of 60%. Given that all questions in LPFQA are collected from real user queries posted on online forums, this result highlights a critical gap between current LLM capabilities and the

Model	Life Sci	Maths	Natural Sci	Social Sci	EE	Applied Tech	CS	Overall
Seed-1.6-Thinking	41.67	52.46	46.88	41.86	42.37	57.89	37.93	45.35
GLM-4.6	54.17	55.74	45.83	48.84	64.41	39.47	43.10	50.47
Gemini-2.5-Pro	47.22	55.74	45.83	46.51	59.32	31.58	43.10	47.67
Gemini-3-Pro	45.83	47.54	55.21	48.84	57.63	57.89	48.28	51.16
Kimi-K2-Thinking	54.17	47.54	47.92	53.49	54.24	36.84	39.66	47.91
GPT-5-High	45.83	63.93	57.29	55.81	61.02	52.63	44.83	54.42
Claude-Sonnet-4.5-Thinking	56.94	49.18	50.00	44.19	57.63	42.11	43.10	49.53
GPT-4o	51.39	31.15	38.14	27.91	37.70	21.05	25.86	35.12
DeepSeek-V3	48.61	34.43	46.39	32.56	40.98	36.84	29.31	39.77
GPT-4.1	45.83	54.10	46.88	41.86	57.63	34.21	41.38	46.51
DeepSeek-V3.2-Thinking	54.17	57.38	48.96	46.51	50.85	47.37	41.38	49.53
Qwen-3-235B-Instruct	43.48	60.00	47.37	46.51	46.55	57.89	44.83	49.09
Qwen-3-235B-Thinking	40.58	48.33	52.63	48.84	51.72	36.84	46.55	46.85
All Models Avg	48.45	50.58	48.41	44.90	52.47	42.51	40.72	47.18

Table 4: Model performance on LPFQA across primary domains. The best score in each domain is in bold.

demands of authentic long-tail problem solving. In particular, it suggests that existing models struggle to reliably handle domain-specific, low-frequency questions that require precise understanding and accurate reasoning beyond memorized patterns. This further underscores the necessity of evaluating LLMs under realistic, user-driven long-tail settings, rather than relying solely on curated or distributionally simplified benchmarks.

Model performance varies substantially across domains, highlighting the strong domain dependence of real-world long-tail question answering. As shown in Table 4, average accuracy differs markedly across domains: Mathematics and Electrical Engineering achieve the highest scores (above 50%), while Computer Science and Applied Technology remain the most challenging, with accuracy dropping to roughly 40-43%. Other domains fall in between but still remain far from reliable. The large cross-domain variance suggests that LPFQA captures heterogeneous and realistic long-tail difficulty distributions rather than a uniformly difficult evaluation setting.

No existing model demonstrates robust and consistent dominance across all domains, highlighting the limited cross-domain generalization of current LLMs under long-tail settings. Although GPT-5-High achieves the highest overall score, its performance varies significantly across fields, with notable drops in Life Sciences and Computer Science. Similarly, Claude-Sonnet-4.5 excels in Life Sciences, while Gemini-3-Pro shows relative strength in Computer Science, yet neither maintains stable superiority across domains. This indicates that performance gains are often domain-

specific and suggests that scaling model capacity alone is insufficient to ensure reliable performance on diverse, forum-derived long-tail questions.

Stability across evaluation trials. To assess whether model performance on LPFQA is sensitive to inference-time stochasticity, we conduct six evaluation trials for representative models. The results, which is detailed in Appendix K, exhibit low variation across trials, with standard deviations ranging from 0.70 to 1.60 percentage points and 95% confidence intervals of approximately ± 0.73 – ± 1.68 points.

4.2 Analysis and Discussions

4.2.1 Error Attribution in LPFQA

To provide actionable insights for improving model performance on professional long-tail tasks, we conduct a fine-grained error attribution analysis. Specifically, for each primary domain, we randomly sample 50 instances incorrectly answered by a model, and systematically categorize the underlying causes of these errors. As illustrated in Figure 3, **knowledge deficiencies and question misinterpretation together account for over 73% of all observed errors**, indicating that current models still struggle to reliably internalize and correctly apply specialized domain knowledge, as well as to accurately comprehend technically phrased questions in long-tail professional scenarios.

In addition, reasoning-related errors constitute 16.3% of the total error cases, further suggesting that, in high-difficulty and highly specialized settings—particularly within science and engineering domains—solving problems requires not only factual knowledge but also non-trivial reasoning capa-

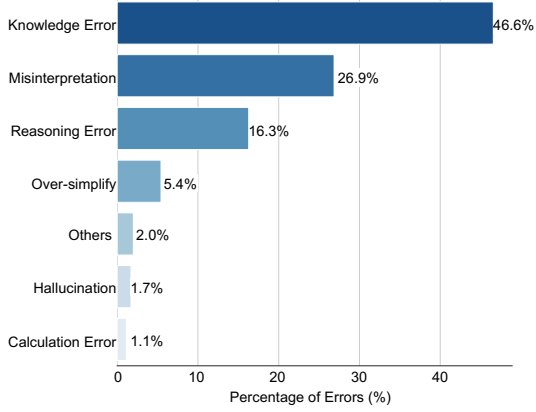


Figure 3: Error type distributions of LPFQA tasks.

bilities. This highlights the persistent gap between existing LLMs and the level of reasoning needed to solve complex real-world professional problems. The detailed distribution of error types for each primary domain can be found in Appendix H.

4.2.2 Long-tail Attribute Validation of LPFQA

To validate the long-tail nature of LPFQA and quantify its exposure under pretrained model distributions, we conduct two complementary analyses on LPFQA and standard academic benchmarks, including MMLU-devset (Hendrycks et al., 2021) and Math500 (Mayilvahanan et al., 2025).

- **Token-level likelihood analysis.** We compute the average token log-probability of benchmark samples under GPT-oss-20B, which measures how likely the benchmark text is under the pretrained distribution.
- **Reconstruction-based contamination analysis.** Following prior studies (Wu et al., 2026), we prompt models with partial input contexts and evaluate their ability to reconstruct the remaining continuation using ROUGE-L score (Lin, 2004). It is calculated as:

$$F_L = \frac{(1 + \beta^2)R_L P_L}{R_L + \beta^2 P_L}$$

Where R_L and P_L represent the recall rate and precision rate:

$$R_L = \frac{LCS(X, Y)}{m} \quad P_L = \frac{LCS(X, Y)}{n}$$

Where m and n denote the lengths of the question remainder sequence and the model-generated sequence, and $LCS(X, Y)$ denotes

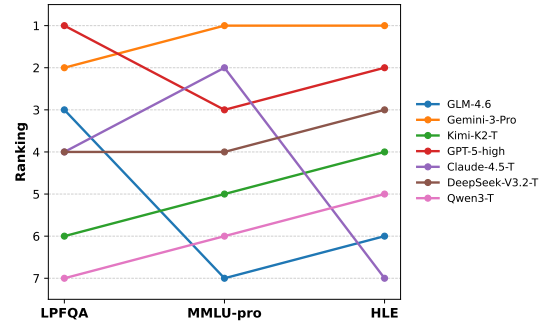


Figure 4: Model performance between LPFQA and other knowledge-oriented benchmarks.

Benchmark	Avg Log-Prob
LPFQA	-2.97
MMLU-devset	-2.63
Math500	-2.32

Table 5: Average token log-probability across benchmarks measured with GPT-oss-20B. Higher values indicate higher likelihood under the pretrained distribution.

the longest common subsequence between sequences. Lower score indicates reduced recoverability under pretrained distributions. We set $\beta = 1.2$ in experiment.

As shown in Table 5 and 6, LPFQA consistently exhibits substantially lower recoverability and lower token likelihood than standard benchmarks across both probes, indicating that its lexical-semantic patterns reside in a lower-density region of the pretrained distribution and are less likely to correspond to highly exposed or frequently repeated training content. These findings provide empirical evidence that LPFQA captures genuinely long-tail professional knowledge distributions rather than high-frequency academic-style benchmark patterns, supporting its suitability for evaluating LLMs in realistic low-exposure professional QA settings.

4.2.3 Relationship of Performance between LPFQA and Existing Benchmarks

To investigate the relationship between model performance on professional long-tail tasks and existing knowledge-oriented benchmarks, we compare both the absolute performance and relative rankings of representative models across LPFQA,

Benchmark	Qwen3-30B -Base	Qwen3-30B -Thinking	GPT-oss -20B
LPFQA	24.02	33.73	25.56
MMLU-devset	40.76	46.17	38.05
Math500	61.35	52.47	35.84

Table 6: ROUGE-L scores across benchmarks. We set the all the proportion of input problems to 40%. Models are deployed in pre-train mode without chat template.

MMLU-pro (Wang et al., 2024), and HLE (Phan et al., 2025), as shown in Figure 4. Top-models like GPT-5-high and Gemini-3-Pro, consistently remain within the top tier across all three benchmarks, suggesting a partial alignment between LPFQA and conventional knowledge-focused evaluations. This indicates that general factual competence and broad-domain reasoning ability still play an important role in professional long-tail tasks.

However, the figure also reveals substantial ranking divergences between LPFQA and other knowledge benchmarks. For example, Gemini-3-Pro, which ranks first on both MMLU-pro and HLE, is surpassed by GPT-5-high on LPFQA. Conversely, Claude-Sonnet-4.5 demonstrates competitive performance on MMLU-pro but exhibit notable degradation on HLE and LPFQA. These discrepancies suggest that excellence on knowledge-intensive benchmarks does not necessarily translate to strong performance in professional long-tail scenarios. LPFQA captures a complementary ability dimension that is seldom reflected by existing knowledge and exploration-oriented benchmarks, further motivating the necessity of evaluating models **under realistic, domain-specific long-tail settings** beyond conventional knowledge-centric benchmarks. The detailed scores of models on MMLU-pro and HLE can be found in Appendix I.

4.3 Ablations: Impact of Tool Augmentation

To investigate whether external tool augmentation can influence model performance on LPFQA, we select several models and conduct a set of ablation experiments by equipping models with additional tools during inference. Specifically, we apply two commonly used categories of tools: (i) Jupyter Code Interpreter (CI) that enables executable reasoning through code execution, and (ii) Search tools, including Google Search and Text Browser, which allow models to access external web content. As shown in Table 7, we observe a clear and consistent performance degradation across all mod-

els when equipped with external tools, regardless of the specific tool type, compared to their corresponding tool-free variants.

Through detailed case analysis, we observe that CI-equipped models tend to over-formalize LPFQA questions as programming tasks and rely excessively on explicit code execution. However, LPFQA mainly requires implicitly grounded, knowledge-driven reasoning that is difficult to formalize programmatically, causing code-based reasoning to become distracting rather than beneficial. Search augmentation shows a similar issue. Due to LPFQA’s long-tail nature, reliable external evidence is often unavailable, while retrieved content is frequently noisy, superficial, or misleading. Such retrieval noise can introduce spurious cues and degrade performance even on questions that models can answer correctly without tool use. These findings are consistent with both our intended characterization of LPFQA —its focus on implicit, knowledge-intensive reasoning beyond explicit proceduralization—and prior observations that equipping models with tools misaligned with the task requirements can lead to performance degradation (Liu et al., 2025b; Shi et al., 2025). Detailed case studies of model performances with/without tools can be found in Appendix J.

5 Conclusion

We present LPFQA, a benchmark constructed from authentic professional forum questions through a human-AI collaborative pipeline for evaluating LLMs on low-frequency, expertise-intensive real-world QA tasks. The results show that current models exhibit limited cross-domain generalization in long-tail professional settings. Further analysis indicates that these limitations arise mainly from insufficient domain expertise, weak contextual comprehension, and inadequate specialized reasoning, while enhancing Search and Code Interpreter even introduces negative gains. These findings offer important insights for the advancement of LLMs toward more robust professional and expertise-intensive capabilities.

Limitations

LPFQA provides a realistic and challenging evaluation of LLMs on long-tail professional knowledge, however, it has two main limitations. First, though LPFQA currently spans 7 primary domains and 50+ forums, it is still limited to publicly avail-

Models \ Tools	No tools	Jupyter CI		Search	
	Score	Score	Δ	Score	Δ
GPT-5-High	54.42	44.50	18.23%↓	44.30	18.60%↓
Gemini-2.5-Pro	47.67	33.50	29.73%↓	30.50	36.02%↓
Deepseek-V3.2-Thinking	49.53	31.78	35.84%↓	31.78	35.84%↓
Kimi-K2-Thinking	47.91	33.50	30.08%↓	30.50	36.34%↓
GLM-4.6	50.47	34.80	31.05%↓	37.30	26.09%↓
Average	50.00	35.62	28.76%↓	34.87	30.26%↓

Table 7: Performances of tool-equipped models. “No tools” denotes the baseline setting without any external tools, while Δ indicates the relative change compared to this baseline.

able English- and Chinese-language communities, potentially missing highly niche or non-public expertise. Second, LPFQA focuses on single-turn QA and does not yet support interactive scenario. These constraints point to clear directions for future extensions.

Ethical Considerations

We have ensured that all data in LPFQA is collected from publicly accessible and vetted forums in compliance with their terms of service. All personally identifiable information has been rigorously anonymized to protect user privacy. Furthermore, we compensated all expert annotators fairly for their professional contributions. While LPFQA covers specialized domains, it is intended strictly for research evaluation for LLM and should not be used as a source of certified professional advice.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62502014.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Zhipu AI. 2025. Glm-4.6. <https://z.ai/blog/glm-4.6>. Accessed: 2025-12-17.
- Anthropic. 2025. Claude-sonnet-4.5. <https://www.anthropic.com/news/claude-sonnet-4-5>. Accessed: 2025-12-17.
- Gavin Brookes and Tony McEnery. 2019. Corpus linguistics for indexing. *The Indexer: The International Journal of Indexing*, 37(2):105–123.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Ameya Godbole and Robin Jia. 2023. Benchmarking long-tail generalization with likelihood splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 963–983.
- Google. 2025. Gemini3. <https://blog.google/products/gemini/gemini-3>. Accessed: 2025-12-17.
- Qiuhan Gu. 2023. Llm-based code generation method for golang compiler testing. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 2201–2203.
- Alex Havrilla, Sharath Raparthi, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravynski, Eric Hambro, and Roberta Raileanu. 2024. Glore: when, where, and how to improve llm reasoning via global and local refinements. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17719–17733.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z Pan. 2025. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems*, page 113648.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pages 15696–15707. PMLR.
- Rohan Kumar, Youngmin Kim, Sunitha Ravi, Haitian Sun, Christos Faloutsos, Ruslan Salakhutdinov, and Minji Yoon. 2024. Automatic question-answer generation for long-tail knowledge. *arXiv preprint arXiv:2403.01382*.
- Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, et al. 2024a. On the role of long-tail knowledge in retrieval augmented large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–126.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.
- Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024c. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18608–18616.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Jian Liu, Kangyun Ning, Yisong Su, Wenjuan Han, Jinan Xu, and Yuanzhe Zhang. 2025b. Wtu-eval: A whether-or-not tool usage evaluation benchmark for large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546.
- Chaitanya Malaviya, Priyanka Agrawal, Kuzman Ganchev, Pranesh Srinivasan, Fantine Huot, Jonathan Berant, Mark Yatskar, Dipanjan Das, Mirella Lapata, and Chris Alberti. 2025. Dolomites: domain-specific long-form methodical tasks. *Transactions of the Association for Computational Linguistics*, 13:1–29.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.
- Prasanna Mayilvahanan, Ricardo Dominguez-Olmedo, Thaddäus Wiedemer, and Wieland Brendel. 2025. Math-beyond: A benchmark for rl to expand beyond the base model. *arXiv preprint arXiv:2510.11653*.
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30.
- Aakanksha Naik, Jill Lehman, and Carolyn Rose. 2022. Adapting to the long tail: A meta-analysis of transfer learning research for language understanding tasks. *Transactions of the Association for Computational Linguistics*, 10:956–980.
- Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.
- OpenAI. 2025a. Gpt-4.1. <https://openai.com/index/gpt-4-1/>. Accessed: 2025-12-17.
- OpenAI. 2025b. *Introducing GPT-5*. Online publication.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*.

- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148.
- Zhengliang Shi, Yuhan Wang, Lingyong Yan, Pengjie Ren, Shuaiqiang Wang, Dawei Yin, and Zhaochun Ren. 2025. Retrieval models aren’t tool-savvy: Benchmarking tool retrieval for large language models. *arXiv preprint arXiv:2503.01763*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.
- Volcengine. 2025. Seed-1.6-thinking. <https://www.volcengine.com/docs/82379/1593703>. Accessed: 2025-12-17.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *EMNLP (Findings)*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, et al. 2026. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33944–33952.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Lu Yang, He Jiang, Qing Song, and Jun Guo. 2022. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872.
- Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35.
- Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816.
- Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6963–6975.
- Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11.
- Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Detailed Domain Distribution of LPFQA

The detailed primary domain and secondary domain distribution of LPFQA is shown at Table 8.

Primary Domain	Secondary Domain
Basic Natural Sciences (Natural Sci)	Physics Chemistry Biology
Life and Medical Sciences (Life Sci)	Biology Medical
Mathematics and Logic (Maths)	Mathematics AI and Machine Learning Data Science
Computer Science (CS)	Computer Science Computer Systems and Software
Electronics Engineering (EE)	Electronic Science and Technology Electronics Information Engineering Electronics&Information Science Info&Communication Engineering
Applied Technologies (Applied Tech)	General Engineering Mechanical and Automation Energy Aerospace
Social Sciences (Social Sci)	Law Finance Miscellaneous

Table 8: Secondary domain list of LPFQA.

B Examples Q&A of LPFQA

Q&A 1, Field: *General Engineering*

Question: When only 110V service is available for a Millport milling machine with a 220V single-phase motor, which power supply solution is recommended, and what key factor must be considered for selecting this equipment?

- A. A voltage regulator with adjustable output, focusing on maximum current capacity alone.
- B. A three-phase to single-phase converter with 110V input, needing to match the motor speed rating.
- C. A Variable Frequency Drive (VFD) with single-phase input and three-phase output, requiring matching the motor's horsepower (HP) rating and current requirements.
- D. A step-up transformer with single-phase input and single-phase output, requiring matching the voltage ratio only.

E. A capacitor-start motor conversion kit, requiring compatibility with motor phase configuration.

F. A DC power supply with inverter function, needing to match the motor frequency range.

Answer: C

Q&A 2, Field: *Biology*

Question: According to the Wright Fisher model of population genetics, please calculate how many generations of living offspring would an average person typically have before the extinction of their genetic lineage?

- A. Approximately 10 generations
- B. Approximately 20 generations
- C. Approximately 25 generations
- D. Approximately 30 generations
- E. Approximately 40 generations
- F. Approximately 60 generations

Answer: B

Q&A 3, Field: *Physics*

Question: Which of the following accurately describes stellar rotation and the key physical factors determining rotation rate and axis orientation?

- A. Not all stars rotate, with key determinants including stellar mass, angular momentum loss during evolution, binary interactions, planetary system formation, and magnetic activity effects
- B. All stars rotate, with key determinants including angular momentum loss through stellar winds, binary interactions, planetary system formation, and magnetic activity effects
- C. Not all stars rotate, with key determinants including initial angular momentum from the parent molecular cloud, angular momentum loss during evolution, binary interactions, planetary system formation, and stellar radius fluctuations
- D. Not all stars rotate, with key determinants including initial angular momentum from the parent molecular cloud, angular momentum loss during evolution, binary interactions, planetary system formation, and magnetic activity effects
- E. All stars rotate, with key determinants including initial angular momentum from the parent molecular cloud, nuclear fusion efficiency, and binary interactions
- F. Not all stars rotate, with key determinants including initial angular momentum from the parent molecular cloud, angular momentum loss during evolution, binary interactions, and magnetic activity effects

Answer: D

Q&A 4, Field: *Computer Science*

Question: In a project management system, when trying to create an external dependency by linking a milestone nested within a phase to an external work package, a 'dependency to parent element exists' error message occurs. Both the phase and the work package are already dependent upon another milestone. The user finds it counter-

intuitive and questions potential circular dependencies, and notes that the dependency works when made from the external milestone directly to the phase. What is the appropriate solution to resolve this error while maintaining valid dependency constraints?

Answer: The error occurs because the system prohibits creating a dependency from a child element (the milestone within the phase) to an external work package when the parent phase is already dependent on another milestone, to prevent indirect circular or redundant dependencies. The correct way to resolve this error is to create the dependency directly between the external milestone and the phase (the parent element), rather than linking the milestone nested inside the phase. By doing so, the dependency respects the system's hierarchical constraints and avoids conflicts with existing parent-level dependencies.

C Prompts of LLMs During Dataset Construction and Judgement.

This benchmark is generated with the assistance of MLLMs and LLMs, and the relevant steps involved the following prompts.

Prompt for image to Q&A

You are a professional expert in data scraping and question generation. Your primary responsibility is to identify the content from forum post screenshots, accurately extract the original poster's question and others' replies, and then generate high-quality questions with a certain level of difficulty and misleading elements based on the scraped content.

Task Steps:

Step 1: Page Validity Check

1. Determine whether the current screenshot is a valid discussion page (not a 404 page, blank page, login page, or advertisement page) 2. If it is an invalid page, return directly: "is_valid": false

Step 2: Question Validity Check

3. If the page is valid, check whether it contains a clear question

4. If no valid question is found (e.g., title only without content, meaningless characters, or simple informational posts without questions), return: "is_valid": false

Step 3: Content Extraction and Processing

5. If there is a valid question and valid replies (substantial content, not just emojis or meaningless characters)

6. Extract the title, question description, and all relevant reply content to form a structured triplet: (question, answer, context)

Step 4: Result Generation

7. Based on the extracted core content, create high-quality academic questions that meet the following criteria:

- The question must be self-contained and independent of external context
 - Rephrase the question using clear and natural language to enhance clarity, fluency, and understandability, avoiding ambiguity or vague expressions
 - Be sufficiently professional — not simple knowledge quizzes or common-sense questions, but those requiring logical reasoning and analysis
 - Require multi-step, cross-disciplinary complex reasoning
 - Involve specialized or cutting-edge knowledge in specific domains, not purely general academic knowledge
 - Avoid simple factual questions; emphasize comprehension and application
 - The answer must be unique and verifiable, capable of being confirmed through reasoning or experimentation; exclude open-ended questions with multiple possible answers or solution paths
 - Use the same language as the original forum post for the question and answer (e.g., if the post is in English, both question and answer should be in English)
8. Summarize and organize the most accurate, comprehensive, and detailed answer from the replies (leave blank if no valid reply exists)
9. Provide a detailed explanation and reasoning process for the answer (leave blank if no valid reply exists)
10. Set the "has_answer" flag: 1 if there are valid replies and an answer can be extracted, 0 otherwise

Please strictly return the result in the following JSON format without adding any extra text:

```
{
  "is_valid": boolean,
  "question": string,
  "answer": string,
  "explanation": string,
  "has_answer": integer
}
```

Prompt for quality filter

Task: Evaluate the quality of the following QA pair and retain only high-quality questions

Evaluation Criteria (all must be satisfied):

- Accuracy: All background knowledge, data, and facts involved in the question are accurate and error-free
- Completeness: The question provides all information necessary for solving it, with no missing or redundant information
- Professionalism: The question is sufficiently professional—not a simple knowledge quiz or common-sense question—but one that requires logical reasoning and analysis to solve
- Timelessness: The question does not involve time-sensitive content; it remains valid and true at all times

Difficulty Criteria: The question must meet the following requirements:

- Depth of Knowledge: Solving the question requires specialized or cutting-edge knowledge in a particular field
- Length of Reasoning Chain: The solution process involves multi-step, cross-disciplinary complex reasoning
- Abstraction and Synthesis Requirements: The question demands high-level abstract thinking and information synthesis
- Deceptiveness and Originality: The question requires creative or unconventional problem-solving approaches

QA pair to be evaluated:

Question: {question}

Answer: {answer}

Explanation: {explanation}

Please return only the result in JSON format, including the following fields:

- is_qualified: boolean (whether the question meets the standards)
- reason: string (reason for not meeting the standards; empty string if qualified)

Prompt for multi-choice

Task: Convert the question, answer, and explanation from the following QA pair into a high-quality multiple-choice question
Requirements:

1. Retain the core of the original question, but adjust the phrasing appropriately to suit the multiple-choice format
2. Generate 5 to 7 options, with exactly one correct answer and the rest as plausible distractors
3. Distractors should be misleading and belong to the same knowledge domain as the correct answer
4. Options must be randomly ordered
5. Ensure all options have a consistent grammatical structure
6. Maintain a high difficulty level, requiring multi-step reasoning to solve
7. The language used must be consistent with the original question—if the original QA pair is in English, the output must also be in English; do not change the language arbitrarily

QA pair:

Question: {question}

Answer: {answer}

Explanation: {explanation}

Please return only the result in JSON format, including the following fields:

- question: string (the multiple-choice question)
- options: array[string] (list of options, including one correct answer)
- correct_answer: string (the content of the correct answer)
- explanation: string (explanation of why this option is correct)

Prompt for short answer

Task: Convert the question, answer, and explanation from the following QA pair into a high-quality short-answer question, and extract the core knowledge points

Requirements:

1. Retain the core of the original question, with appropriate rephrasing to better suit the short-answer format
2. Generate a detailed and comprehensive answer that includes all key information from the original explanation
3. The answer must be substantial, well-supported, and logically structured, avoiding vague or generic statements
4. Use natural, fluent, professional, and accurate language
5. Maintain a high difficulty level, reflecting professional knowledge and reasoning processes
6. Summarize 1–5 core knowledge points—these are the key concepts essential for determining the correctness of the answer
7. The language used must be consistent with the original question—if the original QA pair is in English, the output must also be in English; do not change the language arbitrarily

QA pair:

Question: {question}

Answer: {answer}

Explanation: {explanation}

Please return only the result in JSON format, including the following fields:

- question: string (the short-answer question)
- answer: string (detailed answer, incorporating the explanation)
- core_knowledge_points: array[string] (list of core knowledge points)

Prompt of metric for answer evaluation

From now on, your role is that of a professional grading teacher. Your task is to objectively score the <Student Answer> based on the <Reference Answer> and <Evaluation Points> that I provide. The main steps and rules include the following 6 points:

1. **Scoring Levels:** There are only two levels, 0 and 1. No other scores are allowed, so please do not give any score other than 0 or 1.
2. Since I have not provided you with the <Question>, you only need to judge whether the content of the <Student Answer> meets the <Evaluation Points>. Do not imagine or infer the content of the <Question>. Note that <Evaluation Points> take precedence over the <Reference Answer>. If there is a conflict between them, you must prioritize the requirements of the <Evaluation Points>.
3. You need to first determine whether the <Student Answer> meets the requirements of the <Evaluation Points> (if no <Evaluation Points> are provided, then by default, the results and requirements in the <Reference Answer> are considered as the evaluation points). The answer must meet all conditions in the evaluation points. If any one of them is not satisfied, you should immediately give a score of 0, stop executing Step 4, and output the result according to the format in Step 5. If all conditions are met, continue to Step 4.
4. Determine whether the final answer in the <Student Answer> is correct:
 - You do not need to consider whether the process is correct, only the final answer.
 - If the final answer is identical or synonymous to the <Reference Answer> (e.g., “two times” and “2”), then give a score of 1. Otherwise, if it is inconsistent or not synonymous, give a score of 0.
 - Some answers may contain errors, but if later self-corrected, the final result after correction should be regarded as the actual answer.
 - Some answers may include self-analysis or feedback, possibly repeating outputs multiple times. Do not be misled by intermediate

outputs—only use the last final result for grading.

5. **Format Requirements:** Before outputting your result, first check your analysis process and score to ensure they are reasonable and correct. If there are any errors or omissions, revise them in time. Finally, when you believe everything is correct, provide the output strictly in the following format:
 - **Grading Basis:** (Concise explanation of the grading reason, less than 100 tokens)
 - **Score:** x (output must be “0” or “1”)
 - **JSON:**
{"answer_score": your score}
6. Before you formally start the grading task, to improve your grading accuracy and understanding of the scoring standards, I will first provide a simulated scoring example. By studying this example, you will become more familiar with the grading process and master the grading techniques:

Example

<Reference Answer>:

5. Move to the kitchen.
 3. Find the kettle, check that it has water.
 2. Pick up the kettle and move to the balcony.
 4. Water the flowers.
 Path: 5, 3, 2, 4

<Evaluation Points>:

1. The “watering process” in the student’s answer must be exactly consistent with the reference answer; otherwise, score 0.
2. The “path” in the student answer must be exactly consistent with the reference answer, otherwise score 0.

<Student Answer>:

5. Move to the kitchen.
 1. Find a sponge, check if it has water.
 6. Pick up the sponge and move to the balcony.
 4. Water the flowers.
- Path: 5, 1, 6, 4

Grading Basis: Both process and path are inconsistent with the reference answer, hence score 0.

Score: 0

JSON:

```
{"answer_score": 0}
```

I hope you can fulfill the role of grading teacher. If you perform well, I will give you appropriate rewards. Please strictly follow the output format I provide; otherwise, I will penalize you. Also, always use the final result of the answer for grading, and do not be misled by intermediate outputs.

D Forum list

The forums selected include, but are not limited to, the following:

<https://mathoverflow.net/>
<https://geant4-forum.web.cern.ch/>
<https://root-forum.cern.ch/>
<https://www.physicsforums.com/>
<https://www.biostars.org/>
<https://eaaforums.org/>
<http://www.mjtd.com/>
<https://www.practicalmachinist.com/>
<https://diysolarforum.com/>
<https://cr4.globalspec.com/thread/88025/High-Voltage-Engineering>
<https://www.elitetrader.com/>
<https://answers.justia.com/>
<https://www.lawanswers.com.au/forums/defamation-law-forum.25/>
<https://artofproblemsolving.com/community>
<https://board.asm32.info/>
<https://learn.microsoft.com/en-us/answers/topics/azure-digital-twins.html>
<https://www.biostars.org/>
<http://forums.corvetteforum.com/index.php>
<http://www.cplusplus.com/forum/>
<https://discuss.dvc.org/>

<https://forum.filezilla-project.org/index.php>
<http://www.fluka.org/fluka.php?id=mailinglist&mm2=6>
<https://engx.theiet.org/>
<https://forums.majorgeeks.com/>
<https://community.myfitnesspal.com/en/categories/forums>
<http://www.openedv.com/>
<https://www.physicsforums.com/>
<https://discuss.pytorch.org/>
<https://forums.raspberrypi.com/>
<https://www.reddit.com/r/math/>
<https://root-forum.cern.ch/>
<https://community.spiceworks.com/>
<https://stackoverflow.com/questions/tagged/robotics>
<https://www.statalist.org/forums/forum/general-stata-discussion/general>
<https://www.techpowerup.com/forums/>
<https://pcengine.proboards.com/>
<https://www.vnpy.com/forum/>
<http://forums.vwvortex.com/>
<https://guba.eastmoney.com/>
<https://bbs.pinggu.org/>
<http://www.mjtd.com/>
<http://www.3dportal.cn/>
<http://www.prowildfire.cn/>
<https://www.armbbs.cn/>

All these forums are publicly accessible platforms that either permit or do not prohibit the use of their content for the AI community.

E Expert Qualification and Rules

A total of 209 experts participated in this project, with their qualifications categorized and described across three dimensions: (1) professional field, (2) educational background, and (3) institutional affiliation. All data was self-reported by the experts and verified internally.

1. Professional Domain Distribution. Experts are categorized according to their self-declared primary field. To maintain data integrity, all original categories—including non-academic ones—are retained, totaling 209 individuals.

2. Educational Attainment. The educational background of the expert panel is highly advanced, with the vast majority holding graduate degrees.

3. Institutional Affiliation. The institutional prestige is classified using the QS World University Rankings and inclusion in national elite research

Primary Domain	Count	Pct. (%)
CS & Mathematical Sciences	87	41.6
STEM	60	28.7
Business	20	9.6
Medical	13	6.2
Law	9	4.3
Humanities & Social Sciences	8	3.8
Lifestyle & Hobbies	5	2.4
Education	4	1.9
Languages	4	1.9
Creative Writing	2	1.0
Total	209	100.0

Table 9: Distribution of Experts by Primary Professional Domain

Degree Level	Count	Pct. (%)
Ph.D.	94	45.0
Master’s	97	46.4
Bachelor’s	15	7.2
Not specified / Other	3	1.4
Total	209	100.0

Table 10: Educational Attainment of Experts. For currently enrolled students, degree level refers to the degree they are pursuing; for graduates and working professionals, it refers to the highest degree obtained.

systems. Institutions are grouped as Table 11.

The National-Level Research Institutes include the University of Chinese Academy of Sciences, Chinese Academy of Medical Sciences, People’s Liberation Army General Hospital, etc., which entities with doctoral granting authority and national research mandates.

During the data construction process, the experts are asked to select forums that either permit or do not prohibit the use of their content for the AI community. All experts are paid reasonable remuneration based on their workload in accordance with the market price.

F Models Evaluated on LPFQA

6 open-source models: Qwen-3-235B-Instruct, Qwen3-235B-Thinking (Yang et al., 2025), DeepSeek-V3 (Liu et al., 2024), DeepSeek-V3.2-Thinking (Liu et al., 2025a), Kimi-K2-Thinking (Team et al., 2025) and GLM-4.6 (AI, 2025).

7 closed-source models: GPT-4o (Achiam et al., 2023), GPT-4.1 (OpenAI, 2025a), GPT-5-

Institution Category	Count	Pct. (%)
QS World Top 50 Univ.	91	43.5
QS World Rank 51–200 Univ.	32	15.3
National-Level Research Inst.	54	25.8
Other/Unspecified Inst.	32	15.3
Total	209	100.0

Table 11: Institutional Affiliation of Experts

high (OpenAI, 2025b), Gemini-2.5-Pro (Comanici et al., 2025), Gemini-3-Pro (Google, 2025), Claude-Sonnet-4.5-Thinking (Anthropic, 2025), and Seed-1.6-Thinking (Volcengine, 2025).

G Effectiveness of LLM-Judge on LPFQA

To validate the reliability of the LLM-based judge for evaluating LPFQA tasks, we randomly sampled more than 250 responses generated by different models and asked human experts to assign the ground-truth binary labels. We then compared the human annotations with the scores assigned by the LLM judge. The results show that the LLM judge achieves a **92.9%** agreement rate with human evaluation.

For the vast majority of tasks, evaluation mainly requires comparing the candidate answer against the expert-verified reference answer, making the judgment process relatively objective and less sensitive to the specific choice of judge model.

H Detailed Error Type Distribution on Different Primary Domains in LPFQA.

The detailed distributions of error attributions of different primary domains in LPFQA are shown in Table 12.

I Detailed Scores of LLMs on MMLU-pro and HLE

The detailed scores and rankings of different models on LPFQA, MMLU-pro and HLE are shown in Table 13.

Discipline	Over-simplify	Reasoning	Knowledge	Hallucination	Misinterpretation	Calculation	Others
Natural Sci	6	20	50	4	18	2	0
Applied Tech	8	12	44	4	30	0	2
Maths	4	10	40	0	34	6	6
Life Sci	4	24	52	0	18	0	2
EE	6	18	40	2	34	0	0
Social Sci	6	22	42	2	26	0	2
CS	4	8	58	0	28	0	2

Table 12: Percentage distribution (%) of error types across disciplines in LPFQA. Each discipline contains 50 questions; percentages are obtained by multiplying raw counts by 2.

Model	LPFQA		MMLU-pro		HLE	
	Score	Rank	Score	Rank	Score	Rank
GLM-4.6	50.5	3	83.2	7	17.2	6
Gemini-3-Pro	51.2	2	90.2	1	37.7	1
Kimi-K2-Thinking	47.9	6	84.8	5	23.9	4
GPT-5-High	54.4	1	87.1	3	26.3	2
Claude-Sonnet-4.5-Thinking	49.5	4	87.5	2	13.7	7
DeepSeek-V3.2-Thinking	49.5	4	85.0	4	25.1	3
Qwen3-235B-Thinking	46.9	7	84.5	6	18.2	5

Table 13: Performance comparison of LLMs across LPFQA, MMLU-pro, and HLE. The scores of models on MMLU-pro and HLE are mainly obtained from models’ official technical reports and model cards.

J Case Studies on Model Performance when with/without Tools

J.1 CI Tool

Typical Failure Case: LPFQA JupyterCI Leads to Decreased Model Score

Task Description

Question. Can every real number in the unit interval $[0, 1]$ be expressed as the product of finitely many elements of the Cantor set? If so, what is the minimum number of Cantor set elements required?

- A. Yes; the minimum number required is 2
- B. Yes; the minimum number required is 4
- C. Yes; the minimum number required is 3
- D. Yes; the minimum number required is 1
- E. Yes; the minimum number required is 5
- F. No; there exist real numbers in $[0, 1]$ that cannot be expressed as the product of finitely many Cantor set elements

Ground-truth answer: Yes. The minimum number is 3.

Model Behavior without JupyterCI

Without external tools, the model directly recalls a known theoretical result in fractal geometry and real analysis:

$$C \cdot C \cdot C = [0, 1],$$

while the double product $(C \cdot C)$ does not cover the entire interval. The model therefore selects Option C (3 elements) correctly. This demonstrates that the model possesses the relevant latent mathematical knowledge and can apply it via pure language-based reasoning.

Model Behavior with JupyterCI

Step 1: Tool Invocation and Problem Reformulation

Upon being equipped with a Code Interpreter (CI), the model reformulates the original question as a computational coverage verification task. Specifically, it:

- Constructs finite-level approximations (C_n) of the Cantor set;
- Computes the pairwise product $(C_n \cdot C_n)$;

- Evaluates numerical statistics such as maximum gap, nearest distance to grid points, and empirical coverage over $[0, 1]$.

Step 2: Interpretation of Numerical Outputs

The CI outputs indicate persistent gaps and non-zero distances for all tested finite levels ($n \leq 10$), for example stable maximum gaps around (0.11). The model interprets these results as evidence that the product of two Cantor elements fails to cover $[0, 1]$.

Step 3: Tool-Induced Reasoning Lock-in

After committing to the computational path, the model does not reconsider or override the numerical evidence with known theory. Instead, it:

- Constructs a post-hoc theoretical justification (via logarithmic transformation and thickness arguments);
- Incorrectly concludes that two factors suffice, selecting Option A.

Failure Analysis

With JupyterCI enabled, the model selects Option A (2 elements), yielding an incorrect answer, despite answering correctly in the tool-free setting.

Summary of Failure Mode

This case exemplifies a systematic degradation pattern induced by external tools:

- Tools bias the model toward procedural, simulation-based reasoning, even when inappropriate;
- Finite numerical evidence is misused to answer questions about infinite mathematical objects.

Takeaway

The resulting reasoning lock-in prevents recovery via pure language-based knowledge. Importantly, the error arises not from missing information or computational failure, but from a mismatch between the tool's affordances and the task's epistemic structure.

J.2 Search Tool

Typical Failure Case: LPFQA Search Tool Leads to Decreased Model Score

Question: In a computational study of perovskite solar cell materials, a research team performs an excited-state geometry optimization on the $\text{CH}_3\text{NH}_3\text{PbI}_3$ crystal. The material crystallizes in a tetragonal lattice. The team first constructs a $2 \times 2 \times 1$ supercell and maps the primitive-cell coordinates to the supercell. After building the structure, they carry out an S_1 excited-state optimization using the ORCA software package at the PBE0/def2-SVP level of theory. Upon convergence with the VeryTightOpt criterion, a single significant imaginary frequency of -32 cm^{-1} is observed.

Which of the following statements regarding this study is correct?

- The mapping from the primitive cell to the $2 \times 2 \times 1$ supercell fails because the axial-angle characteristics of the tetragonal lattice prevent the use of simple translation vectors.
- To resolve the imaginary frequency, one should immediately search for a transition state using the Opt=TS keyword in Gaussian.
- Errors in the supercell coordinate construction have a direct causal relationship with the appearance of the imaginary frequency in the excited-state optimization.
- The most reasonable way to handle the imaginary frequency is to apply ORCA's iterative optimization method by including the %compound iterativeOptimization.cmp setting.
- Owing to the high symmetry of the tetragonal lattice, the coordinates of the $2 \times 2 \times 1$ supercell can be obtained from the primitive-cell coordinates via simple matrix multiplication.

Ground-truth answer: E — Supercell coordinates in a tetragonal lattice can be ob-

tained via a simple linear (matrix) transformation from the primitive cell.

Model Behavior without Search

When no external search tool is available, the model selects **E**. Its reasoning focuses on the crystallographic principle that supercell construction is a linear transformation independent of vibrational artifacts arising in excited-state optimization. The imaginary frequency is correctly treated as irrelevant to the coordinate-mapping question posed.

Model Behavior with Search

When equipped with a search tool, the model invokes external retrieval and obtains ORCA manual pages and forum discussions describing the *iterative optimization* compound script for removing imaginary frequencies. After incorporating this retrieved information, the model selects **D and E**, incorrectly elevating **D** as a correct choice. In particular, the model treats the existence of an ORCA workflow for eliminating imaginary frequencies as evidence that it is the most reasonable action in this scenario, despite the question not asking how to resolve the frequency issue.

Failure Analysis

This case exemplifies a **search-induced failure mode** in LPFQA, where tool usage degrades performance relative to the tool-free setting. First, the retrieved content is **semantically misaligned with the task objective**. The search tool returns method-level and documentation-level descriptions of how to remove imaginary frequencies, whereas the question requires a contextual judgment about crystallographic supercell mapping. Second, the model **misuses existence as normativity**. The availability of a technically valid ORCA procedure is incorrectly promoted to a normative recommendation for this specific problem, without considering whether the procedure is relevant to what is being asked. Third, **external retrieval disrupts internal domain reasoning**. Exposure to detailed procedural information diverts the model's attention away

from its original, correct crystallographic reasoning path. Finally, tool availability induces an **over-engineering bias**: instead of answering a conceptual multiple-choice question, the model proposes an executable computational workflow, leading it to select an option that is technically plausible but conceptually out of scope.

Takeaway

This example demonstrates that, for low-retrievability professional questions, external search can introduce technically correct but contextually inappropriate information.

K Result of 6-Trial Run Experiment on LPFQA

The detailed result of representative models' performance on LPFQA is shown at Table 14.

Model	Original (%)	6-trial Mean (%)	Std.	95% CI
Gemini-3-Pro	51.16	51.20	1.17	[49.97, 52.43]
GLM-4.6	50.47	50.39	0.70	[49.65, 51.12]
DeepSeek-V3.2-Thinking	49.53	48.10	1.37	[46.67, 49.54]
GPT-4.1	46.51	46.28	1.60	[44.60, 47.96]
Seed-1.6-Thinking	45.35	46.24	1.28	[44.90, 47.58]

Table 14: Performance stability across 6 evaluation trials for representative models on LPFQA.