ELAIPBENCH: A Benchmark for Expert-Level Artificial Intelligence Paper Understanding

Anonymous ACL submission

Abstract

While large language models (LLMs) excel at many domain-specific tasks, their ability to deeply comprehend and reason about full-length academic papers remains underexplored. Existing benchmarks often fall short of capturing such depth, either due to surface-level question design or unreliable evaluation metrics. To address this gap, we introduce ELAIPBENCH, a benchmark curated by domain experts to evaluate LLMs' comprehension of artificial intelligence (AI) research papers. Developed through an incentive-driven, adversarial annotation process, ELAIPBENCH features 403 multiple-choice questions from 137 papers. It spans three difficulty levels and emphasizes non-trivial reasoning rather than shallow retrieval. Our experiments show that the best-performing LLM achieves an accuracy of only 39.95%, far below human performance. Moreover, we observe that frontier LLMs equipped with a thinking mode or a retrieval-augmented generation (RAG) system fail to improve final results—even harming accuracy due to overthinking or noisy retrieval. These findings underscore the significant gap between current LLM capabilities and genuine comprehension of academic papers.

1 Introduction

In recent years, large language models (LLMs) have made remarkable advancements in acquiring domain-specific knowledge and enhancing reasoning capabilities. These models encapsulate extensive expertise across various fields, enabling performance at or beyond human-expert levels on multiple scientific benchmarks (Achiam et al., 2023; Lála et al., 2023; Laurent et al., 2024; Du et al., 2025). Furthermore, LLMs have demonstrated strong reasoning abilities on tasks that require logical deduction (Liu et al., 2024; Yang et al., 2025), including mathematics (Hendrycks et al., 2021) and programming (Hui et al., 2024; Jain et al., 2025).

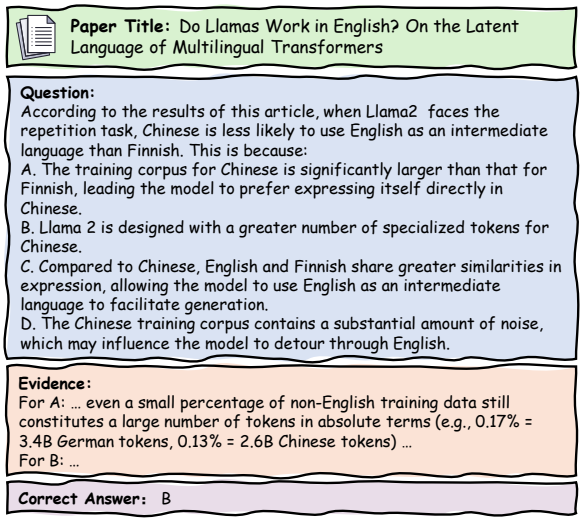

Figure 1: An example from ELAIPBENCH. Option A, while potentially factually correct, is not substantiated by the text. Option C is a deliberately constructed distractor with no direct supporting evidence in the text. Option D is also not explicitly stated in the text, although its latter part is derived from the paper.

Given that artificial intelligence (AI) research often requires a strong command of both mathematics and algorithmic programming, these advancements have led AI researchers to increasingly employ LLMs as tools for reading, interpreting, and even reviewing academic papers. However, this raises a critical question: **To what extent can LLMs truly comprehend academic content—specifically, are they capable of deep understanding, learning, and reasoning based on lengthy scholarly texts?**

Unfortunately, existing domain-specific benchmarks fail to adequately assess the deep comprehension of LLMs for academic papers. Some benchmarks focus on content extraction tasks where answers are explicitly stated in the papers (Lála et al., 2023; Auer et al., 2023; Laurent et al., 2024). These answers can be easily retrieved by LLMs due to their strong recall abilities. Others employ open-ended questions requiring the summarization

of phenomena or methods described in the papers (Saikh et al., 2022; Lee et al., 2023; Wan et al., 2024; Wellawatte et al., 2025); however, their evaluation metrics (e.g., ROUGE) fail to accurately and reliably measure fine-grained details or key technical knowledge. Crucially, existing approaches relying on human annotation struggle with quality control—although some incorporate question validation mechanisms (Welbl et al., 2017; Bai et al., 2024), crowd-sourced annotation schemes may still produce superficial or low-quality questions (Kittur et al., 2008; Snow et al., 2008).

To address these limitations, we aim to construct a benchmark with the following features: (1) **Reliability**: all evaluations should be conducted in the form of multiple-choice questions, ensuring a precise and standardized format; (2) **Difficulty**: the questions should be challenging enough that even experienced researchers cannot answer all of them correctly in a short time; (3) **High Quality**: all questions must be answerable, and each must be accompanied by evidence-supported answers to ensure both validity and interpretability.

With the above goal in mind, we introduce ELAIPBENCH, a benchmark designed to evaluate **E**xpert-**L**evel **A**rtificial **I**ntelligence **P**aper understanding. ELAIPBENCH encompasses papers from the domains of machine learning (ML), computer vision (CV), and natural language processing (NLP). It comprises 403 manually crafted multiple-choice questions in both single-answer and multiple-answer formats, thereby ensuring the **reliability** of evaluation. The benchmark, presented in English, provides each question with its full-length source paper, the correct answers, and the corresponding evidence excerpt from the paper.

To ensure the **difficulty** and **high quality** of ELAIPBENCH, we employ a competitive incentive annotation process to create challenging questions. We recruit 20 human annotators, each holding at least a master's degree in computer science and having prior experience in publishing academic papers. Each annotator is assigned one of three roles: *Question Writer*, *Evidence Verifier*, or *Answer Verifier*.The *Question Writer* is responsible for creating data instances. Each instance comprises an English academic paper, a set of questions with corresponding answer choices, the correct answer, and supporting evidence excerpts from the paper. Subsequently, as a preliminary difficulty filter, we use three LLMs for automated review. If any model answers the question correctly, it is discarded as too easy. The filtered questions are assigned to Evidence Verifiers, who answer them using the paper and evidence excerpts. A question is retained only if the *Evidence Verifier* derives the correct answer using the provided evidence. Finally, the *Answer Verifier* attempts to answer the questions based solely on the paper, question, and options. Throughout the process, annotators receive a base salary and compete for performance bonuses based on response time and accuracy. This mechanism categorizes questions into three difficulty levels and motivates annotators to create challenging yet answerable questions. Figure 1 presents an example question from ELAIPBENCH. Its options not only incorporate explicit content from the paper but also require LLMs to perform reasoning based on this content combined with scientific common knowledge. Overall, our contributions are summarized as follows:

- We introduce ELAIPBENCH, a benchmark containing 403 expert-created questions with three difficulty levels. The questions are precise and unambiguous, and they require deep understanding beyond simple paper retrieval.
- We propose a competitive incentive annotation mechanism that uses performance-linked bonuses to generate high-quality questions. Comprehensive evaluations show the benchmark's difficulty: human experts achieve only 48.14% accuracy within 20 minutes, while the best-performing LLMs reach just 39.95%.
- In our experiments, we observe that LLMs augmented with either reasoning or retrieval capabilities generally underperform non-augmented baselines on ELAIPBENCH. This suggests that current naive augmentation methods may lead to misinterpretations of papers, which highlights an unexplored gap for future research.

2 Related Works

2.1 Scientific Comprehension Benchmarks

To advance the evaluation of LLMs in the domain of scientific comprehension, a variety of benchmark datasets have emerged, with increasing emphasis on task complexity and domain specificity. These benchmarks can be broadly categorized into two groups. The first group either does not require contextual information or relies on only brief text

passages. Sorted by publication date, these benchmarks include emrQA (Pampari et al., 2018), PubMedQA (Jin et al., 2019), ScienceQA (Saikh et al., 2022), SciQA (Auer et al., 2023), SciBench (Wang et al., 2023), ChemLit-QA (Wellawatte et al., 2025), and HLE (Phan et al., 2025). They primarily assess models' intrinsic knowledge within a specific domain while neglecting the critical ability to comprehend contextual information—an essential skill in academic research or paper review. The second group incorporates full academic papers as context (Lála et al., 2023; Skarlinski et al., 2024; Bai et al., 2024); however, the associated questions tend to be superficial, enabling LLMs to locate answers through direct text matching rather than requiring synthesis, inference, or critical engagement with the academic paper context.

2.2 Academic Benchmark Annotation Methods

Current academic question answering benchmarks are primarily generated using two annotation approaches. The first relies on crowdsourced human annotation, where data labeling is conducted through pipeline-based workflows and quality control is typically managed by a review group (Lála et al., 2023; Bai et al., 2024; Skarlinski et al., 2024; Asai et al., 2024). This approach, however, heavily depends on the efficiency and expertise of the reviewing team, with limited mechanisms for providing feedback on the review outcomes. The second approach leverages LLMs to automatically generate datasets by equipping them with various specialized tools and prompting frameworks (Lee et al., 2023; Wan et al., 2024; Kim et al., 2025; Yu et al., 2025). While scalable, this method is constrained by the generative model's own comprehension ceiling. The resulting questions, therefore, often lack diversity in reasoning types and fail to achieve sufficient depth in academic problem-solving.

3 Construction of ELAIPBENCH

In this section, we present the construction process of ELAIPBENCH. First, we define the task overview. Next, we introduce the incentive competition annotation mechanism designed for generating high-quality, challenging data. Finally, we provide statistical information about the benchmark.

3.1 Task Overview

Task Definition. The task is formally defined as follows: given an academic paper text p from the AI domain, published in a conference or journal, and a multiple-choice question q, the model is required to identify all correct answer choices based on the content of the paper. Any response containing incorrect options or failing to include all correct choices receives zero points.

Annotation Team. To collect high-quality and challenging question-answering (QA) data derived from academic papers, we recruit a team of 20 annotators from top universities. We also develop an online annotation platform designed to facilitate the annotation process. All annotators have previously published papers in the field of AI and are fluent in English. Specifically, the team comprises one professor, one postdoctoral researcher, ten Ph.D. students, and eight master's students, all specializing in AI. Further details about the annotation platform and team are provided in Appendix A.2.

3.2 Competitive Incentive Annotation Mechanism

As illustrated in Figure 2, our incentive-driven competitive mechanism consists of three key stages: QA pair generation, Evidence Verification, and Answer Verification. These stages involve three distinct roles (*Question Writer*, *Evidence Verifier*, and *Answer Verifier*), staffed with 10, 4, and 6 annotators, respectively. The roles operate independently yet pursue competing objectives within the workflow. Compensation is structured as a combination of base pay and performance-based bonuses. Annotators receive a base payment of 30 CNY per question they generate. During the verification stages, they could earn additional performance bonuses of 30 CNY (Level-1) or 60 CNY (Level-2), awarded based on the quality of their annotations. The total cost for all data collection and annotation was approximately 50,000 CNY.

3.2.1 QA Pair Generation

In the QA Pair Generation stage, each *Question Writer* is required to first upload a long paper from a top conference or journal with which they are thoroughly familiar. The annotation platform employs a PDF-to-text tool such as PyMuPDF[1] to transform the uploaded file into plain text. This preprocessed text is then automatically checked for length, and any document containing fewer than 4,096 tokens is rejected due to insufficient content for generating challenging questions. Following

[1] https://github.com/pymupdf/PyMuPDF

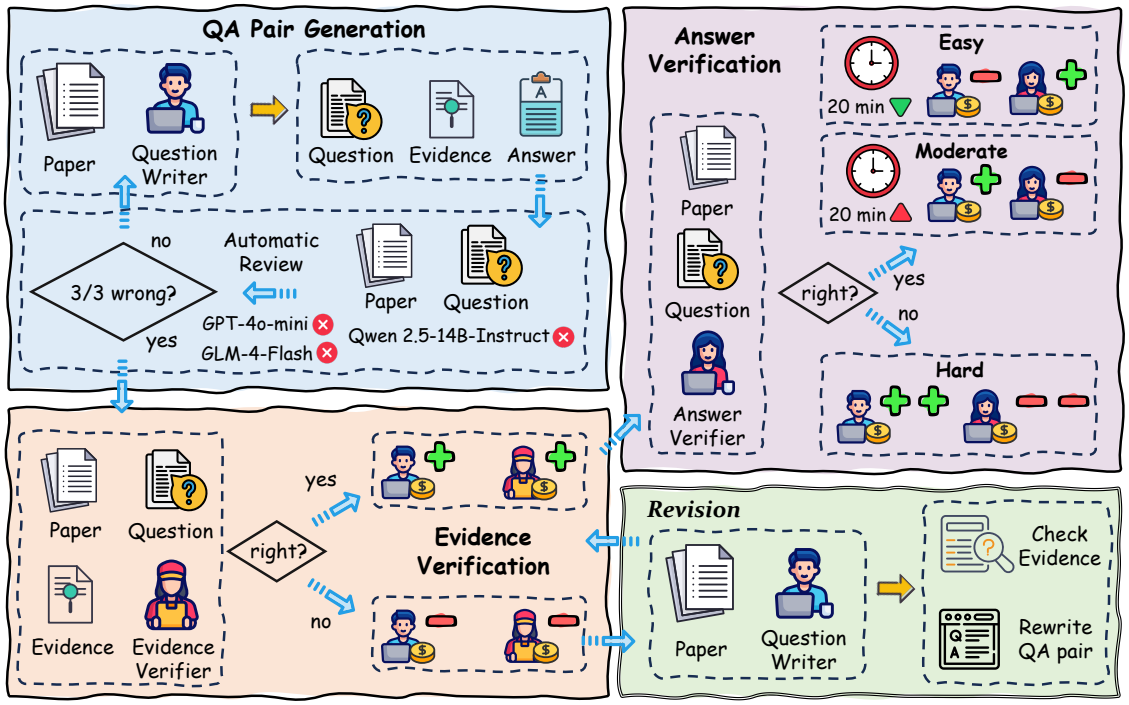

Figure 2: Construction of ELAIPBENCH.

successful preprocessing, the *Question Writer* creates both single-answer multiple-choice questions (SA-MCQ) and multiple-answer multiple-choice questions (MA-MCQ) according to detailed annotation guidelines. Each SA-MCQ must have exactly one correct answer, while each MA-MCQ must have one or more correct answers. All correct answers must be grounded in factual content from the paper and require corresponding supporting evidence. This supporting evidence must consist of verbatim excerpts from one or more paragraphs of the source paper. Its sufficiency is determined by whether an AI expert could correctly answer the question based solely on this evidence.

Our annotation guidelines are as follows: (1) Avoid selecting highly-cited or canonical academic papers to ensure diversity and mitigate the risk of models relying on memorized knowledge. (2) Questions must be objectively formulated in English, free of subjectivity or ambiguity. (3) The correct answer must not be obtainable through simple keyword matching or direct retrieval from the text; instead, answering the question should require reasoning, summarization, or integration of multiple pieces of information from different sections of the paper. (4) Questions and options should not introduce domain-specific terms or abbreviations unless they are explicitly defined or already appear in the paper, as such content may confuse LLMs or introduce unintended ambiguity. (5) Distractors should be carefully designed to include misleading phrasing or partial truths—statements that align with some aspects of the paper but are ultimately incorrect—thereby increasing the cognitive demand and necessitating careful comparison and deep reasoning. (6) For SA-MCQ, exactly one option must be correct. (7) For MA-MCQ, the number of correct options must be either two or three, a constraint designed to ensure a balanced level of difficulty and discourage random guessing. The detailed annotation guidelines are provided in Appendix A.1.

The *Question Writer* must provide: (1) a paper, (2) a question, (3) verbatim evidence from the source paper corresponding to each option, and (4) the correct answer. After the initial annotation, the *Question Writer* evaluates the question's difficulty during a process we term the Automatic Review stage. In this stage, the writer prompts three LLMs on our platform—`GPT-4o-mini`, `Qwen 2.5-14B-Instruct` and `GLM-4-Flash`—to answer the question based solely on the provided paper. Questions incorrectly answered by all three models proceed to Evidence Verification. Conversely, if any of the three models answers the question correctly during this stage, it is deemed insufficiently challenging and must be revised by the *Question Writer* to increase its difficulty.

3.2.2 Evidence Verification

In this stage, *Evidence Verifiers* assess whether the provided evidence excerpts are sufficient to support a correct response. This requires: (1) analyzing and reasoning over the evidence, and (2) attempting to answer the question using only this evidence. To incentivize emphQuestion Writers to provide accurate and logically aligned evidence, we implement a competitive mechanism: if the *Evidence Verifier* arrives at the correct answer using exclusively the provided evidence, both receive Level-1 performance bonuses. Conversely, if verification fails, penalties apply: the *Question Writer* and the *Evidence Verifier* lose their bonus allocation. This failure triggers a *Revision* cycle where the *Question Writer* must improve either the evidence or the entire QA pair. The revised version must pass both Automatic Review and subsequent Evidence Verification before either party is rewarded. This dual-incentive mechanism simultaneously maintains question difficulty standards and compels *Question Writers* to provide logically sound evidence supporting the answer. During the annotation process, 67% of the questions initially failed the Automated Review, and 38% failed the initial Evidence Verification.

3.2.3 Answer Verification

In the Answer Verification phase, two constraints govern the process: *correctness* and *time limitation*. Specifically, the *Answer Verifier* attempts to answer each question within 20 minutes using only the provided paper and without any external evidence. Based on the outcome of this process, questions are classified into three levels: (1) **Easy**: The question is classified as easy if the verifier answers correctly within the time limit. In this case, the *Question Writer* forfeits their bonus, while the *Answer Verifier* receives a Level-1 bonus, indicating that the question posed an insufficient challenge. (2) **Moderate**: The question is classified as moderate if the *Answer Verifier* responds correctly but exceeds the 20-minute threshold. In this scenario, the *Question Writer* receives a Level-1 bonus, while the *Answer Verifier* receives no bonus, reflecting the verifier's inadequate efficiency. (3) **Hard**: The question is classified as hard if the *Answer Verifier* fails to provide the correct answer. In this case, the *Question Writer* receives a Level-2 bonus, and the *Answer Verifier* again receives no bonus.

This mechanism establishes a dynamic adversarial relationship between the *Question Writer* and the *Answer Verifier*: *Question Writers* are incentivized to maximize question difficulty for higher compensation, while *Answer Verifiers* are motivated to provide accurate responses efficiently to secure their bonuses.

3.3 Benchmark Statistics

Tables 1 presents the distribution of data. For more statistics and examples, please see Appendix A.3.

Statistics	Numbers	Average Length
Paper	137	15,012.85
Question	403	140.17
SA-MCQ	88	163.80
MA-MCQ	315	133.57
Easy Question	85	140.05
Moderate Question	109	139.82
Hard Question	209	140.40
ML Question	129	138.06
CV Question	54	139.44
NLP Question	220	141.59

Table 1: Statistics of ELAIPBENCH.

4 Experiments

In this section, we evaluate the performance of frontier LLMs on ELAIPBENCH and identify five key findings: (1) Large reasoning models (LRMs) with an explicit thinking mode often underperform their base-model counterparts—a phenomenon we term *reasoning paralysis* (Section 4.2). (2) Over half of *reasoning paralysis* cases stem from harmful verification, in which models overturn initially correct answers during the reasoning process (Section 4.3). (3) Longer reasoning chains do not necessarily improve accuracy; instead, excessively long chains often lead to incorrect results (Section 4.4). (4) Models fail to adapt their reasoning depth to the difficulty of a question, producing chains of similar length regardless of its complexity (Section 4.5). (5) Retrieval-augmented generation (RAG) yields marginal gains at best, as retrievers struggle to find relevant evidence in academic papers while models fail to sufficiently integrate the retrieved content (Section 4.6). These findings underscore the challenges ELAIPBENCH poses for academic context understanding in LLMs.

4.1 Baseline LLMs

We evaluate seven base models and seven LRMs. The base models include Llama3.3-70B-Instruct (Dubey et al., 2024), Qwen3-8B, Qwen3-

Paradigm	Models	Total	Question Type		Difficulty			Discipline		
			SA-MCQ	MA-MCQ	Easy	Moderate	Hard	ML	CV	NLP
Base Models	Llama3.3-70B-Instruct	38.71	43.18	37.46	80.00	**79.82**	0.48	36.43	**38.89**	42.93
	Qwen3-8B	35.48	23.86	38.73	69.41	77.06	0.00	31.01	35.19	40.98
	Qwen3-14B	36.23	28.41	38.41	77.65	73.39	0.00	32.56	35.19	41.46
	Qwen3-32B	37.72	31.82	39.37	82.35	75.23	0.48	33.33	**38.89**	42.93
	DeepSeek-V3	**39.95**	32.95	**41.90**	**85.88**	78.90	**0.96**	**37.21**	**38.89**	**44.88**
	GPT-4o-0806	37.47	22.73	41.59	80.00	75.23	0.48	34.11	37.04	42.44
	GPT-5	38.71	31.82	40.63	**85.88**	74.31	**0.96**	36.43	**38.89**	42.93
CoTs	Llama3.3-70B-Instruct + CoT	25.81	25.00	26.03	80.00	32.11	0.48	24.81	24.07	28.78
	Qwen3-8B + CoT	35.98	25.00	39.05	72.94	76.15	0.00	30.23	35.19	42.44
	Qwen3-14B + CoT	31.76	19.32	35.24	77.65	56.88	0.00	31.01	31.48	34.63
	Qwen3-32B + CoT	32.75	23.86	35.24	82.35	56.88	0.00	32.56	33.33	35.12
	DeepSeek-V3 + CoT	29.78	32.95	28.89	84.71	42.20	**0.96**	27.13	29.63	33.66
	GPT-4o-0806 + CoT	25.31	12.50	22.54	78.82	31.19	0.48	22.48	25.93	28.78
	GPT-5 + CoT	35.24	22.73	38.73	**85.88**	62.39	0.48	31.78	35.19	40.00
LRMs	QwQ-32B	9.68	12.50	8.89	31.76	11.01	0.00	9.30	9.26	10.73
	Qwen3-235B-A22B-thinking	19.35	20.45	19.05	57.65	22.94	0.48	17.83	18.52	21.95
	DeepSeek-R1	25.81	31.82	24.13	61.18	45.87	**0.96**	23.26	25.93	29.27
	GPT-5-thinking-all	37.22	**46.59**	34.60	83.53	70.64	**0.96**	**37.21**	**38.89**	39.51
	GPT-o1-mini	35.48	28.41	37.46	78.82	67.89	**0.96**	34.11	35.19	39.02
	Gemini-2.5-flash-thinking	19.35	18.18	19.68	60.00	23.85	0.48	17.05	20.37	21.95
	Claude-3.7-Sonnet-thinking	38.46	43.18	37.14	**85.88**	73.39	**0.96**	35.66	**38.89**	42.93
Human	-	48.14	56.82	45.71	100.00	100.0	0.00	45.74	51.85	52.20

Table 2: Accuracy of LLMs on the ELAIPBENCH. **Bold** indicates the best result among LLMs. The **Human** results were given by the *Answer Verifier* during the annotation process.

14B, Qwen3-32B (Yang et al., 2025), DeepSeek-V3 (Liu et al., 2024), GPT-4o-0806 (Hurst et al., 2024), and GPT-5[2]. The LRMs comprise QwQ-32B (Wake et al., 2024), Qwen3-235B-A22B-thinking (Yang et al., 2025), DeepSeek-R1 (Liu et al., 2024), GPT-5-thinking-all, GPT-o1-mini (Jaech et al., 2024), Gemini-2.5-flash-thinking (Comanici et al., 2025), and Claude-3.7-Sonnet-thinking[3]. To reduce variability introduced by stochastic decoding, we report the mean results from 3 independent API calls. Evaluation metric, prompts and detailed parameter configurations of all LLMs are provided in Appendix A.4.

4.2 Results for LLMs

We employ three primary paradigms: a base model paradigm, a chain-of-thought (CoT) paradigm, and a LRM paradigm. In the base model paradigm, models are instructed to output only the final answer option. In the CoT paradigm, we add the prompt *"Let's think step by step."* to elicit stepwise reasoning before the final answer. In the LRM paradigm, the model is allowed to produce an unstructured reasoning chain before the answer.

As shown in Table 2, the overall performance of evaluated models on ELAIPBENCH is poor, with the highest accuracy reaching only 39.95%, highlighting their limitations in handling complex academic reasoning. Moreover, integrating reasoning strategies into base models yields inconsistent improvements: apart from Qwen3-8B, all other LLMs exhibit degraded performance with CoT prompting. This trend persists in specialized LRMs—GPT-5-thinking-all underperforms GPT-5, while DeepSeek-R1 trails DeepSeek-V3. This phenomenon, which we term *reasoning paralysis*, indicates that current reasoning paradigms are insufficient for ELAIPBENCH's academic challenges.

In both SA-MCQ and MA-MCQ tasks, LLMs consistently underperform human experts. In SA-MCQ, the LRMs demonstrate moderate improvement over the base model (GPT-5-thinking-all achieves 46.59%, surpassing GPT-5's 31.82%). However, in the MA-MCQ setting, the accuracy of LLMs drops significantly. We hypothesize that this decline is attributable to the increased complexity of the MA-MCQ format, which, by requiring the selection of multiple correct answers, makes it more challenging for LLMs to reliably identify all correct options. Furthermore, no LLM exceeds human expert performance on questions of easy or moderate difficulty. On hard questions, the marginal advantages exhibited by a few models (a mere 0.96%) are insufficient to warrant their reliable use by human experts for academic paper understanding. Finally, the consistent underperformance of LLMs

[2]www.openai.com/index/introducing-gpt-5
[3]www.anthropic.com/news/claude-3-7-sonnet

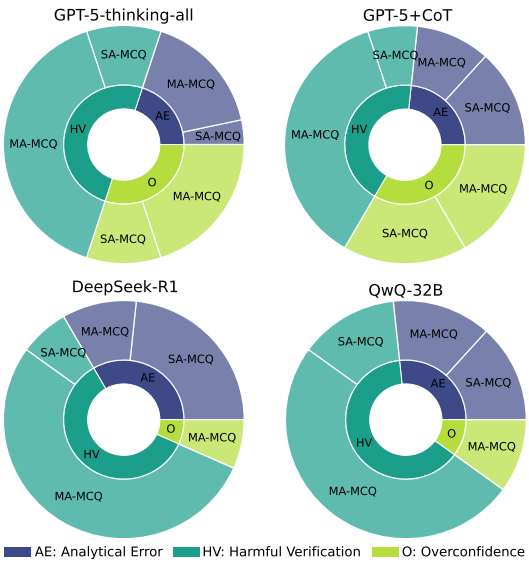

Figure 3: Reasoning error distribution of ELAIPBNECH on different LLMs.

compared to human experts across all AI-related disciplines indicates that their limitations in paper comprehension are generalizable.

4.3 Reasoning Error Analysis

To investigate the causes of *reasoning paralysis*, we randomly selected 30 incorrectly answered questions from the results produced under both the LRM and CoT paradigms and conducted a manual examination of the outputs. Based on our analysis of these errors, we classify the causes into three categories: (1) **Analytical Error**, where the model consistently performs incorrect reasoning to produce wrong answers; (2) **Harmful Verification**, where the model's flawed corrective measures during its verification process alter originally correct answers to incorrect ones; and (3) **Overconfidence**, where the model determines answers based on its own prior knowledge rather than the provided text. Figure 3 illustrates the distribution of these failure types. Notably, across all integrated reasoning settings, **Harmful Verification** accounts for over half of all error cases, indicating that the model often initially generates the correct answer but subsequently invalidates it through excessive re-analysis. This finding suggests that **Harmful Verification** is the primary reason for the underperformance of the reasoning paradigms compared to the base models. Furthermore, models exhibit a higher tendency toward *reasoning paralysis* in MA-MCQ tasks. Detailed examples of these three error types are provided in Appendix A.5.

4.4 Accuracy vs. Reasoning Length

Figure 4 illustrates the relationship between reasoning length and accuracy across different LLMs. Although QwQ-32B generates the longest reasoning chains, its accuracy does not improve with increased length—in fact, it achieves the lowest performance among the models. DeepSeek-R1 attains the highest accuracy among the three LRMs while maintaining a relatively short average reasoning chain, demonstrating superior reasoning efficiency. Nevertheless, its performance remains significantly lower than that of Qwen3-32B+CoT, suggesting that longer reasoning chains do not necessarily lead to better performance and may even be detrimental. We observe that self-verification behaviors occupy a substantial portion—approximately 50%—of the LRM's reasoning chain. These behaviors typically initiate with phrases such as *Wait* or *Alternatively*, and are followed by re-statements of the original question and subsequent restarts of the reasoning chain. Such repetitive patterns significantly lengthen the reasoning trajectory, introducing ineffective verification steps. Representative cases are provided in Appendix A.5.

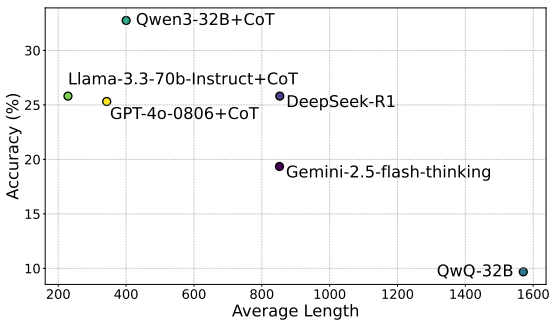

Figure 4: Completion Tokens vs.Performance.

4.5 Difficulty vs. Reasoning Length

Figure 5 presents the reasoning lengths of LLMs across different question types. We observe that all LRMs generate longer outputs than the CoT paradigm. However, across question categories—including SA-MCQ, MA-MCQ, easy, moderate, and hard difficulty levels—none of the baselines exhibit meaningful variation in reasoning length. This lack of adaptive behavior suggests that current LLMs fail to recognize the relationship between distractors in the options and the paper content, and consequently do not engage in deeper analysis. This shallow reasoning process often leads to incorrect inferences, which in turn

degrades final performance, resulting in reasoning-enhanced LLMs sometimes underperforming non-reasoning baselines. These findings underscore a critical limitation in current LLMs: while they can produce longer responses, they lack the ability to strategically adjust reasoning depth based on question complexity.

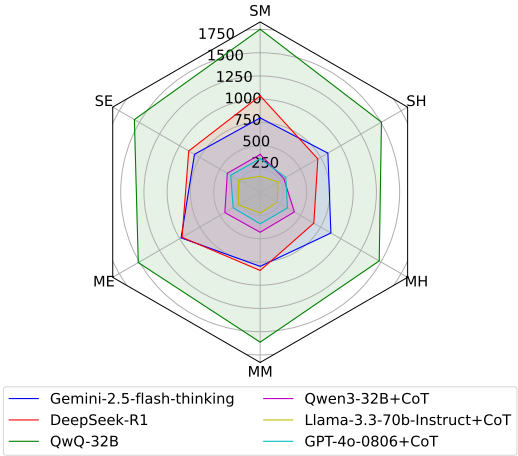

Figure 5: The reasoning lengths across different question types (SE: SA-MCQ + easy, SM: SA-MCQ + moderate, SH: SA-MCQ + hard, ME: MA-MCQ + easy, MM: MA-MCQ + moderate, MH: MA-MCQ + hard).

4.6 RAG Performance

In our RAG-based evaluation, we adopt two retrieval paradigms: intra-paper retrieval and web-based retrieval. For intra-paper retrieval, we segment each paper into non-overlapping passages of 512 tokens using sentence boundaries as delimiters. We then retrieve the top-five most relevant passages for each question using two methods: dense passage retrieval (DPR) based on the `BGE-m3` (Chen et al., 2024) encoder and the BM25 algorithm (Robertson et al., 2009). For web-based retrieval, we query the Google API using the question and paper title as input, retaining the content of the top-five most relevant web pages. In the prompt, we explicitly inform the LLM that the retrieved content constitutes auxiliary knowledge, which it may consult alongside the original paper. Detailed hyper-parameters, RAG configurations and prompts are provided in Appendix A.6.

Figure 6 presents the impact of RAG on answer accuracy across different LLMs. We observe that both `BGE-m3`-based DPR and BM25 intra-paper retrieval lead to performance degradation. This stems from the retrievers' inability to effectively align the questions with paper content, making it

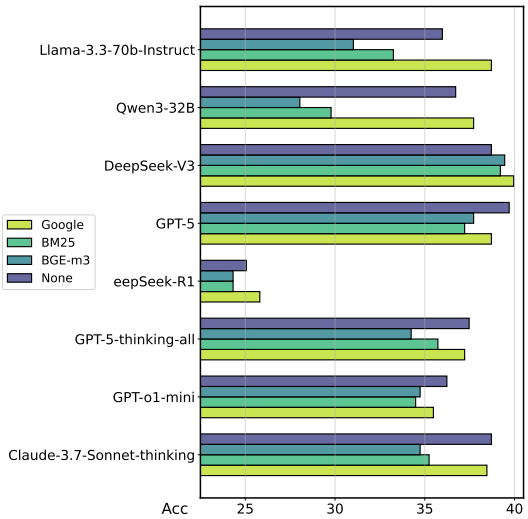

Figure 6: RAG Performance. The label **None** indicates that the LLM generates answers directly, and its accuracy corresponds to the **Total** reported in Table 1.

difficult to pinpoint evidence strongly correlated with the answer options. Within the web-based paradigm, improvements in answer accuracy remain marginal. Our analysis of the retrieved results reveals that nearly all returned documents are other relevant academic papers, compelling the LLM to perform multi-document QA while simultaneously analyzing complex scholarly content. Moreover, many models struggle to extract and synthesize information from noisy web sources, impeding coherent and contextually grounded reasoning. Although models such as GPT-5, GPT-5-thinking-all, and Claude-3.7-Sonnet-thinking show modest performance gains, these incremental improvements fall short of the expert-level academic understanding expected. We believe that a carefully engineered retriever, coupled with a structured organization of retrieved external knowledge, could enhance the capacity of LLMs for academic paper comprehension. Nevertheless, optimizing RAG system configurations lies beyond the scope of our study, and thus we defer such investigations to future work.

5 Conclusion

In this work, we introduce ELAIPBENCH, an expert-level benchmark designed to evaluate the understanding of AI research papers by LLMs. Our evaluation reveals that even state-of-the-art LLMs achieve only moderate performance (best at 39.95%), highlighting a substantial gap in current models' ability to perform fine-grained, evidence-grounded reasoning over complex scientific texts.

6 Limitations

We acknowledge several limitations of our work. (1) The scope of our benchmark is constrained in two ways: it is limited to the AI domain and covers only a single modality—text. However, expanding the benchmark is hindered by the practical challenges and high cost of annotation; for instance, recruiting expert annotators to collect the current 403 questions required six months and over 50,000 CNY (~7,000 USD). (2) The benchmark is English-only, which limits the evaluation of models' cross-lingual capabilities. In the future, we aim to expand the benchmark in two directions: including other scientific disciplines such as biology, physics, and chemistry, and incorporating multi-modality to enable a broader evaluation of scientific comprehension.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, and 1 others. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *Preprint*, arXiv:2412.15204.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Preprint*, arXiv:2502.14739.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *Preprint*, arXiv:2409.12186.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Min-Woo Kim, Hyo-Bin Park, Hee-Jin Ahn, Woo-Ram Park, Jae-Wan Jeon, Kyong-Ha Lee, Ryong Lee, and Dong-Geol Choi. 2025. Autopaperbench: An mllm-based framework for automatic generation of paper understanding evaluation benchmarks. *Electronics*, 14(6):1175.

Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *Preprint*, arXiv:2312.07559.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *Preprint*, arXiv:2407.10362.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pages 19036–19052. PMLR.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *Preprint*, arXiv:2409.13740.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Alan Wake, Bei Chen, CX Lv, Chao Li, Chengen Huang, Chenglin Cai, Chujie Zheng, Daniel Cooper, Fan Zhou, Feng Hu, and 1 others. 2024. Yi-lightning technical report. *arXiv preprint arXiv:2412.01253*.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *Preprint*, arXiv:2405.09939.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. SCIBENCH: Evaluating college-level scientific problem-solving abilities of large language models. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark.

Geemi P Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. 2025. Chemlit-qa: a human evaluated dataset for chemistry rag tasks. *Machine Learning: Science and Technology*, 6(2):020601.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jing Yu, Yuqi Tang, Kehua Feng, Mingyang Rao, Lei Liang, Zhiqiang Zhang, Mengshu Sun, Wen Zhang, Qiang Zhang, Keyan Ding, and 1 others. 2025. Scicueval: A comprehensive dataset for evaluating scientific context understanding in large language models. *Preprint*, arXiv:2505.15094.

A Appendix

A.1 Annotation Guidelines

All annotators utilized the annotation platform anonymously, ensuring their identities remained concealed from one another and thereby eliminating the possibility of collusion for reward acquisition. Different annotators are provided with distinct guidelines based on their assigned tasks. The content of these guidelines is shown below:

Guideline for Question Writer

Overview Thank you for participating in this annotation project. Your task is to read the assigned academic paper carefully and write well-designed, challenging questions based on its content. These questions will be used to evaluate the academic reasoning capabilities of large language models (LLMs) in long-context settings. Your contribution is highly valuable to the advancement of AI benchmarks. We ask you to approach this task with care and precision, as your work directly supports the development of more accurate and responsible AI systems.

General Guidelines

- Prefer documents between 8,192 and 128,000 words.
- Read the paper thoroughly before writing questions.
- Questions must not be answerable by simple keyword matching or shallow retrieval.

Question Design Requirements

- Avoid selecting overly well-known or commonly cited academic papers to ensure diversity and reduce the risk of models relying on memorized knowledge.
- Questions must be objectively formulated in English, free of subjectivity or ambiguity.
- The correct answer must not be obtainable through simple keyword matching or direct retrieval from the text; instead, answering should require reasoning, summarization, or integration of multiple pieces of information across the paper.
- To prevent confusion or unintended ambiguity, questions and options must not introduce domain-specific terms or abbreviations unless they are explicitly defined within the source paper.
- Distractor selections should be carefully designed to include misleading phrasing or partial truths—statements that align with some aspects of the paper but are ultimately incorrect—to increase cognitive demand and necessitate careful comparison and deep reasoning.
- For SA-MCQ, exactly one option must be correct.
- For MA-MCQ, either two or three options must be correct, ensuring a balanced level of difficulty and discouraging random guessing.

Reward Rules Annotators will receive **30 CNY** for each well-written and verified question. However, the reward may be reduced under the following conditions:

- If the question fails to pass evidence verification.
- If the question is too easy and can be answered correctly without reasoning.
- If the question or options violate any of the design guidelines above.

If your question fails the evidence verification stage, you will be required to revise and resubmit it based on the verifier's feedback. In such cases, your reward will be reduced accordingly.

Final Note We sincerely appreciate your effort and participation in this project. Your thoughtful contributions are essential to building more capable and trustworthy AI systems. Thank you for your cooperation!

Guideline for Evidence Verifier

Overview Thank you for participating in the evidence verification process. Your task is to carefully read the provided academic paper, the corresponding question, and the proposed evidence. Based on these inputs, you must attempt to answer the question as accurately as possible. This process is designed to evaluate whether the evidence is sufficient and relevant to support the question.

General Workflow

- Carefully read the assigned paper segment, the question, and the associated evidence.
- Try to answer the question solely using the provided paper content and evidence.
- Submit your answer through the platform interface within the given time limit.

Answer Judgment and Feedback

- If your answer is correct, you will receive a reward for this verification task.
- If your answer is incorrect, a penalty will be applied by deducting part of the reward.
- Incorrect answers indicate that the evidence may not sufficiently support the question.
- In such cases, the corresponding question will be sent back to the original question author for revision.

Guidelines for Verification

- Base your judgment strictly on the provided evidence and the content of the paper.
- Do not incorporate any external knowledge or assumptions.
- Make sure to complete the task within the time limit.
- Be precise in your answer; partial correctness may still lead to rejection.

Reward Rules

- You will receive a base reward for each correctly verified question.
- If your answer is incorrect:
  - A partial deduction will be made from your total reward for this task.
  - The question will be marked as needing revision and returned to the question annotator.

Final Note Your participation is essential to ensuring the quality and reliability of our benchmark. Accurate evidence verification directly enhances the overall dataset quality and model evaluation fairness. We appreciate your careful attention and thoughtful work.

Guideline for Answer Verifier

Overview Thank you for participating in the answer verification process. Your task is to answer a question based on a given academic paper, simulating the perspective of a well-informed reader. Your answer quality and response time will be used to assess the difficulty level of the question and to determine the reward distribution for both you and the original question author.

General Workflow

- Read the provided paper and the associated question carefully.
- Answer the question using only the paper content; external knowledge is not allowed.
- Submit your answer within the interface, where a timer will be running.

Timing and Reward Mechanism

- A timer starts when the question is displayed.
- If you submit a correct answer within the expected time window, you will receive the maximum reward.
- If you exceed the time limit, a partial penalty will be applied.
- If your answer is incorrect, a more significant penalty will be imposed.

Impact on Question Difficulty

- Your performance is used to estimate the difficulty of the question:
  - Quick and accurate answers suggest the question is easy.
  - Slow or incorrect responses indicate higher question difficulty.
- These estimates will influence quality control, benchmarking statistics, and bonus calculations for the original question writer.

Answering Guidelines

- Base your answer strictly on the information provided in the paper.
- Avoid speculative reasoning or assumptions beyond the document.
- Be concise but precise—ambiguous answers may be penalized.

Final Note Your role is crucial in validating the reliability and effectiveness of each question. Accurate and timely answers not only help refine the question set but also contribute directly to fairer model evaluation benchmarks. We appreciate your rigorous participation and attention to detail.

A.2 Annotation Details

A.2.1 Annotator Statistics

To better understand the background and composition of our annotators, we collected information on their age, gender, and academic status during registration. Figure 7 summarizes the distribution across these three dimensions. In terms of age, the majority of annotators fall within the 22–26 (41%) and 26–30 (36%) age groups, while those under 22 (14%) and over 30 (9%) constitute smaller proportions. Regarding gender, 80% of annotators identify as male and 20% as female. All annotators hold or are currently pursuing a master's degree or higher: specifically, the cohort includes 8 master's students, 2 individuals with completed master's degrees, 8 PhD students, 1 postdoctoral researcher, and 1 professor. Furthermore, every annotator has authored at least one academic paper, and 55% have published more than five papers, underscoring the high level of academic expertise within the annotation team.

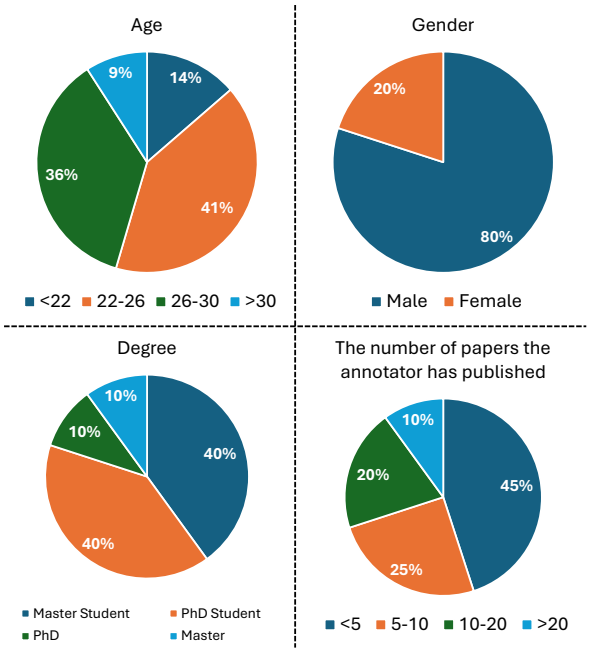

Figure 7: Distribution of our annotators across ages, genders, and academic status.

A.2.2 Annotation Platform

Our annotation platform is designed to support the construction pipeline for ELAIPBENCH. As shown in Figure 8, annotators begin by logging into the system, where they can upload papers and track their annotation progress, including the number of QA pairs submitted and the corresponding rewards. Once a paper is uploaded, *Question Writers* can browse its contents and start creating QA pairs along with the associated supporting evidence on the right side of the interface (Figure 9). The input interface allows *Question Writers* to write and submit their QA pairs in a structured format (Figure 10). After submission, they can view their previous submissions and accumulated earnings (Figure 11). Submitted QA pairs then undergo an evidence verification stage, where another *Evidence Verifier* is assigned to judge whether the provided evidence adequately supports the question. Annotators can check the verification status of each question (Figure 12). If the evidence is deemed insufficient or incorrect, the original *Question Writer* is required to revise the QA pair and resubmit it (Figure 13). The evidence verification stage (Figure 14) presents the verifier with the original paper, the question, and the proposed evidence. The verifier must analyze whether the evidence logically supports the question. If the verification is successful, the QA pair proceeds to the subsequent answer verification stage. In the answer verification stage (Figure 15), a designated *Answer Verifier* is tasked with answering the question based on the content of the provided paper. A timer runs during the answering process, and both the accuracy and response time are used to calculate the verifier's reward. The outcome also determines whether the original question writer receives a performance bonus.

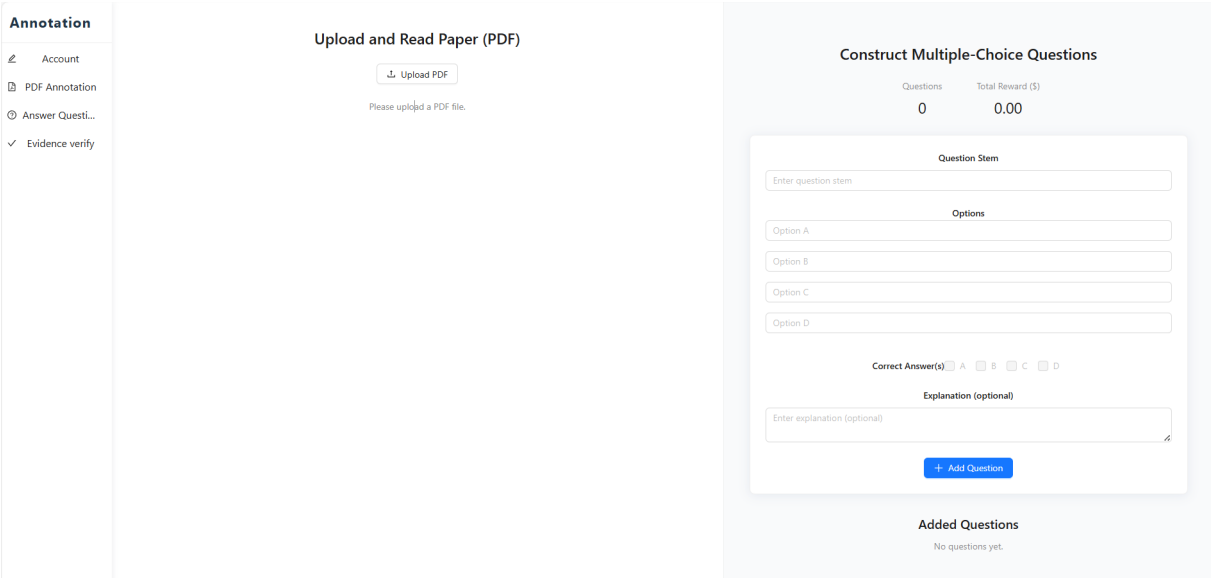

Figure 8: Screenshot of the QA pair annotation page. After logging in, annotators can upload papers and perform annotations on this page. They can also see the number of questions they have annotated and the amount of money they have earned.

Figure 9: Screenshot of the QA pair annotation page. After uploading the papers, *Question Writers* can browse the document and create QA pairs along with corresponding evidence on the right side of the page.

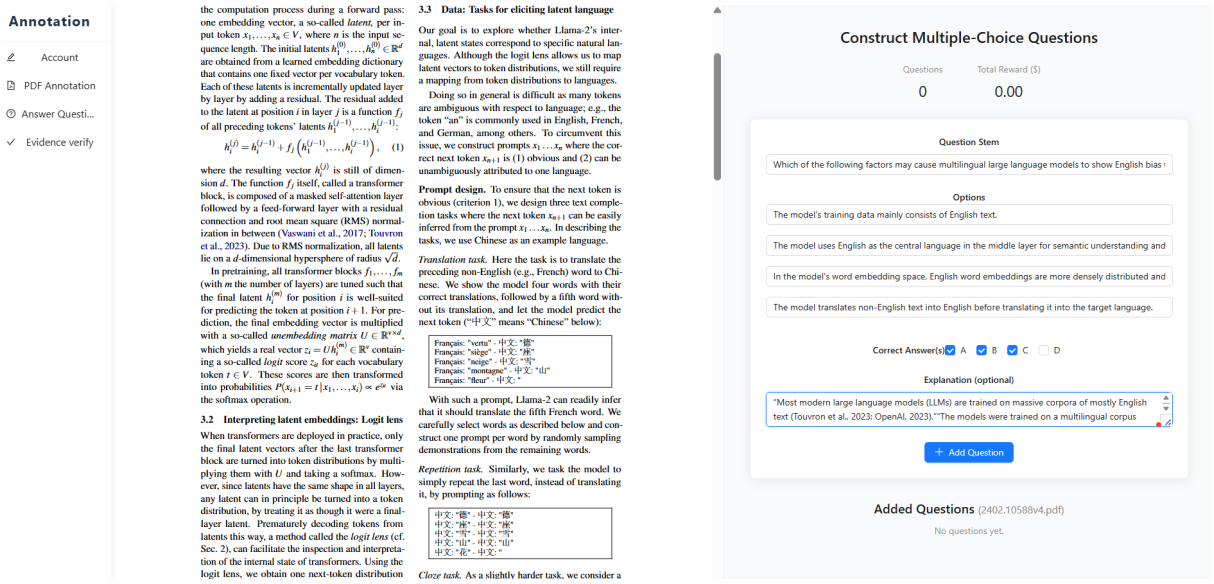

Figure 10: Screenshot of the QA pair annotation page. Annotators input the created QA pairs.

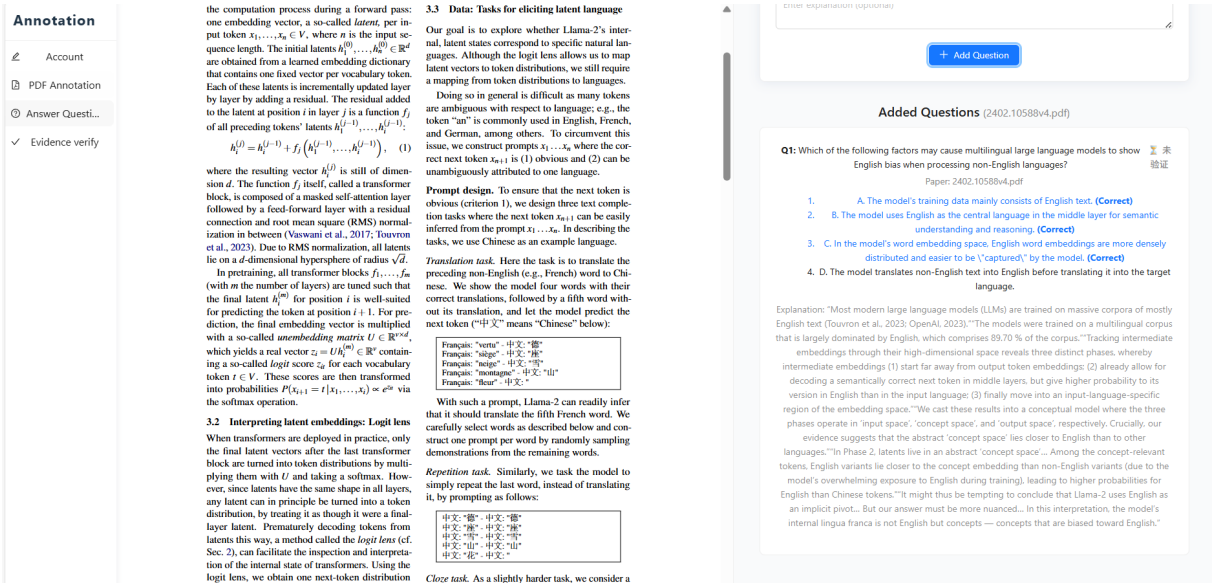

Figure 11: Screenshot of the QA pair annotation page. After annotating and submitting questions, annotators can view the questions they have submitted and the amount of money they have earned.

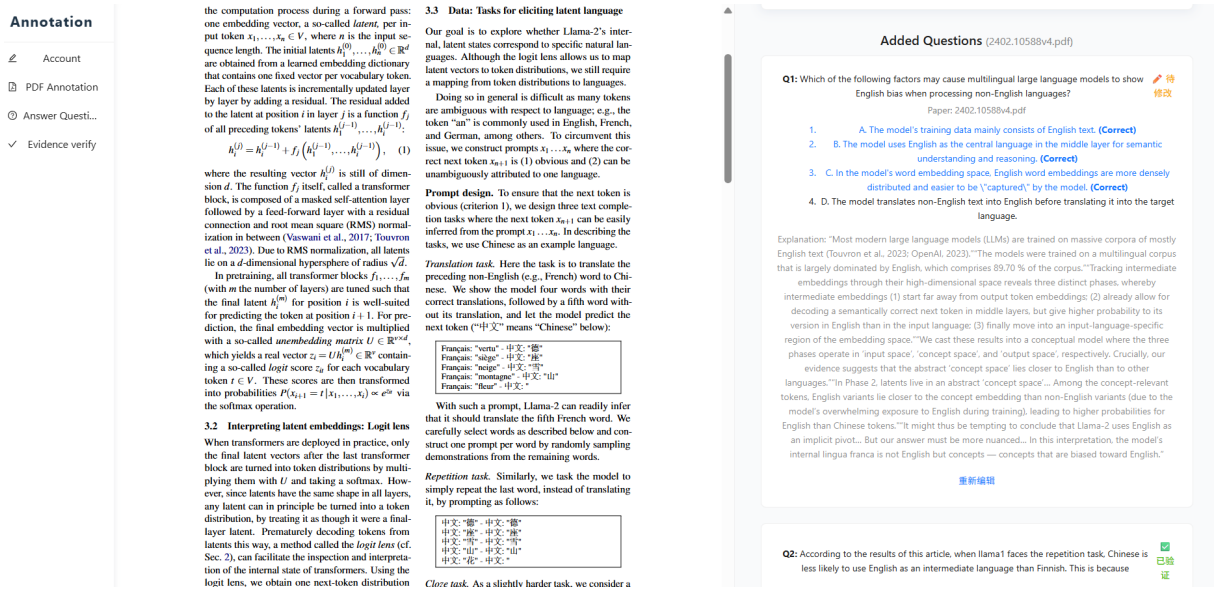

Figure 12: Screenshot of the QA pair annotation page. After verification of evidence, annotators can check the status of the questions submitted and revise those that did not pass the verification.

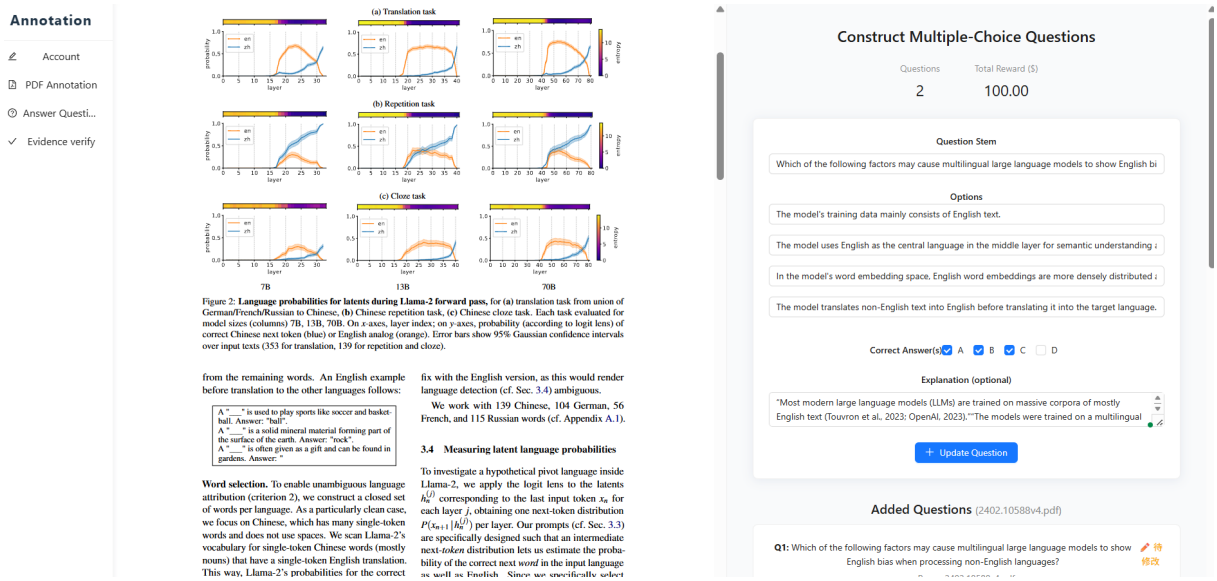

Figure 13: Screenshot of the QA pair annotation page. Annotators can revise the questions they have created and resubmit them.

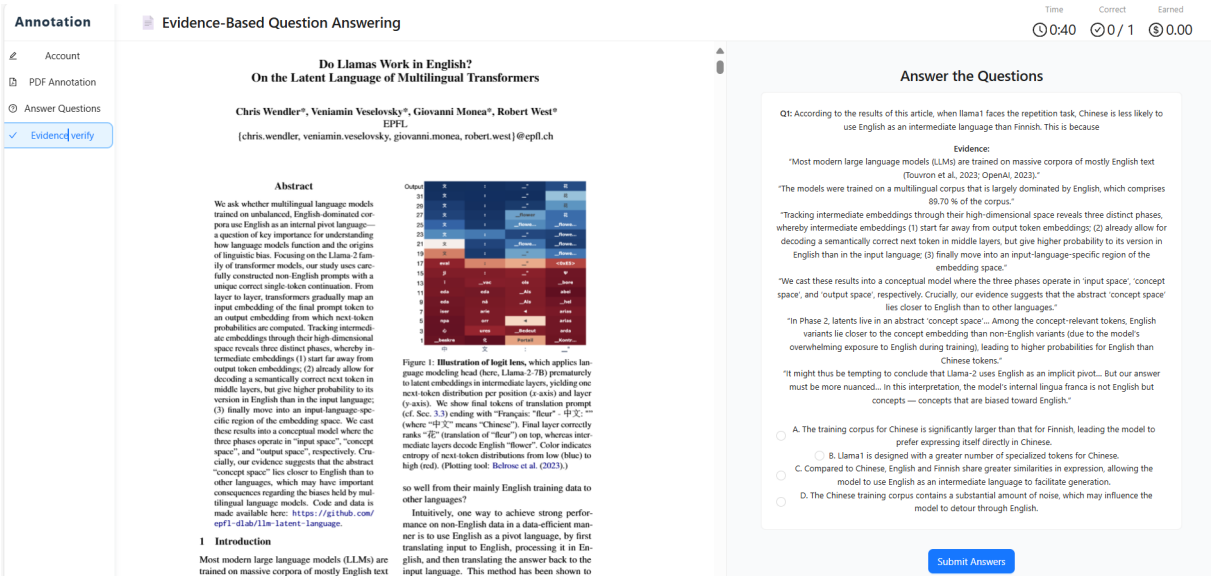

Figure 14: Screenshot of the evidence verification page. On this page, annotators are provided with a paper, a question, and the corresponding evidence for the question. They are then asked to judge whether the evidence supports the question. Rewards are given based on the correctness of their judgment.

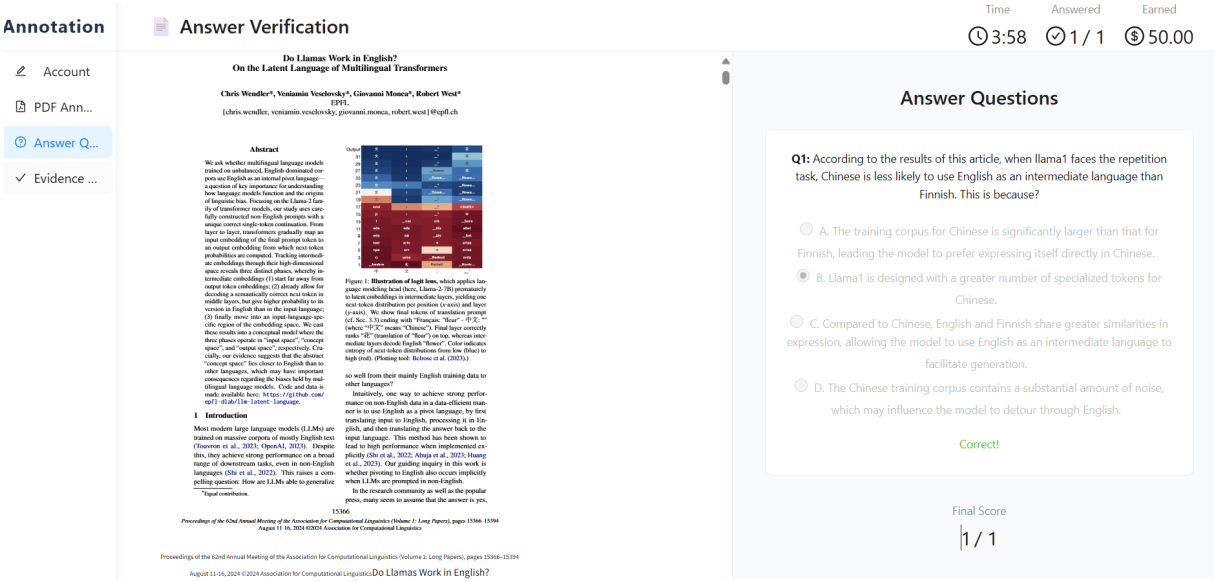

Figure 15: Screenshot of the answer verification page. *Answer Verifiers* are tasked with answering questions based on the content of the provided paper. During this process, a timer tracks the duration, and rewards are calculated for the *Answer Verifiers* based on both the accuracy of their answers and the response time.

A.3 Benchmark Statistics

In ELAIPBENCH, the answer options A, B, C, and D account for 26.45%, 26.10%, 24.20%, and 23.25% of all questions, respectively, indicating a relatively balanced distribution among the four choices. On average, each article corresponds to 0.64 SA-MCQs and 2.31 MA-MCQs, with 76% of the papers containing at least one question classified as **Hard** difficulty. Among all questions, the shortest contains 29 tokens and the longest 492 tokens; across all source papers, the shortest document spans 5,479 tokens, while the longest reaches 31,333 tokens. We visualize the word clouds of questions and papers from ELAIPBENCH in Figures 16 and 17. These figures indicate that the benchmark contains a diverse vocabulary for describing academic content.

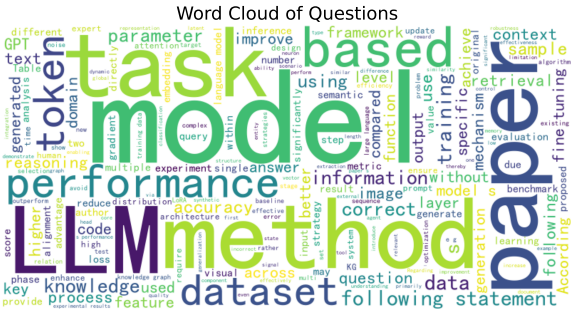

Figure 16: The word cloud visualization of questions in ELAIPBENCH.

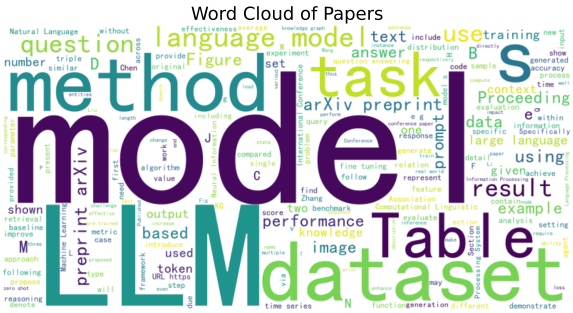

Figure 17: The word cloud visualization of papers in ELAIPBENCH.

Table 3 shows the statistics of the papers. Tables 4 through 9 present representative SA-MCQ and MA-MCQ examples from the three disciplines: ML, CV, and NLP.

A.4 Evaluation Metric and LLM Settings

Evaluation Metric. Our evaluation employs strict accuracy metrics: for MA-MCQ, responses must identify all correct options without errors; SA-MCQ requires exact match.

Discipline	Year		
	2021<	2021-2023	>2023
ML	11	9	15
CV	9	4	11
NLP	15	17	46

Table 3: Paper statistics.

Prompts. For the base models:

SA-MCQ Prompt

Please select the correct answer based on the paper content. Each question has only 1 correct option. Format your response as follows:
The correct answer is boxed {insert answer here}.
Examples:
The correct answer is boxed {A}
Paper: {paper}
Question: {question}

MA-MCQ Prompt

Please select all correct answers based on the paper content. Each question has 2-3 correct options. Format your response as follows:
The correct answer is boxed {insert answer here}.
Examples:
The correct answer is boxed {ACD}
Paper: {paper}
Question: {question}

For the CoT paradigm, we add *"Let's think step by step."* to the prompt. For the LRMs, the prompts are the same as the base models.

Parameter Settings. All evaluated LLMs feature context windows of at least 32k tokens, enabling comprehensive coverage of full papers and questions. In the evaluation setup for base models, we configure the generation sampling parameters with temperature=0.1 and max_new_tokens=32. For the base models with CoT setting, we employ temperature=0.1 and max_new_tokens=4096. The same parameter configuration is adopted for LRMs to

Paper Title	Do Llamas Work in English? On the Latent Language of Multilingual Transformers
Question	According to the results of this article, when llama1 faces the repetition task, Chinese is less likely to use English as an intermediate language than Finnish. This is because: A. The training corpus for Chinese is significantly larger than that for Finnish, leading the model to prefer expressing itself directly in Chinese. B. Llama1 is designed with a greater number of specialized tokens for Chinese. C. Compared to Chinese, English and Finnish share greater similarities in expression, allowing the model to use English as an intermediate language to facilitate generation. D. The Chinese training corpus contains a substantial amount of noise, which may influence the model to detour through English.
Answer	B

Table 4: ML Question (SA-MCQ).

Paper Title	Machine Unlearning Fails to Remove Data Poisoning Attacks
Question	Regarding the two main failure reason hypotheses proposed by the authors and how they were validated experimentally, which of the following descriptions is accurate and incisive? A. Hypothesis 1: Large Displacement refers to the model parameter movement distance required for unloading poisoned samples being significantly greater than that required for unloading an equivalent number of randomly selected clean samples. B. To validate Hypothesis 2: Orthogonal Subspace, the authors calculated the cosine similarity between the expected offloading direction vector and the direction vector obtained from gradient descent with clean data in a linear regression task and found that this value is close to 1, indicating that the two directions are nearly parallel. C. When experimentally validating Hypothesis 1, the authors utilized a logistic regression model trained on CIFAR-10 features, as the solution to the convex optimization problem is unique, allowing for clear calculation and comparison of distances between different models. D. The deeper implication of Hypothesis 2: Orthogonal Subspace is that due to the poisoning effect and normal learning effect being mutually exclusive in the model parameter space, any offloading method that only uses good data to repair the model will ultimately be futile.
Answer	AC

Table 5: ML Question (MA-MCQ).

Paper Title	Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models
Question	Which option most accurately summarizes the strategies and findings of the paper regarding dataset construction, model evaluation, and the discussion of current LVLM limitations? A. The paper finds that although the BLEU-2 score for multi-figure captioning tasks significantly improved for the Qwen-VL-Chat model after fine-tuning on the ArXivCap dataset, GPT-4V demonstrated absolutely leading performance across all vision-to-text tasks, including single-figure, multi-figure, contextualized, and paper title generation, highlighting its powerful zero-shot generalization ability in scientific figure comprehension, far surpassing all open-source models. B. In the quality control phase of the ArXivCap dataset, to ensure high image quality, the research team not only excluded images with extreme aspect ratios but also filtered out images with the shortest edge shorter than 224 pixels and removed images with pixel numbers exceeding the decompression bombs threshold. These stringent filtering criteria, combined with text cleaning for LaTeX expressions, were key to the dataset's ultimate high quality. C. The paper conducts a detailed analysis of generated caption error types, where "contextual misinterpretation" is identified as the most prevalent issue for all models, typically stemming from the models' inability to effectively integrate visual information from images with contextual clues from text. "Oversimplification," on the other hand, indicates a tendency for models to generate overly generic descriptions, failing to capture specific details and complex semantics in scientific figures, which reflects the current LVLMs' limitations in deep semantic understanding. D. All of the above.
Answer	B

Table 6: CV Question (SA-MCQ).

Paper Title	SynTab-LLaVA: Enhancing Multimodal Table Understanding with Decoupled Synthesis
Question	Regarding the construction and characteristics of the SynTab-LLaVA multimodal table understanding synthesis method, which of the following statements are correct? A. The method decouples the synthesis process into two independent steps: table image rendering and question-answer pair generation, thereby improving efficiency and robustness. B. Compared with human annotation and traditional MLLM synthesis, the method is more cost-effective, constructing a large-scale dataset with only 200 US dollars. C. By processing table images generated by Doubao, the LLM can reduce hallucinations and improve the accuracy of Q&A generation. D. SynTab-LLaVA adopts a hybrid multi-resolution visual encoder, combining high-resolution and low-resolution image information to capture both local textual content and global structural relationships.
Answer	AD

Table 7: CV Question (MA-MCQ).

Paper Title	Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data
Question	Under the Domain-Specific Fine-Tuning framework, according to GPT-4's evaluation criteria, which model in the OPT series is the most sensitive to different Table-to-Text methods? A. OPT-1.3B B. OPT-2.7B C. OPT-6.7B D. OPT-13B
Answer	A

Table 8: NLP Question (SA-MCQ).

Paper Title	Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models
Question	Which approach most effectively mitigates the negative consequences of flawed information retrieval in RAG systems, as suggested by recent research? A. A primary approach involves exclusively refining the initial retrieval mechanism through techniques like advanced query expansion and document re-ranking to guarantee the relevance of the provided external knowledge. B. One compelling approach is the sophisticated integration of the language model's pre-existing internal knowledge with the retrieved external information, employing strategies to identify and resolve potential conflicts in a source-aware manner. C. A straightforward approach to enhance RAG system resilience focuses on increasing the sheer volume of retrieved documents, assuming that the probability of encountering correct information will proportionally rise with a larger context window. D. An alternative approach leverages a multi-stage process where the language model first generates potential answers based solely on its internal knowledge, and subsequently uses retrieved documents merely as a validation layer without actively integrating conflicting information.
Answer	BD

Table 9: NLP Question (MA-MCQ).

accommodate their extended reasoning capabilities. This standardized parameter setup ensures fair comparison across different model configurations while allowing sufficient generation length for complex reasoning tasks.

A.5 Error Analysis

Figure 18 presents a case of **Analytical Error**, in which the model commits an incorrect analysis at the outset and prematurely settles on a flawed conclusion. Figure 19 illustrates an instance of **Harmful Verification**, where the model expends a substantial number of tokens during the verification phase yet ultimately revises a correct answer into an incorrect one. Figure 20 demonstrates **Overconfidence**, in which the model disregards the specific context or arguments presented in the source paper and instead relies on its internal priors to directly select an option, bypassing contextual reasoning.

Paper Title: Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models

Question:
Which option most accurately summarizes the strategies and findings of the paper regarding dataset construction, model evaluation, and the discussion of current LVLM limitations?
A. The paper finds that although ...
B. In the quality control phase of the ArXivCap ...
C. The paper conducts a detailed analysis of generated...
D. All of the above....

Reasoning Chain:
<think>
Hmm, this is a complex query requiring careful analysis of a technical paper summary. The user asks which option best summarizes the paper's strategies and findings regarding dataset construction ...
Option C correctly identifies key error types from the manual evaluation...
The paper emphasizes contextual misinterpretation and oversimplification as major limitations, matching C's description. I should provide a concise justification highlighting these points without unnecessary details.
</think> \\box{C}

Correct Answer: B

Figure 18: An example of **Analytical Error** from DeepSeek-R1, where the LLM commits to an incorrect answer from the outset.

A.6 RAG Settings

A.6.1 Hyper-parameter Setting

For the dense passage retrieval in the intra-paper setting, we employ the `BGE-m3` (Chen et al., 2024) encoder with its official checkpoint.[4] Each segmented passage is embedded into a 1024-dimensional vector space using the encoder's default configuration. During retrieval, we adopt a cosine similarity metric to measure relevance between a question embedding and candidate passage

[4] www.huggingface.co/BAAI/bge-m3

Paper Title: IRCAN: Mitigating Knowledge Conflicts in LLM Generation via Identifying and Reweighting Context-Aware Neurons

Question:
Based on the experimental section of the article, which of the following parameters are hyperparameters?
A. The neuron attribution scores threshold.
B. The number of identified context-aware neurons.
C. The activation value.
D. The enhancement strength.

Reasoning Chain:
<think> Okay, I need to figure out which of the options A-D are hyperparameters based on ... Looking at the options: A. Threshold (t) for neuron attribution scores: The paper mentions in Section 3.2 that they set t to 10%. **Since this is a parameter chosen by the user and not learned, it's a hyperparameter. But wait,** in the hyperparameter tuning part ... The threshold t was set to 10%, which is a fixed value, not part of the hyperparameter tuning. So maybe A is not a hyperparameter in this context because they fixed it at 10% and didn't tune it. **So A is a parameter set by them but not part of the hyperparameter search here. Hmm, tricky.** B. ... Therefore the correct answers are B and D.</think> \\box{BD}

Correct Answer: ABD

Figure 19: An example of **Harmful Verifications** from QwQ-32B, where the LLM first identifies the correct answer but then incorrectly rejects it during reasoning, ultimately selecting a wrong alternative.

Paper Title: Large Language Models Can Better Understand Knowledge Graphs Than We Thought

Question:
Which of the following statements about LLMs understanding different input formats of KGs are correct?
A. LLMs with larger parameters are more robust to ...
B. LLMs are better able to capture answer-related ...
C. Converting knowledge graphs to natural language ...
D. D. Different large language models have ...

Reasoning Chain:
<think>
I need to find relevant information in the provided PDF about how LLMs understand different input formats of knowledge graphs. Let me search for that information ...
B: This is supported by the attention level evaluation, which shows that LLMs consistently demonstrate a greater capacity to capture answer information from linearized triples compared ...
D: The text describes consistent LLM preference for unordered linearized triples across formats, **with no mention of model-specific preferences for triple organization** ...
</think> \\box{BC}

Correct Answer: BD

Figure 20: An example of **Overconfidence** from Claude-3.7-Sonnet-thinking, where the LLM fails to carefully review the paper, assumes that no content relevant to the options is present, and ultimately selects an incorrect choice.

embeddings. We set the `top_k` parameter to 5, returning the five highest-scoring passages per query.

For the BM25-based retrieval, we utilize the `rank_bm25` implementation with the default hyper-parameters $k_1 = 1.5$ and $b = 0.75$. Queries are tokenized using the same sentence boundary segmentation as in the DPR setup, ensuring consistency.

In the web-based retrieval setting, we interact with the Google Custom Search API, forming queries by concatenating the question and paper title. We restrict the API to return at most 10 results, from which we select the top-5 after relevance filtering. All retrieved HTML documents are converted to plain text using the `trafilatura` library and truncated to the first 4096 tokens for efficiency.

A.6.2 RAG Configuration

In both retrieval paradigms, retrieved passages or web documents are treated as external knowledge sources. Before being fed into the LLM, these snippets undergo the following preprocessing steps:

- **Text Normalization**: Remove excessive whitespace, HTML tags, and non-textual artifacts.
- **Sentence Trimming**: Ensure that each snippet begins and ends at complete sentence boundaries to preserve semantic integrity.
- **Token Budgeting**: Maintain a combined context length (paper + knowledge snippets) within 32,768 tokens to fit the LLM's input limits; in cases exceeding this limit, we perform a round-robin truncation over retrieved items.
- **Ordering**: Within each query, snippets are ordered in descending retrieval score before concatenation into the prompt.

A.6.3 RAG Prompts

In prompts, we explicitly instruct the LLM to treat this content as relevant context and to consult it when formulating its response. The prompt is structured as follows:

SA-MCQ Prompt

Please select the correct answer based on the content of the paper and the retrieved knowledge snippets provided below. Each question has only 1 correct option. Format your response as follows:
The correct answer is boxed {insert answer here}.
Examples:
The correct answer is boxed {A}
Paper: {paper}
Knowledge snippets: {retrieved knowledge}
Question: {question}

MA-MCQ Prompt

Please select all correct answers based on the content of the paper and the retrieved knowledge snippets provided below. Each question has 2-3 correct options. Format your response as follows:
The correct answer is boxed {insert answer here}.
Examples:
The correct answer is boxed {ACD}
Paper: {paper}
Knowledge snippets: {retrieved knowledge}
Question: {question}