# SCALAR: Scientific Citation-based Live Assessment of Long-context Academic Reasoning

**Renxi Wang**[1,2*] **Honglin Mu**[1,2*] **Liqun Ma**[1] **Lizhi Lin**[2,3] **Yunlong Feng**[4]
**Timothy Baldwin**[1,2,5] **Xudong Han**[1,2] **Haonan Li**[1,2†]
[1]MBZUAI [2]LibrAI [3]Tsinghua University
[4]Alibaba Group [5]The University of Melbourne

## Abstract

Long-context understanding has emerged as a critical capability for large language models (LLMs). However, evaluating this ability remains challenging. We present SCALAR, a benchmark designed to assess citation-grounded long-context reasoning in academic writing. SCALAR leverages academic papers and their citation structure to automatically generate high-quality ground-truth labels without human annotation. It features controllable difficulty levels and a dynamic updating mechanism that mitigates data contamination. The benchmark includes two tasks: a multiple-choice QA format and a cloze-style citation prediction. We evaluate a range of state-of-the-art LLMs and find that the multiple-choice task effectively distinguishes model capabilities. While human experts achieve over 90% accuracy, most models struggle. The cloze-style task is even more challenging, with no model exceeding 50% accuracy. SCALAR provides a domain-grounded, continuously updating framework for tracking progress in citation-based long-context understanding.[1]

## 1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in processing texts of increasing lengths (Achiam et al., 2023; Anthropic, 2024; Dubey et al., 2024; Yang et al., 2025). While capable of handling contexts of hundreds of thousands of tokens, evaluating their true understanding of long documents remains challenging.

Previous evaluations of long-context understanding have often relied on synthetic datasets or simple retrieval tasks like "needle in a haystack" variations (Kamradt, 2023; Kuratov et al., 2024b; Wang et al., 2024c; Roberts et al., 2024). While such tasks can test a model's ability to locate information in long sequences, they fail to assess genuine comprehension and are readily solvable by current LLMs (Team et al., 2024). Moreover, creating high-quality benchmarks traditionally requires extensive human annotation, which is both time-consuming and costly. Some work transforms short-context tasks into long context by combining them with passages or long documents, such as long-document QA (Kamradt, 2023), summarization (Chang et al., 2024), reasoning (Kuratov et al., 2024a) and reranking (Yen et al., 2024). However, such datasets suffer from two key issues: *data contamination* and *shortcut exploitation*, as LLMs can solve problems using their own knowledge rather than the long context. See more related work in Section 2.

In this work, we present SCALAR, a benchmark designed to evaluate large language models' (LLMs) citation-grounded long-context reasoning within the scientific domain (Figure 1. Our benchmark leverages recently published academic papers and their citations, which are implicitly annotated by domain experts through citation behavior, offering high-quality, unambiguous supervision without manual labeling. By focusing on papers accepted to top-tier venues (e.g., ICLR or ACL) and publicly available through arXiv, SCALAR ensures **data quality**, **transparency**, and **reproducibility**.

To ensure the benchmark remains relevant and minimizes data contamination from model pretraining corpora, we perform a detailed contamination analysis using a 4-year span of ICLR papers (Section 4). Our findings show that papers prior to 2023 are often memorized by models, while more recent ones (2024–2025) remain largely unseen, underscoring the importance of a live and continuously updated evaluation set. SCALAR is thus designed with a dynamic updating mechanism, allowing automatic integration of the latest high-quality publications.

SCALAR includes two complementary task for-

[*] Equal contributions.
[†] Corresponding author.
[1]Code and data are available at https://github.com/LibrAIResearch/scalar



*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 7817–7830

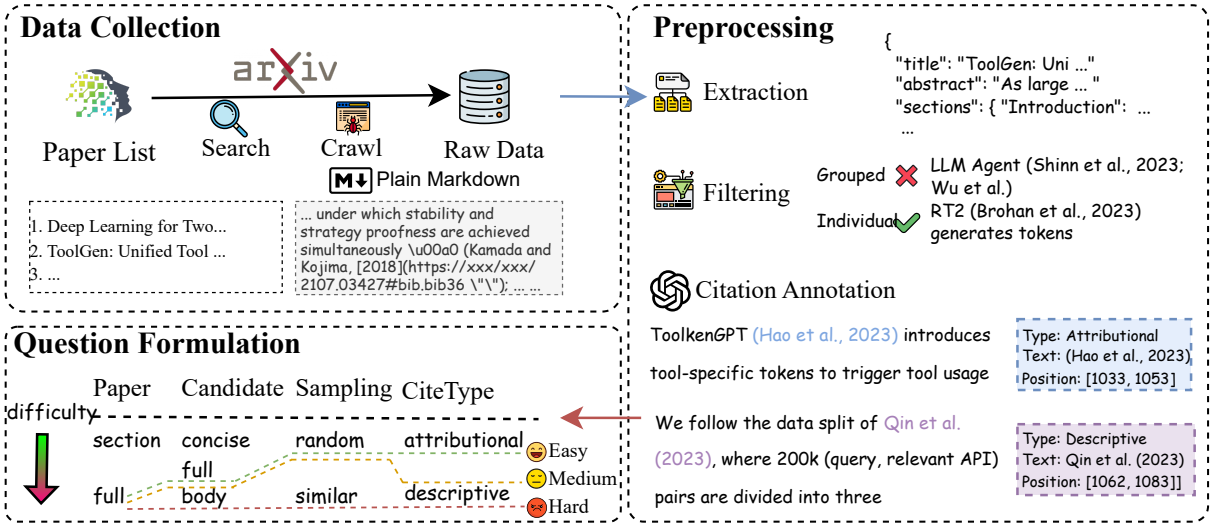


Figure 1: The overall process of building SCALAR. We start with a paper list sourced from arXiv but restrict to papers that have received high scores and been accepted at peer-reviewed conferences. We crawl the raw markdown data (top left). Then we parse them into structured data, sampling citations to mask and other citations as candidates (right). Finally, questions are formulated by masking citations in selected paper sections, choosing candidate spans, applying different sampling strategies, and assigning citation types, aligned with difficulty levels (bottom left).

mats. The first is a multiple-choice citation question answering (MCQA) task, which evaluates a model's ability to identify the correct cited paper from a set of plausible distractors. The second is a cloze-style citation prediction task, where the model must simultaneously predict four masked citations within a paper by selecting the appropriate references from a shared candidate list. Both tasks are designed to assess citation-grounded reasoning and long-context comprehension across entire scientific documents

To accommodate a wide range of model capabilities, we introduce a flexible difficulty control framework that adjusts both semantic complexity and context length. We define three levels of difficulty (Easy, Medium, and Hard) by systematically varying distractor construction and citation types. This design ensures the benchmark remains suitable for both small and large models.

Extensive experiments across state-of-the-art LLMs reveal substantial performance gaps, especially under longer contexts and harder reasoning settings. To validate task difficulty, we conduct human evaluation on the MCQA task, demonstrating that while humans perform near-perfectly, models still struggle, highlighting critical deficiencies in current long-context modeling.

## 2 Related Work

**Long-context Evaluation** Many approaches have been proposed to evaluate the ability of language models to utilize a longer context (Zhang et al., 2024b; Li et al., 2023; Dong et al., 2023; Wang et al., 2024b; Song et al., 2024). Real-world evaluations cover long-document QA and summarization (Shaham et al., 2023; Laban et al., 2024), mathematics and code understanding (An et al., 2024; Zhao et al., 2024; Wang et al., 2024a), domain specific analysis (Reddy et al., 2024), and retrieval tasks (Yen et al., 2024), in various languages (Qiu et al., 2024), formats (Zhang et al., 2024a). However, most of these benchmarks repurpose existing corpora, which introduces potential data contamination and limits their ability to measure genuine contextual reasoning.

Meanwhile, benchmarks using synthetic data focus on atomic abilities such as retrieval (Kamradt, 2023), state tracking (Kuratov et al., 2024a), data aggregation (Hsieh et al., 2024), multi-hop reasoning (Bai et al., 2023) like code understanding. Although such datasets provide controllable difficulty, they often diverge from realistic long-context usage.

**SCALAR complements these efforts** by focusing on *citation-grounded reasoning* within authentic scientific papers. Unlike prior benchmarks, it (i) derives supervision from expert-authored citation links instead of synthetic prompts, (ii) up-

| Benchmark | Timely Update | w/o Potential Issues | | Difficulty Control | Expert Created | Length | Domain |
|---|---|---|---|---|---|---|---|
| | | Contaminate | Shortcut | | | | |
| Needle-in-haystack (Kamradt, 2023) | ✗ | ✗ | ✗ | ✓ | ✗ | Static | N/A |
| BABILong (Kuratov et al., 2024a) | ✗ | ✓ | ✗ | ✓ | ✗ | Dynamic | Reasoning |
| RULER (Hsieh et al., 2024) | ✗ | ✓ | ✗ | ✓ | ✗ | Dynamic | General |
| LongBench (Bai et al., 2023) | ✗ | ✗ | ✗ | ✗ | ✗ | Dynamic | General |
| LongBench-v2 (Bai et al., 2025) | ✓ | ✗ | ✓ | ✓ | ✓ | Dynamic ($\leq$2M) | General |
| HELMET (Yen et al., 2024) | ✗ | ✓ | ✓ | ✓ | ✓ | Dynamic | Multi |
| SCALAR (ours) | ✓ | ✓ | ✓ | ✓ | ✓ | Dynamic ($\leq$1M) | Academic |

Table 1: Comparison of long-context benchmarks regarding update frequency, potential data contamination, shortcut susceptibility, difficulty control, expert involvement, length properties, and domain coverage.

dates dynamically to reduce data contamination, and (iii) provides *explicit difficulty control* through citation type, sampling strategy, and context length. This design bridges the controllability of synthetic datasets with the realism of document-based tasks, enabling a more faithful evaluation of long-context understanding in academic reasoning. For a detailed comparison, see Table 1.

**Citation-based Benchmarking** Although scholar literature corpus has been extensively used in language model pretraining (Lo et al., 2020; Gao et al., 2020), its potential to evaluate long context utilization is not fully explored. A number of datasets focus on generating and recommending citations (Funkquist et al., 2023; Färber and Jatowt, 2020; Gu et al., 2022). Ajith et al. (2024) creates a retrieval benchmark by constructing questions for inline citations using GPT-4 and manually creating questions. There are benchmarks testing models' abilities to answer questions based on papers. QASPER (Dasigi et al., 2021) focuses on answering questions about NLP papers, and LitQA (Lála et al., 2023) examines models' knowledge of biology works.

These benchmarks capture valuable aspects of academic understanding but are largely confined to local reasoning or single-document contexts. **SCALAR extends this line of work to the long-context setting** by treating citations as implicit supervision signals for multi-document reasoning. Rather than introducing new annotation pipelines or synthetic prompts, SCALAR systematically repurposes citation structures from recent peer-reviewed papers to assess whether models can link claims in one work to supporting evidence in others. This approach situates citation analysis within a realistic, scalable framework for measuring long-context comprehension.

## 3 *SCALAR*

This section describes our benchmark construction pipeline. We begin with data collection and preprocessing, including structural parsing and citation filtering. We then detail task formulation and introduce configurable dimensions for difficulty control—covering question scope, citation types, candidate sampling, and candidate representation. We define three difficulty levels, outline the final dataset, and conclude with a data contamination analysis to assess potential training overlap.

### 3.1 Data Collection

We begin by collecting research papers and their corresponding citations from arXiv,[2] a widely-used open-access repository for scientific papers across various domains. However, due to its open submission policy, it may include preliminary or non-peer-reviewed low-quality articles. To ensure the quality and reliability of our dataset, we filter for papers that have been accepted by a top-tier venue. More specifically, we select ICLR 2024 and ICLR 2025 accepted papers as original paper list, and search them in arXiv.[3] We crawl the markdown version as the raw data for further processing.

### 3.2 Preprocessing

**Structuralization** We then extract structured data from raw markdown texts. This includes identifying the title, abstract, and sections of the paper, cleaning links, and extract citations. Each citation is marked with its position and mapped to its corresponding reference in the bibliography. After that, we use the same data collection process to gather information about the cited papers. We separate papers into top-level sections, finding that papers in

[2] https://arxiv.org/

[3] Note that our methods can be generally applied to papers from other venues as well.

our dataset contain an average of 6.1 sections. For all papers, we remove references and appendices, leaving out the main content for task formulation.

**Citation Filtering** To ensure high-quality evaluation data, we distinguish between two categories of citations: grouped citations and individual citations. **Grouped Citations** refer to multiple sources together in a single parenthetical reference, typically used when summarizing general information, such as "*prior works (Liu et al., 2024; Hsieh et al., 2024a; Zhang et al., 2024)*". **Individual Citations**, in contrast, are used to refer to single source separately, usually when discussing specific methods or results, for instance "*needle in a haystack' test (Kamradt, 2023)*". Since grouped citations may have multiple correct answers when masked one of them, during the processing in Section 3.3, we exclusively sample and mask individual citations to maintain clear ground truth labels.

### 3.3 Task Formulation

We define two citation resolution tasks designed to evaluate a language model's ability to identify masked references in scientific papers: **Multiple-choice QA** (MCQA) and **Cloze Test**. Both tasks involve replacing in-text citations with placeholders and require the model to recover the correct references using the surrounding context. While Multiple-choice QA focuses on identifying a single masked citation from a set of options, the Cloze Test requires resolving multiple masked citations jointly.

For both tasks, each test case comprises three distinct parts:

- **Instruction** provides task-specific guidance, including how the LLM should complete the task, the expected answer format, and the roles of other components.
- **Question Paper** contains either the full text or a specific section of a paper, with one or more citations replaced by placeholders such as `**[MASKED_CITATION]**`.
- **Candidates** list four reference options. In the case of Multiple-choice QA, this includes one correct answer and three distractors.

To ensure high annotation quality, we draw all candidate references from the reference list of the question paper. This design leverages the expertise of the original authors, who have carefully selected relevant prior work for citation. By restricting candidates to this curated list, we avoid introducing external papers that, while potentially topically similar, may not align with the author's intent and could be more suitable for the citation than the actual ground truth.

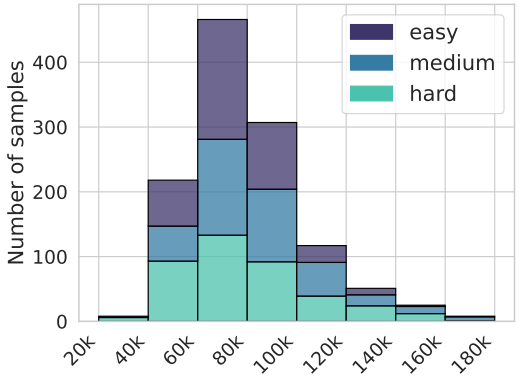


Figure 2: Token length distribution of the three subsets of our dataset. Samples are tokenized using GPT-4o tokenizer from tiktoken (OpenAI, 2023).

In our prompt implementation, we define several XML elements to separate different elements. The details of prompt templates are shown in Appendix C.

### 3.4 Difficulty Configuration Dimensions

We define four core configuration dimensions that govern each benchmark instance: question paper scope, citation type, distractor sampling strategy, and candidate representation. These dimensions allow fine-grained control over both semantic complexity and context length. A visual summary is shown in Appendix A.

**Question Paper Scope.** We use two modes for presenting the source paper to the model. In the Single Section setting, we provide only the section that contains the masked citation, thereby limiting contextual cues. In the Full Paper setting, the entire paper is included as context, offering broader information for reasoning.

**Citation Type.** We distinguish between two types of citations based on how they are presented in the text. **Attributional Citations** explicitly mention a specific model, method, or dataset (e.g., *BERT (Devlin et al., [2018]) is used for embedding texts...*). In contrast, **Descriptive Citations** embed the reference in a more narrative form without explicitly naming it (e.g., *Pretraining a bidirectional transformer (Devlin et al., [2018]) is time-*

*consuming...*).

**Candidate Sampling.** To construct challenging MCQA and cloze test options, we consider two candidate sampling strategies. In the Random Sampling setting, candidates are randomly drawn from the reference list of the question paper. In the Nearest Sampling setting, we sample four references that are cited at the same section. For MCQA, we then randomly mask one citation corresponding to one reference, and take other references as distractors. For cloze test, we mask four citations in the question paper, each corresponding to one of the references.

**Candidate Representation.** Each candidate reference can be presented in one of three formats. The Concise format includes only the title and abstract. The Full format includes the complete content of the cited paper. The Body format removes the title, abstract, introduction, and conclusion, leaving only the main body text to test the model's reasoning under limited cues.

### 3.5 Difficulty Levels and Final Dataset

Building on these dimensions, we define three standard difficulty levels to facilitate controlled benchmarking. These levels progressively increase in complexity by varying citation types, candidate selection strategies, and input lengths: (I) **Easy** level samples candidates randomly, and only mask attributional citations. (II) **Medium** level also samples candidates randomly while only masks descriptive citations. Both easy and medium level use the full paper for the question paper and candidates. (III) **Hard** level masks descriptive citations, but candidates are sampled from nearest references. Additionally, we use only the body of the paper for candidates, to avoid the model easily get answer from titles, abstracts, introductions, and conclusions.

The final dataset (per year) consists of 600 questions evenly distributed across two tasks and three difficulty levels, with each question containing four candidates. To ensure diversity, we limit at most five questions per paper. All papers, including question paper and candidates, are limited to 100,000 characters to accommodate model context limitations, with questions formatted using the template in Appendix C.

We collect data from papers published between 2022 and 2025; while data from 2022 and 2023 are primarily used for retrospective analysis, our benchmark focuses on 2024 and 2025 to emphasize its live and forward-looking nature. The length distribution is shown in Figure 2. Most samples have tokens between 60k and 80k. Hard set are slightly shorter than Easy and Medium set, since we only use the body content of the candidate papers. Figure 3 shows the submission or publication years of cited papers for specific venues. As shown in figure, most papers tend to cite more recent papers.

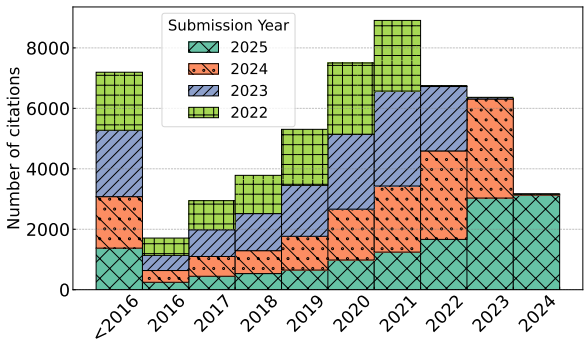


Figure 3: Distribution of cited papers' publication years across ICLR submission years used in SCALAR.

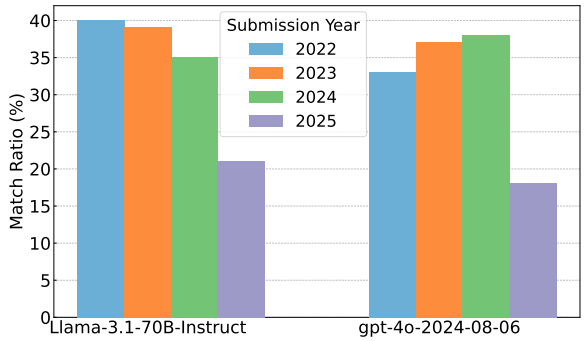


Figure 4: Data contamination prediction across submission years. We report the match rate (%), defined as the proportion of ground-truth citations successfully predicted by the model when prompted with the citation context. Higher match rates suggest a greater likelihood of data contamination.

## 4 Data Contamination Analysis

To evaluate potential data contamination, we conduct a targeted analysis using a subset of hard-level questions drawn from our benchmark. For each question, we input the full context from the question paper up to, but not including, the masked citation. We then prompt two language models to generate 20 tokens and assess whether the ground truth author name is explicitly predicted within the generated sequence.

The models we evaluate include *LLaMA-3.1-70B-Instruct* (training cutoff: December 2023) and

| Model | ICLR 2024 | | | ICLR 2025 | | | Average |
|---|---|---|---|---|---|---|---|
| | Easy | Medium | Hard | Easy | Medium | Hard | |
| Llama-3.1-8B | 0.21 | 0.22 | 0.23 | 0.30 | 0.28 | 0.24 | 0.25 |
| Llama-3.1-70B | 0.37 | 0.25 | 0.30 | 0.37 | 0.29 | 0.16 | 0.29 |
| Llama-3.3-70B | 0.43 | 0.28 | 0.29 | 0.37 | 0.36 | 0.23 | 0.33 |
| Qwen2.5-7B-1M | 0.31 | 0.34 | 0.22 | 0.49 | 0.38 | 0.28 | 0.34 |
| Phi-3-Mini | 0.19 | 0.21 | 0.26 | 0.39 | 0.21 | 0.20 | 0.24 |
| Phi-3-Medium | 0.12 | 0.17 | 0.12 | 0.20 | 0.12 | 0.08 | 0.14 |
| Claude-3.5-Haiku | 0.71 | 0.56 | 0.44 | 0.77 | 0.61 | 0.42 | 0.58 |
| GPT-4o-Mini | 0.75 | 0.51 | 0.38 | 0.81 | 0.56 | 0.48 | 0.58 |
| GPT-4o | 0.92 | 0.66 | 0.56 | 0.95 | 0.72 | 0.50 | 0.72 |
| QwQ-32B | 0.50 | 0.42 | 0.45 | 0.47 | 0.46 | 0.46 | 0.46 |
| DeepSeek-R1-0528 | 0.93 | 0.78 | 0.71 | 0.96 | 0.89 | 0.66 | 0.82 |
| Kimi-k2-Thinking | 0.95 | 0.87 | 0.82 | 0.99 | 0.91 | 0.76 | 0.88 |
| Gemini-2.5-Flash | 0.95 | 0.73 | 0.74 | 0.94 | 0.77 | 0.61 | 0.79 |
| Gemini-2.5-Pro | 0.98 | 0.92 | 0.93 | 0.99 | 0.92 | 0.85 | 0.93 |
| GPT-5-mini-medium | 0.99 | 0.85 | 0.90 | 0.99 | 0.92 | 0.82 | 0.91 |
| GPT-5-mini-high | 0.99 | 0.87 | 0.90 | 0.99 | 0.92 | 0.81 | 0.91 |
| GPT-5-high | 0.99 | 0.95 | 0.94 | 0.99 | 0.95 | 0.92 | 0.96 |
| o4-mini-medium | 0.98 | 0.88 | 0.81 | 0.99 | 0.88 | 0.72 | 0.88 |
| o4-mini-high | 0.97 | 0.87 | 0.86 | 0.99 | 0.88 | 0.74 | 0.89 |

Table 2: Multiple-choice question answering accuracy on ICLR 2024 and ICLR 2025 datasets, evaluated across three difficulty levels. Cell colors indicate performance from low (red) to high (green).

*GPT-4o* (training cutoff: June 2024). We record the proportion of questions where the correct author name is successfully predicted by each model.

Figure 4 illustrates the results across publication years. All models demonstrate a high success rate in predicting citations from papers in earlier years, with markedly reduced accuracy for papers published in 2025. This temporal trend strongly suggests that earlier papers are more likely to have been part of the models' training corpora. These findings highlight the persistent risk of data contamination in static benchmarks and reinforce the importance of using a live, dynamically updated benchmark to ensure fair and forward-looking evaluation.

## 5 Experiments

We benchmark a set of open-source and proprietary models with long-context capabilities ($\geq 128k$ tokens) on SCALAR.

**Open-source models** (1) *Llama-3.1-8B-Instruct*, *Llama-3.1-70B-Instruct*, and *Llama-3.3-70B-Instruct* are variants from Meta's Llama series (Touvron et al., 2023; Dubey et al., 2024), each supporting 128K-token contexts. (2) *Qwen2.5-7B-1M* (Yang et al., 2025) is a model from Alibaba with long context support. (3) *Phi-3-Mini-128K-Instruct* and *Phi-3-Medium-128K-Instruct* (Abdin et al., 2024) are lightweight models from Microsoft designed for efficiency, both supporting 128K tokens.

**Proprietary models** (4) *Claude-3.5-Haiku* (Claude, 2023) is Anthropic's high-speed long-context model with a 200K-token context window. (5) *GPT-4o-Mini* and *GPT-4o* (Achiam et al., 2023) are proprietary models from OpenAI, each with 128K-token context limits. GPT-4o represents OpenAI's most advanced model to date.

Recently, a new class of reasoning-focused models has emerged, such as DeepSeek-R1 (Guo et al., 2025) and others (Yu et al., 2025; Face, 2025), which emphasize reasoning capabilities through reinforcement learning. These models are designed specifically to enhance logical inference, multi-hop question answering, and mathematical reasoning. Although some of them were released within three months of our benchmark creation, we also include their performance for reference.

### 5.1 Main Results

**Multiple-choice QA** Table 2 presents model performance across difficulty levels. Generally, all

| Model | ICLR 2024 | | | ICLR 2025 | | | Average |
|---|---|---|---|---|---|---|---|
| | Easy | Medium | Hard | Easy | Medium | Hard | |
| Llama-3.1-8B | 0.24 | 0.24 | 0.25 | 0.24 | 0.20 | 0.20 | 0.23 |
| Llama-3.1-70B | 0.28 | 0.19 | 0.21 | 0.23 | 0.23 | 0.19 | 0.22 |
| Llama-3.3-70B | 0.27 | 0.18 | 0.22 | 0.22 | 0.22 | 0.18 | 0.22 |
| Qwen2.5-7B-1M | 0.30 | 0.21 | 0.24 | 0.25 | 0.28 | 0.24 | 0.25 |
| Phi-3-Mini | 0.20 | 0.19 | 0.16 | 0.18 | 0.16 | 0.21 | 0.18 |
| Phi-3-Medium | 0.14 | 0.11 | 0.12 | 0.12 | 0.12 | 0.13 | 0.12 |
| Claude-3.5-Haiku | 0.37 | 0.26 | 0.26 | 0.31 | 0.30 | 0.31 | 0.30 |
| GPT-4o-Mini | 0.35 | 0.22 | 0.28 | 0.35 | 0.28 | 0.28 | 0.29 |
| GPT-4o | 0.45 | 0.30 | 0.35 | 0.42 | 0.36 | 0.32 | 0.37 |
| QwQ-32B | 0.40 | 0.29 | 0.34 | 0.34 | 0.29 | 0.34 | 0.33 |
| DeepSeek-R1-0528 | 0.35 | 0.21 | 0.25 | 0.28 | 0.25 | 0.22 | 0.26 |
| Gemini-2.5-Flash | 0.42 | 0.37 | 0.38 | 0.35 | 0.44 | 0.40 | 0.39 |
| Gemini-2.5-Pro | 0.44 | 0.43 | 0.39 | 0.35 | 0.49 | 0.48 | 0.43 |
| GPT-5-mini-medium | 0.43 | 0.34 | 0.32 | 0.34 | 0.40 | 0.35 | 0.36 |
| GPT-5-mini-high | 0.43 | 0.35 | 0.32 | 0.35 | 0.41 | 0.37 | 0.37 |
| GPT-5-high | 0.44 | 0.34 | 0.38 | 0.35 | 0.43 | 0.42 | 0.39 |
| o4-mini-medium | 0.43 | 0.35 | 0.26 | 0.34 | 0.45 | 0.36 | 0.37 |
| o4-mini-high | 0.42 | 0.36 | 0.31 | 0.34 | 0.40 | 0.36 | 0.37 |

Table 3: Cloze-style citation prediction scores for non-reasoning and reasoning LLMs across two venues (ICLR 2024 and 2025). Cell colors indicate performance from low (red) to high (green).

LLMs' performance downgrades when difficulty increases. Reasoning LLMs and non-reasoning LLMs show a large gap in performance, indicating that our benchmark requires reasoning capability to finish.

For non-reasoning LLMs, on the easy level, *SCALAR* can already differentiate their long context capability, where the best model GPT-4o achieves 95% accuracy, while most models get a score lower than 60%, compared to the random baseline of 25%. For the hard level set, half of the models achieve random performance, and even current SOTA models obtain less than half the correct results, demonstrating how challenging our dataset is. The below-random performance of the Phi models is largely attributed to their failure to adhere to the expected output format, which leads to unsuccessful answer extraction. The hard level proves particularly challenging. Even state-of-the-art models like GPT-4o achieve only 50% accuracy, while most other models perform near random chance.

For reasoning LLMs, most of them achieved better performance than non-reasoning models. And the performance drop as the difficulty level increases is not as high as non-reasoning LLMs. The best model, GPT-5 with high reasoning effort achieved over 90% accuracy on all subsets of our MCQ benchmark.

| Annotator | Easy | Medium | Hard |
|---|---|---|---|
| Human 1 | 1.00 | 0.90 | 0.90 |
| Human 2 | 1.00 | 0.80 | 0.80 |

Table 4: Performance of human experts on the multiple-choice QA task. A total of 30 questions were evaluated, evenly drawn across three difficulty levels. Human experts achieved perfect or near-perfect accuracy.

**Cloze-style QA** As depicted in Table 3, the cloze-style test setting presents a greater difficulty to language models compared to multiple-choice formulations, primarily because they require the prediction of four citations at once. This increased challenge is evident in performance shifts: the accuracy of most open-source models degrades by approximately 10% relative to their multiple-choice QA performance, while proprietary models can see their scores reduced by as much as half. Overall, even the best-performing model achieves an average accuracy of less than 40%.

**Human Performance on MCQA** To better understand the task difficulty from a human perspective, we evaluated expert performance on the ICLR 2025 portion of our benchmark. Specifically, two

| Question | Candidate | Easy | | Hard | |
|---|---|---|---|---|---|
| | | GPT | Qwen | GPT | Qwen |
| Section | Concise | 0.86 | 0.88 | 0.58 | 0.48 |
| Full | Concise | 0.86 | 0.72 | 0.52 | 0.40 |
| Section | Full | 0.80 | 0.40 | 0.64 | 0.22 |
| Full | Full | 0.80 | 0.42 | 0.50 | 0.24 |

Table 5: Impact of context length on model performance. Results compare GPT-4o-mini and Qwen2.5-7B-Instruct-1M across different configurations. **Easy** level uses attributional citations with random candidates, while **Hard** level uses descriptive citations with nearest-neighbor candidates.

| Cite Type | Sampling | Candidate | GPT | Qwen |
|---|---|---|---|---|
| Attributional | Random | Full | 0.82 | 0.42 |
| Attributional | Nearest | Full | 0.82 | 0.38 |
| Descriptive | Random | Full | 0.56 | 0.34 |
| Descriptive | Nearest | Full | 0.52 | 0.22 |
| Descriptive | Nearest | Body | 0.44 | 0.28 |

Table 6: Impact of semantic reasoning difficulty on model performance. Results of GPT-4o-mini and Qwen2.5-7B-Instruct-1M.

PhD candidates in computer science were asked to complete a multiple-choice QA task consisting of 30 questions evenly sampled across the three predefined difficulty levels. As shown in Table 4, both annotators achieved perfect accuracy on the easy subset and high accuracy on the medium and hard subsets. These results demonstrate that while the task challenges current LLMs, it remains straightforward for human experts familiar with scientific content.

### 5.2 Error Analysis

We manually examined 72 incorrect predictions across models and difficulty levels. The analysis reveals several consistent failure patterns: (1) Over-reliance on semantic similarity. In over 70% of the incorrect cases, models selected papers sharing the same research field or similar methods, even when they were irrelevant to the actual citation. (2) Failure in citation-level reasoning: some cases misinterpret content near the citation location—over half of the incorrect predictions had irrelevant or weakly related citation context. (3) There are some model-specific trends. For example, LLaMA series frequently confused weakly related papers, even when the correct answer had distinct content. Also showed occasional format-following issues. GPT-4o, Claude, and Qwen are relatively more robust, but still failed when distractors shared close topical similarity or when keywords near the citation misled the model.

These findings suggest that while models can capture global topical alignment, they often struggle with fine-grained, context-sensitive citation understanding—underscoring the value of SCALAR in evaluating deep reasoning over long contexts.

### 5.3 Difficulty Analysis

Following our discussion of difficulty control in Section 3.5, where we established three difficulty levels, we now conduct a detailed analysis of how different configuration combinations affect task difficulty. We categorize these configurations into two primary dimensions: context length and semantic complexity. For each configuration setting in our analysis, we evaluate model performance on a sample of 50 questions.

**Context Length** We analyze the impact of context length by varying both question and candidate paper representations, as shown in Table 5. For questions, we compare using either the full paper or just the section containing the masked citation. For candidates, we compare using either the full paper or just the abstract. This creates four distinct length configurations while controlling for semantic difficulty.

The results show that both models generally perform better with more concise contexts. GPT-4o-mini maintains relatively stable performance across configurations, while Qwen2.5-7B shows significant degradation when given full candidate papers (dropping from 88% to 40-42% in the easy setting). This pattern persists in the hard setting, though with overall lower performance, suggesting that focused, relevant context may be more beneficial than comprehensive but potentially noisy full-paper information.

**Semantic Complexity** Table 6 demonstrates how different semantic factors affect model performance on citation prediction. We analyze three key dimensions: citation type (attributional vs. descriptive), candidate sampling method (random vs. nearest), and candidate representation (full paper vs. body only).

The results show a clear hierarchy of difficulty. Both models perform best with attributional citations, likely due to their more straightforward na-

ture. Performance drops significantly with descriptive citations, particularly when combined with nearest-neighbor sampling. While GPT-4o-mini maintains above-random performance across all configurations, Qwen's performance on the most challenging setting (descriptive, nearest sampling, full paper) drops to 22%, below random chance (25%). Interestingly, when using body-only candidates, Qwen's performance improves slightly to 28%, suggesting that the previous drop might be due to context length limitations rather than semantic difficulty alone. These patterns validate our benchmark's difficulty levels while highlighting the importance of considering both semantic and context length effects.

## 6 Conclusion

In this work, we introduced SCALAR, a novel benchmark designed to evaluate LLMs long-context understanding while mitigating data contamination. By leveraging citations from scholarly papers, we generate high-quality ground truth labels while controlling task difficulty. Our experiments with state-of-the-art LLMs reveal that while models can effectively handle simple citation matching, they struggle with deeper comprehension of complex, context-rich references. SCALAR offers a sustainable and evolving benchmark to track advancements in long-context processing, providing insights for future model development.

## Limitation

SCALAR currently focuses on citation-based QA and cloze-style matching tasks, which capture a narrow but meaningful dimension of long-context understanding, identifying and reasoning over citation evidence in scholarly texts. It does not evaluate broader academic reasoning skills such as summarization, synthesis, or argument tracking. SCALAR currently focuses on MCQA and cloze-style citation-matching tasks, which may not fully assess broader comprehension and reasoning abilities. Additionally, its scope is limited to computer science, restricting its applicability to other academic domains. To address these limitations, we plan to introduce diverse evaluation formats, while also expanding SCALAR to fields like biomedical and legal research for a more comprehensive assessment of long-context understanding.

## References


Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

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*.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *ICLR*.

Anthropic. 2024. Introducing the next generation of claude.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks.

Yushi Bai et al. 2023. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv:2308.14508*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Claude. 2023. Claude 2.1 model card. Technical report, Claude Inc.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv:2309.13345*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,

Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2023. Citebench: A benchmark for scientific citation text generation.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *Advances in Information Retrieval*, pages 274–288, Cham. Springer International Publishing.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models?

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. *Github*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024a. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024b. In search of needles in a 11m haystack: Recurrent memory finds what llms miss.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems.

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.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv:2311.04939*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2orc: The semantic scholar open research corpus.

OpenAI. 2023. tiktoken. `https://github.com/openai/tiktoken`.

Zexuan Qiu, Jingjing Li, Shijue Huang, Xiaoqi Jiao, Wanjun Zhong, and Irwin King. 2024. Clongeval: A chinese benchmark for evaluating long-context large language models.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Roberts, Kai Han, and Samuel Albanie. 2024. Needle threading: Can llms follow threads through near-million-scale haystacks?

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *EMNLP*.

Mingyang Song, Mao Zheng, and Xuan Luo. 2024. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang, Shan Dong, Yuhui Xu, Hanze Dong, Yalu Wang, Amrita Saha, Ee-Peng Lim, Caiming Xiong, and Doyen Sahoo. 2024a. Mathhay: An automated benchmark for long-context mathematical reasoning in llms.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024c. Needle in a multimodal haystack. In

*The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. Dapo: An open-source llm reinforcement learning system at scale.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024a. Marathon: A race through the realm of long context with large language models.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. $\infty$bench: Extending long context evaluation beyond 100k tokens. *arXiv:2402.13718*.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. Docmath-eval: Evaluating math reasoning capabilities of llms in understanding long and specialized documents.

## A Difficulty Control in SCALAR

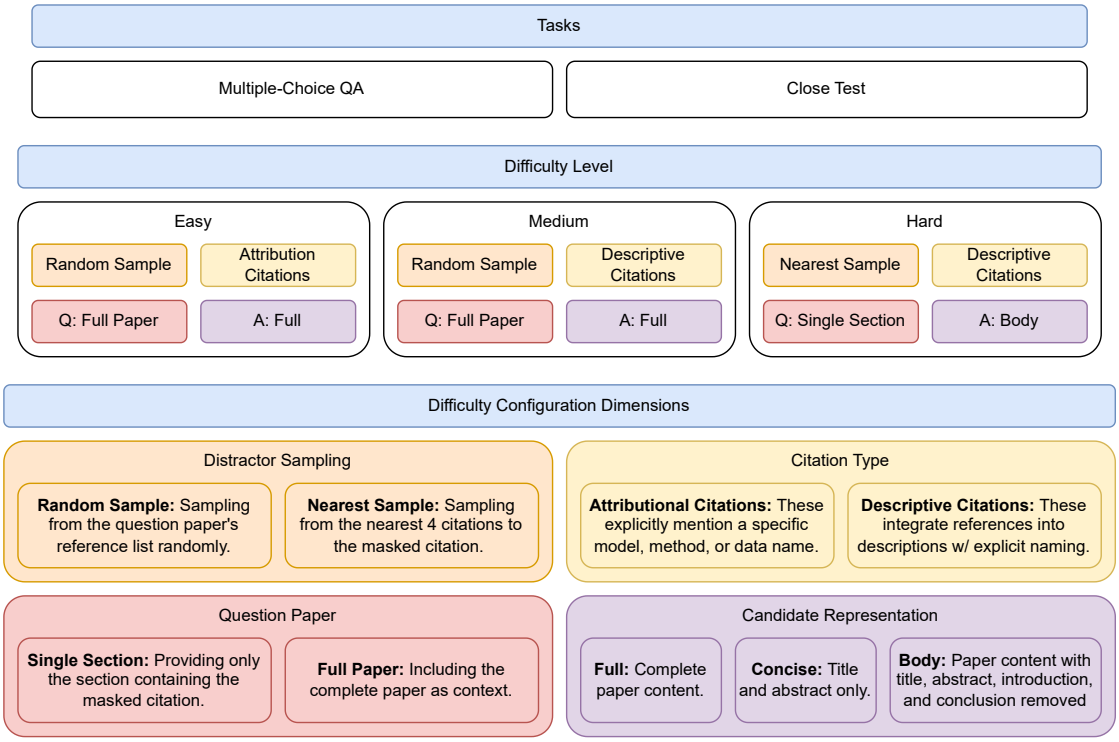


Figure 5: Overview of our configurable framework for difficulty control. The benchmark supports two tasks (Multiple-Choice QA and Cloze Test) and categorizes difficulty into three levels by varying distractor sampling strategies, citation types, question paper context, and candidate representations. These dimensions allow for fine-grained semantic and contextual adjustments, as illustrated in the diagram.

## B Model Context Length and Price

We list the context length and price of the models in this paper as in Table 7.

## C Prompt use in SCALAR

The prompt template used for the multiple choice QA and cloze test are in Figure 6 and Figure 7, respectively.

| Model | Context | Price |
|---|---|---|
| Llama-3.1-8B | 128K | $0.05* |
| Llama-3.1-70B | 128K | $0.65* |
| Llama-3.3-70B | 128K | $0.65* |
| Qwen2.5-7B-1M | 1M | $0.05* |
| QwQ-32B | 128K | $0.1* |
| Qwen3-8B | 128K | $0.035* |
| Phi-3-Mini-128K-Instruct | 128K | $0.03* |
| Phi-3-Medium-128K-Instruct | 128K | $0.10* |
| Claude-3.5-Haiku | 200K | $0.80 |
| GPT-4o-Mini | 128K | $0.15 |
| GPT-4o | 128K | $2.50 |

Table 7: Model context length and price per 1M tokens. * denotes price estimated via third-party inference service.

You are given a paper with a placeholder "**[MASKED_CITATION]**" in its content. Your task is to select the most appropriate reference from the provided reference list to replace the mask.
- The paper content is enclosed within <Paper> and </Paper>.
- The reference list is enclosed within <References> and </References>, with each reference wrapped in <Candidate> and </Candidate>.
- After selecting the best-suited reference, output the index of that reference in the following format: <answer>index</answer>.
<Paper>
... BERT (*[MASKED_CITATION]**) or ...
</Paper>
<References>
<Candidate>Candidate [0]:
... candidate content ...
</Candidate>
<Candidate>Candidate [1]:
... candidate content ...
</Candidate>
<Candidate>Candidate [2]:
... candidate content ...
</Candidate>
<Candidate>Candidate [3]:
... candidate content ...
</Candidate>
</References>
Remember to output the index of the selected reference enclosed within <answer> and </answer>.

Figure 6: The prompt template used for the multiple choice QA.

You are given a paper with four placeholders **[MASKED_CITATION_0]**, **[MASKED_CITATION_1]**, **[MASKED_CITATION_2]**, and **[MASKED_CITATION_3]**, each hiding a citation, plus exactly four reference candidates.
- The paper content is enclosed within <Paper> and </Paper>.
- The reference list is enclosed within <References> and </References>, with each reference wrapped in <Candidate> and </Candidate>.
- Return **one** <answer> tag with four separated integers giving the candidate index (0-3) for placeholders in the order of **[MASKED_CITATION_0]**, **[MASKED_CITATION_1]**, **[MASKED_CITATION_2]**, and **[MASKED_CITATION_3]**. For example: <answer>
2
1
3
0
</answer>.
<Paper>
... BERT (*[MASKED_CITATION_0]**) or ...
</Paper>
<References>
<Candidate>Candidate [0]:
... candidate content ...
</Candidate>
<Candidate>Candidate [1]:
... candidate content ...
</Candidate>
<Candidate>Candidate [2]:
... candidate content ...
</Candidate>
<Candidate>Candidate [3]:
... candidate content ...
</Candidate>
</References>
Remember: output four integers wrapped inside a single <answer> tag.

Figure 7: The prompt template used for Cloze-style test.