# Long-Context Reasoning Through Proxy-Based Chain-of-Thought Tuning

**Anonymous ACL submission**

## Abstract

Recent large language models support inputs of up to 10 million tokens, yet they perform poorly on long-context tasks that require complex reasoning. Such tasks can be solved using only a subset of the input — a proxy context — rather than the full sequence. Despite sharing the same underlying reasoning process, models exhibit a significant performance disparity between proxy and full contexts. To improve long-context reasoning, we propose ProxyCoT, a novel training framework that transfers reasoning capabilities from short proxy contexts to full long contexts. Specifically, we first obtain high-quality chain-of-thought reasoning traces on proxy contexts through reinforcement learning or distillation from a larger teacher model, and then ground the generated traces in full long contexts with supervised fine-tuning. Experiments across different datasets demonstrate that ProxyCoT consistently outperforms strong baselines with reduced computational overhead. Furthermore, models trained with ProxyCoT generalize their long-context reasoning capabilities to out-of-domain tasks. [1]

## 1 Introduction

Large language models (LLMs) have been developed with increasingly expansive context windows, now reaching lengths of up to 10 million tokens (Gemini Team, 2025; Llama Team, 2025; Yang et al., 2025c). These models promise advancements for long-context tasks that demand complex reasoning, such as synthesizing insights from multiple medical reports or addressing analytical questions spanning several financial documents. Successfully performing these tasks requires LLMs not only to locate relevant information within extensive inputs but also to reason effectively over the extracted knowledge to produce correct responses.

To enhance the reasoning capabilities of LLMs, prior work has primarily relied on chain-of-thought distillation (Li et al., 2023; Ho et al., 2023) and reinforcement learning (DeepSeek-AI et al., 2025) to elicit visible, step-by-step reasoning traces. These approaches have been successful on short-context tasks, but exhibit notable limitations when applied to long-context settings. Chain-of-thought distillation, for instance, depends on the high-quality reasoning traces from a teacher model, which is typically large and therefore slow and costly to query (Li et al., 2023; Ho et al., 2023; DeepSeek-AI et al., 2025).[2] Moreover, even strong teacher models may produce unreliable traces on complex long-context tasks. As an illustration, on SciTrek, a recently released long-context question answering benchmark over (full-text) scientific articles, the best performing open-source model only achieves 48.8% exact match (Li et al., 2025a). Reinforcement learning with policy-gradient methods, often used when no suitable teacher is available, is also challenging to scale to long contexts because it requires extensive sampling, making training slow

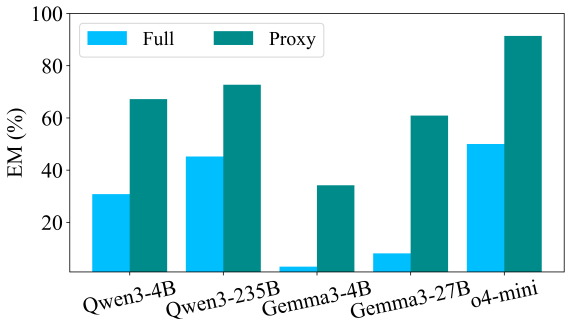


Figure 1: The disparity of model performance in the zero-shot setting on SciTrek (Li et al., 2025a) when showing *full long contexts* vs *short proxy contexts* in terms of exact match. The full context includes 128K tokens, while there are on average only around 650 tokens in a proxy context. Models work better on proxy contexts, while the proxy context requires the same reasoning process as their corresponding full long context.

[1] Our code, data, and models are available at `xxx.yyy.zzz`.

[2] Closed-source models from OpenAI and Google do not even provide access to their reasoning traces.

and computationally expensive.

However, there are indications that much of the computational cost of long-context processing is unnecessary: in many long-context tasks, only a small fraction of the input provides the evidence needed to the correct output. For instance, in multi-hop QA, such as HotpotQA (Yang et al., 2018), systems may retrieve entire articles, yet the answer typically depends on only a handful of relevant sentences. We refer to such subsets as *proxy contexts*: compact snippets that contain sufficient information to derive the correct answer. We hypothesize that the underlying reasoning should be *invariant* to the choice of context representation, i.e., a model should follow the same reasoning steps, whether it is conditioned on the full long context or on the corresponding proxy context.

Although the underlying information and required reasoning are the same, our experiments reveal a substantial performance gap between full and proxy contexts. As shown in Figure 1, LLMs across scales and model families perform markedly better when conditioned on proxy contexts. When given the full context, Li et al. (2025a) report that models often produce plausible *high-level* reasoning structures, yet hallucinate the *specific* facts needed to execute those steps correctly. In contrast, we find that the same models achieve much higher accuracy on each reasoning step when given proxy contexts. This suggests that LLMs struggle to correctly ground their reasoning in the relevant evidence in long inputs.

Because performance is often substantially higher on proxy contexts and reinforcement learning on them is far less computationally expensive, this motivates us to use the proxy contexts as a means for improving long-context reasoning. Figure 2 provides an overview of our training framework. We first obtain chain-of-thought reasoning traces (CoTs) based on proxy contexts (e.g., in SciTrek, metadata can serve as a proxy for the full-text articles). These traces can be obtained by reinforcement learning with verifiable rewards or by sampling from a larger teacher model. We then perform CoT distillation via supervised fine-tuning (SFT) training the target model to reproduce the proxy-derived reasoning traces *when given the full long contexts*.

This two-stage procedure first teaches the model to reason in a computationally efficient setting and then transfers that reasoning behaviour to long inputs. Compared to reinforcement learning directly on full contexts, our framework significantly reduces training cost and avoids requiring teacher-generated traces over long contexts. We summarize our contributions as follows:

- We introduce and formalize *proxy contexts* for long-context tasks, revealing a significant performance disparity between short proxy and full long contexts.
- We propose ProxyCoT, a novel training framework that leverages the short proxy contexts to acquire high-quality chain-of-thought reasoning traces, which in turn are used to enhance reasoning over full long contexts.
- Through extensive experiments across multiple models and datasets, we demonstrate that ProxyCoT consistently outperforms strong baselines while generating shorter reasoning traces, and generalizes to out-of-domain long-context reasoning tasks.

## 2 Related Work

**Reasoning in Language Models** Reinforcement learning and chain-of-thought distillation from teacher models are widely-used approaches for improving language model reasoning (Kumar et al., 2025). DeepSeek-R1 (DeepSeek-AI et al., 2025) showed that reasoning abilities can be developed through pure reinforcement learning without requiring supervised fine-tuning as a first step. This is especially useful for frontier models where 'teachers' may not exist to provide reasoning traces. However, this often comes at significant computational cost even for short-context tasks like mathematics.

DeepSeek-R1 has inspired many subsequent efforts in training reasoning models (Yang et al., 2025b; Bakouch et al., 2025; Mistral-AI et al., 2025). For smaller or non-frontier models, DeepSeek further showed that distilling reasoning patterns from larger models into smaller ones via SFT on reasoning traces (Li et al., 2023; Ho et al., 2023) outperforms applying reinforcement learning directly. It has since become standard practice to build SFT reasoning datasets through collecting traces from teacher models (Guha et al., 2025; Hugging Face, 2025; Li et al., 2025b).

However, both approaches face significant challenges when applied to long-context reasoning. Reinforcement learning becomes prohibitively expensive as context length increases due to extensive sampling over long sequences during training. CoT distillation avoids sampling costs, but still requires

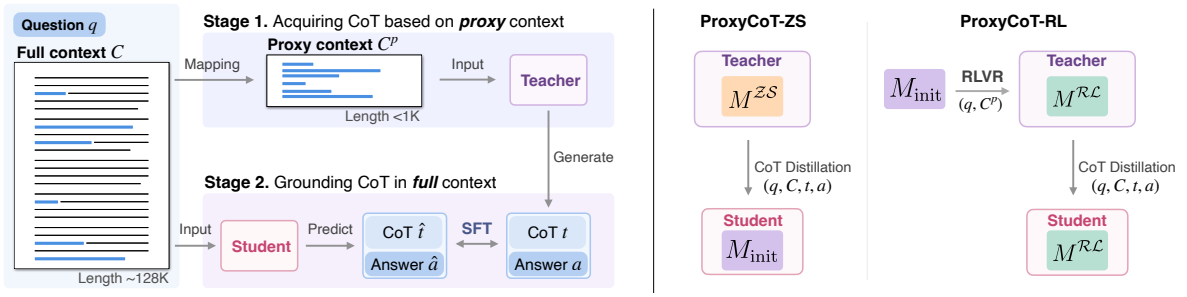


Figure 2: General two-stage pipeline of ProxyCoT (left), and two instantiations (right): ProxyCoT-ZS and ProxyCoT-RL. Given a target model $M_{\text{init}}$, ProxyCoT-ZS employs a large off-the-shelf model $M^{\mathcal{ZS}}$ as the teacher to generate CoTs from proxy contexts, and then fine-tunes $M_{\text{init}}$ as the student to generate the CoTs on corresponding long contexts. ProxyCoT-RL first optimizes $M_{\text{init}}$ using RLVR to obtain CoTs on proxy contexts, and subsequently fine-tunes the RL-optimized model $M^{\mathcal{RL}}$ as the student to ground these CoTs in corresponding long contexts.

querying large teacher models on full long contexts to generate reasoning traces, which is both time-consuming and computationally intensive. Moreover, even strong teacher models may fail to produce reliable traces for difficult long-context tasks. Our approach addresses these limitations by using proxy contexts to generate intermediate training signals, enabling efficient training without requiring long-context inference from teacher models.

**Long-context Language Models** The ability to fully utilize long sequences has been a long-standing challenge for language models (Liu et al., 2025). A core difficulty is representing token positions over long sequences. Rotary Position Embeddings (RoPE; Su et al. 2024) replace absolute positional embeddings with rotational transformations, and subsequent extensions (e.g., YaRN) rescale positional frequencies to support longer contexts without full retraining (Peng et al., 2024). Another bottleneck is the quadratic cost of Transformer attention. Sparse-attention models (e.g., Longformer) reduce computation by restricting attention patterns to selected entries of the full matrix, improving both prefilling and inference efficiency (Beltagy et al., 2020; Jiang et al., 2024; Fu et al., 2024). Many contemporary long-context LLMs interleave sparse and full-attention layers to balance efficiency and quality (Dubey et al., 2024; Yang et al., 2025a; Gemma Team, 2025).

Many modern LLMs also incorporate long-context-specific training data and post-training procedures. For example, Qwen2.5-1M (Yang et al., 2025c) and Qwen3 (Yang et al., 2025a) use synthetic long-context data during pre-training and multi-stage supervised fine-tuning tailored to long contexts. Similarly, OLMo 3 (Olmo et al., 2025) includes curated long-context data and synthetic aggregation-style tasks. However, due to the expense and evaluation challenges of long-context supervision, such training often targets generic long-context understanding rather than eliciting faithful, step-by-step reasoning on downstream tasks.

In this work, we *reformulate* long-context tasks into a setting that enables collecting high-quality reasoning traces (via proxy contexts), and then train models to reproduce these traces when conditioned on the original long context.

## 3 ProxyCoT Training

This section introduces ProxyCoT, our two-stage training framework designed to enhance long-context reasoning in question answering. Given proxy contexts containing the minimal information required to answer each question, ProxyCoT operates in a teacher-student paradigm with two stages, shown in Figure 2. In Stage 1, a teacher model generates high-quality reasoning traces over *proxy* contexts. In Stage 2, these reasoning traces are used to fine-tune a student model on the corresponding *long* contexts via chain-of-thought distillation. As the reasoning traces can be obtained from a large off-the-shelf model or based on reinforcement learning, ProxyCoT has two variants: ProxyCoT-ZS and ProxyCoT-RL (Figure 2 right).

### 3.1 Acquiring CoTs on Short Proxy Contexts

For any long-context question answering task, a *proxy context* $C^p$ denotes a compact version of the long input $C$ that preserves answerability. CoTs over proxy contexts should transfer to the corresponding full long contexts. Formally, we denote each example as a question–context pair $(q, C)$ with ground-truth answer $a$. Each context has a corresponding proxy, $C^p$, which is substantially

shorter ($|C^p| \ll |C|$) while containing sufficient information to answer the question. Our goal is to obtain a dataset $\mathcal{D} = \{(q_i, C_i^p, t_i, a_i)\}$ of reasoning traces $t$ generated conditioned on proxy contexts.

**Large Teacher Generation** We query a large off-the-shelf teacher model $M^{\mathcal{ZS}}$ to generate reasoning traces conditioned on proxy contexts: $t \sim p_\phi(t \mid q, C^p)$, where $\phi$ denotes the teacher model parameters. We retain only traces that produce correct answers, yielding high-quality demonstrations. The teacher estimates the distribution over reasoning traces given the question and proxy context. Compared to generating traces over full long contexts, inference on shorter proxies is substantially faster and more cost-effective.

**Reinforced Self-exploration** If a capable off-the-shelf large teacher model is unavailable (e.g., it is too expensive to run or performs poorly on our specific task), we train the target model $M_{\text{init}}$ directly on proxy contexts using reinforcement learning with verifiable rewards (RLVR). RLVR optimizes $p_\theta(t \mid q, C^p)$ to maximize the probability of generating traces that lead to correct answers. After training, the resulting model $M^{\mathcal{RL}}$ with parameters $\theta_{\text{RL}}$ estimates $p_{\theta_{\text{RL}}}(t \mid q, C^p)$, from which we sample reasoning traces for the dataset $\mathcal{D}$. This approach enables the model to learn task-specific reasoning structure while focusing exclusively on the necessary information in proxy contexts.

### 3.2 Grounding CoTs in Full Long Contexts

Stage 2 transfers the reasoning patterns learned on proxy contexts to full long-context inputs. The student model must learn to reproduce reasoning traces while conditioning on the full context $C$ rather than the proxy $C^p$. We achieve this through supervised fine-tuning (SFT) on reasoning traces conditioned on $(q, C)$. Crucially, the two ProxyCoT variants (illustrated in the right part of Figure 2) differ in their Stage 2 objectives:

**ProxyCoT-ZS** The off-the-shelf teacher model $M^{\mathcal{ZS}}$ generates traces from proxy inputs: $t \sim p_\phi(t \mid q, C^p)$. We then use the target model $M_{\text{init}}$ as the student to reproduce the teacher's reasoning structure while grounding it in the full long context. We fine-tune $M_{\text{init}}$ by minimizing:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,C,C^p,t)} \left[\log p_\theta(t \mid q, C)\right] \quad (1)$$

**ProxyCoT-RL** The RL-optimized model $M^{\mathcal{RL}}$ acts as the teacher generating reasoning traces from proxy inputs: $t \sim p_{\theta_{\text{RL}}}(t \mid q, C^p)$. We then *further train* $M^{\mathcal{RL}}$ as the student to minimize:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,C,C^p,t)} \left[\log p_{\theta_{\text{RL}}}(t \mid q, C)\right] \quad (2)$$

This continues training from the RL checkpoint, teaching the model to apply its learned reasoning patterns to full long contexts. As shown in our ablations (Table 8), starting from $M^{\mathcal{RL}}$ rather than $M^{\mathcal{S}}$ is crucial for ProxyCoT-RL's performance.

| Specification | Qwen3-4B | Gemma3-4B |
|---|---|---|
| Number of Parameters | 4B | 4B |
| Native Context Size | 256K | 32K |
| Supporting Context Size | 256K | 128K |
| Attention Architecture | Dense | Sparse |
| Release Date | July 2025 | March 2025 |

Table 1: Details of experimental models: Qwen3-4B-Instruct-2507 and Gemma3-4B-IT. Gemma3-4B-IT supports the 128K-token context based on RoPE scaling.

## 4 Experimental Setup

This section describes the models and datasets we use in our experiments. In addition to our experimental setting, we discuss implementation details and evaluation metrics.

### 4.1 Models and Datasets

We conduct experiments with two widely-used open-source long-context language models: Qwen3-4B-Instruct-2507 (Yang et al., 2025a) and Gemma3-4B-IT (Gemma Team, 2025).[3] The two models differ substantially in architecture and modality, as shown in Table 1. Qwen3-4B-Instruct-2507 is dense and text-only, while Gemma3-4B-IT is sparse and multimodal.

We report results on two question answering benchmarks: SciTrek (Li et al., 2025a) and HotpotQA (Yang et al., 2018). **SciTrek** is a long-context question answering benchmark, testing model reasoning capabilities over multiple scientific articles. Each instance consists of an *article collection* formed by concatenating several articles to a target length (released at 64k, 128k, 512K, and 1M tokens), with collections constructed either by random sampling within topical clusters or by traversing citation graphs to preserve citation structures. SciTrek's questions generally require many steps of reasoning, such as comparison, sorting, filtering, and aggregating. As these questions

[3] We obtained models from www.huggingface.co: Qwen/Qwen3-4B-Instruct-2507 and google/gemma-3-4b-it. Qwen3-4B-Instruct abbreviates Qwen3-4B-Instruct-2507.

| | SciTrek |
|---|---|
| $q$ | What is the smallest number of authors for any article in the collection? |
| $C$ | {7 full-text scientific articles} |
| $C^p$ | Article title: Existence and uniqueness for Legendre curves<br>There are 6 words in the title (separated by spaces).<br>There are 2 authors: Tomonori Fukunaga, Masatomo Takahashi<br>There are 9 references in the reference section.<br>The other provided articles are not cited by this article.<br>*{we cut down the text because of limited space.}*<br>Article title: Effect of higher-order interactions on synchronization of neuron models with electromagnetic induction<br>There are 12 words in the title (separated by spaces).<br>There are 4 authors: Mohanasubha Ramasamy, Subhasri Devarajan, Suresh Kumarasamy, Karthikeyan Rajagopal<br>There are 37 references in the reference section.<br>The other provided articles are not cited by this article. |
| | **HotpotQA** |
| $q$ | Luis Gianneo was teacher of which chief exponent of Argentine folk music? |
| $C$ | {67 full-text Wikipedia articles} |
| $C^p$ | Luis Gianneo (1897–1968) was an Argentine composer, pianist and conductor.<br>As music educator, he was the teacher of composers Ariel Ramirez, Juan Carlos Zorzi, Virtú Maragno, Pedro Ignacio Calderón and Rodolfo Arizaga, among others. |

Table 2: Example questions ($q$), abbreviated long contexts($C$), and their proxy contexts ($C^p$) for SciTrek (top) and HotpotQA (bottom).

(and their answers) are generated based on metadata about titles, authors, and references for each collection, we simply use a textual representation of the metadata as the proxy context of the full long context (see the example in Table 2).

**HotpotQA** is a multi-hop question answering benchmark based on Wikipedia articles, requiring models to reason and derive the answer from multiple documents. Each instance comprises of a collection of Wikipedia articles and a question-answer pair. To create inputs of varying length up to 128K tokens, we extended each instance by appending additional full Wikipedia articles that are cited by the original context articles as needed. Details about how we construct long contexts for HotpotQA are provided in Appendix A. For each question, HotpotQA provides human-annotated supporting sentences, which are a subset of sentences from the corresponding input articles. We use the concatenation of these annotated evidence sentences as the proxy context, since they contain the information required to answer the question (see Table 2).

Due to limited computational resources, we restrict our experiments to data with contexts of up to 128K tokens (7,290/413/840 instances for SciTrek and 4,136/400/600 for HotpotQA in training/development/testing).[4] More statistics for SciTrek and HotpotQA are in Table 3.

| Dataset | Full Context | Proxy | Question | Answer |
|---|---|---|---|---|
| SciTrek | 83,018.3 | 659.4 | 18.9 | 16.7 |
| HotpotQA | 77,764.4 | 301.1 | 24.5 | 4.5 |

Table 3: Dataset statistics for SciTrek and HotPotQA: average number of tokens in full contexts, proxy contexts, questions, and ground-truth answers, based on the tokenizer of Qwen3-4B-Instruct.

### 4.2 Implementation Details

**Large Teacher Sampling** We use Qwen3-235B-A22B-Thinking as the teacher model for both SciTrek and HotpotQA to generate reasoning traces on proxy contexts for Stage 1. It uses a Mixture-of-Experts architecture with 235B total parameters, of which only ~22B are active during inference, and supports contexts of up to 256K tokens. Among the open-source models we evaluated, it achieved the strongest overall performance, motivating its use as our teacher model. No training is required to collect reasoning traces from it; we sample three times and select only traces that result in correct answers. We run inference with vLLM (Kwon et al., 2023), and temperature sampling set to 0.7, nucleus sampling with $top\_p = 0.8$, and top-$k$ sampling with $top\_k = 20$. We set $min\_p$ to 0, and cap generations to a maximum of 32,768 tokens.

**RLVR Training** For RL training in Stage 1 (see Figure 2), we optimize our models using DAPO (Yu et al., 2025). At each training step, DAPO samples a group of reasoning traces from the policy and computes advantages for each trace through group-based normalization of their rewards. The model parameters are then updated based on these advantage estimates. As both of our tasks require short strings as answers, we adopt a simple sum of F1 and exact match to represent the reward as a function of the ground-truth answer $a$ and our predicted answer $\hat{a}$:

$$R(a, \hat{a}) = \text{F1}(a, \hat{a}) + \mathbb{1}_{a==\hat{a}} \quad (3)$$

We perform RL training with OpenRLHF (Hu et al., 2024) with a batch size of 64 and a maximum generation length of 2,048 tokens. We set the actor learning rate to 5e-7, apply dynamic reward filtering with a range of (0.3, 2.0), and use clipping

[4] All experiments were conducted on 8 NVIDIA HGX H200 GPUs.

| Model | Training Strategy | Proxy↑ | Full↑ |
|---|---|---|---|
| Qwen3-235B-Instruct | Zero-shot | 72.7 | 45.2 |
| Qwen3-235B-Thinking | Zero-shot | 85.6 | 48.8 |
| Qwen3-4B-Instruct | Zero-shot | 67.2 | 30.8 |
| | SFT on $C$ | 39.0 | 19.5 |
| | RLVR on $C$ | 66.1 | 32.9 |
| | SFT on $C$, CoT* | 45.8 | 31.6 |
| | ProxyCoT-ZS | 67.8 | 38.8 |
| | ProxyCoT-RL | 88.5 | 46.5 |
| Gemma3-4B-IT | Zero-shot | 34.2 | 3.0 |
| | SFT on $C$ | 19.1 | 12.7 |
| | RLVR on $C$ | 39.9 | 5.5 |
| | SFT on $C$, CoT* | 53.1 | 36.9 |
| | ProxyCoT-ZS | 64.2 | 36.5 |
| | ProxyCoT-RL | 69.8 | 43.7 |

Table 4: Performance of Qwen3-4B-Instruct and Gemma3-4B-IT on SciTrek, in terms of exact match (%). Models are evaluated with Proxy and Full contexts as input. $C$: full long contexts, and CoT*: chain-of-thought reasoning traces generated by Qwen3-235B-A22B-Thinking on full long contexts.

| Model | Training Strategy | Proxy↑ | Full↑ |
|---|---|---|---|
| Qwen3-235B-Instruct | Zero-shot | 92.1 | 60.8 |
| Qwen3-235B-Thinking | Zero-shot | 93.2 | 50.7 |
| Qwen3-4B-Instruct | Zero-shot | 91.3 | 44.5 |
| | SFT on $C$ | 92.6 | 48.8 |
| | RLVR on $C$ | 88.6 | 48.1 |
| | SFT on $C$, CoT* | 84.5 | 40.2 |
| | ProxyCoT-ZS | 91.4 | 50.3 |
| | ProxyCoT-RL | 92.1 | 52.7 |

Table 5: Performance of Qwen3-4B-Instruct on HotpotQA evaluated with GPT5-mini as judge in terms of accuracy (%). Models are evaluated with Proxy and Full contexts as input. $C$: full long contexts, and CoT*: chain-of-thought reasoning traces from Qwen3-235B-A22B-Thinking on full long contexts.

parameters (0.2, 0.3) for the policy update. For each prompt, we sample 8 trajectories. Models are trained for 10 epochs.

**SFT on CoTs** In Stage 2, both ProxyCoT-ZS and ProxyCoT-RL use supervised fine-tuning (SFT) to distil chain-of-thought traces. We implement SFT with OpenRLHF (Hu et al., 2024), using a batch size of 64 and a learning rate of 5e-6. We apply linear learning-rate warm up over the first 10% of training steps ($rl\_warmup\_ratio = 0.1$).

### 4.3 Evaluation Metrics

We generate final answers using the default decoding settings for Qwen3-4B-Instruct and Gemma3-4B-IT. For models based on Qwen3-4B-Instruct, we decode with temperature 0.7, $top\_p = 0.8$, $top\_k = 20$, $min\_p = 0$, and a maximum generation length of 2,048 tokens. For Gemma3-4B-IT based models, we use temperature 1.0, $top\_p = 0.95$, $top\_k = 64$, and a maximum of 2,048 generated tokens.

We evaluate model performance based on the quality of the generated answers. For SciTrek, following Li et al. (2025a), we use exact match and F1 comparing the generated answer against the ground-truth answer. For HotpotQA, where acceptable answers can be more variable, we adopt a model-based evaluation protocol following previous work (Sun et al., 2024; Perez-Beltrachini and Lapata, 2025). We use GPT5-mini as the judge, assessing the generated answer against the ground-truth answer in the prompt. The evaluation prompt and more details are provided in Appendix B.

## 5 Main Results

In this section, we present experimental results comparing our training framework against a range of baselines. We further analyze the reasoning traces produced by each method and evaluate robustness via an out-of-domain transfer task.

**ProxyCoT improves long-context reasoning across model families.** We compare our training framework with several baselines, which are based on Qwen3-4B-Instruct or Gemma3-4B-IT and do not make use of proxy contexts. These baselines include: (1) supervised fine-tuning (SFT) on full long contexts without generating reasoning traces, (2) reinforcement learning with verifiable rewards (RVLR) on full contexts, and (3) SFT on full long contexts using reasoning traces generated by Qwen3-235B-A22B-Thinking. As an upper bound, we also compare the smaller models against zero-shot Qwen3-235B-A22B-Thinking.

Results in Table 4 show that ProxyCoT-ZS and ProxyCoT-RL consistently outperform these training alternatives improving performance on both full long contexts and short proxy contexts, and across model architectures. Notably, ProxyCoT-RL substantially improves Qwen3-4B-Instruct, achieving performance competitive with the much larger Qwen3-235B-A22B-Thinking model. Moreover, ProxyCoT-RL consistently outperforms ProxyCoT-ZS, suggesting that reasoning traces generated via RLVR can be more effective distillation targets than zero-shot traces from a large teacher model.

| Training Strategy | CoT Tokens↓ | EM↑ |
|---|---|---|
| Zero-shot | 1,744 | 30.8 |
| RLVR on $C$ | 937 | 32.9 |
| SFT on $C$, CoT* | 6,683 | 31.6 |
| ProxyCoT-ZS | 5,520 | 38.8 |
| ProxyCoT-RL | 617 | 46.5 |

Table 6: Models trained with different training strategies use different numbers of chain-of-thought (CoT) tokens when showing full long contexts. Results are based on Qwen3-4B-Instruct on SciTrek. CoT*: chain-of-thought reasoning traces are generated by Qwen3-235B-A22B-Thinking on full long contexts.

**ProxyCoT performs well across datasets.** Table 5 reports results on HotpotQA, corroborating the trends observed on SciTrek: ProxyCoT consistently improves long-context question answering. Although the baseline Qwen3-4B-Instruct model is already strong on proxy inputs, ProxyCoT yields clear gains when evaluated on full long contexts, outperforming both SFT and RVLR on full contexts. In particular, ProxyCoT-RL achieves the best performance on full contexts, indicating that distilling high-quality traces obtained on proxy inputs remains effective for improving long-context reasoning on HotpotQA.

**ProxyCoT-RL reduces inference compute.** Beyond eliminating the need to generate teacher traces on full long contexts and requiring less computation during reinforcement learning, ProxyCoT-RL also induces substantially shorter reasoning traces at inference time. As shown in Table 6 (Qwen3-4B-Instruct on SciTrek), ProxyCoT-RL uses fewer CoT tokens on average while achieving better accuracy than the other optimization strategies. These results suggest that ProxyCoT-RL improves long-context performance while simultaneously reducing inference-time budget.

**ProxyCoT-RL also pushes out-of-domain long-context capabilities.** To further validate the out-of-domain generalization capabilities of ProxCoT, we report results on Loong (Wang et al., 2024), a benchmark designed to evaluate long-context language models through extended multi-document question answering (every document in each test case must be considered to derive the final answer). Loong features four long-context task types: Spotlight Locating, Comparison, Clustering, and Chain of Reasoning, and represents the domains of Financial Reports, Academic Papers, and Legal Cases (with context lengths ranging from 10K to beyond 200K tokens). We evaluate ProxyCoT-RL (trained on SciTrek) on Loong, considering English instances only up to 128K tokens without any further adaptation.[5]

| Model | Training | Financial↑ | Academic↑ |
|---|---|---|---|
| Qwen3-4B-Instruct | Zero-shot | 37.76 | 24.91 |
| | ProxyCoT-RL | 40.83 | 42.51 |
| Gemma3-4B-IT | Zero-shot | 25.85 | 3.55 |
| | ProxyCoT-RL | 32.05 | 24.32 |

Table 7: Results on Loong (domains of Financial Reports and Academic Papers) when the model is zero-shot or trained on SciTrek with ProxyCoT-RL. Evaluation is based on GPT5-mini as the judge.

As shown in Table 7, both Qwen3-4B-Instruct and Gemma3-4b-IT demonstrate improved performance on Loong across different domains after training with ProxyCoT-RL.[6] On financial reports, both models show gains, with Gemma3-4B-IT exhibiting particularly substantial improvement (25.85 → 32.05). More notably, on academic papers, a domain represented in Loong but with different questions than those in SciTrek, both models achieve substantial performance increase, with Gemma3-4B-IT improving from 3.55 to 24.32. These cross-domain gains demonstrate that ProxyCoT-RL enhances general long-context reasoning capabilities of the models rather than simply memorizing task-specific patterns.

## 6 Analysis and Ablations

We further analyse the significance of the two training stages in our framework, and conduct experiments with alternative proxy contexts to examine whether their quality impacts performance.

### 6.1 Ablations on Two-stage Training

Table 8 summarizes ablation studies with both Qwen3-4B-Instruct and Gemma3-4B-IT. We report performance on proxy and full contexts in three settings: (a) our full training with both RVLR and SFT; (b) only RVLR-based training; and (c) only SFT-based training but with reasoning traces from RLVR. Models trained with RVLR learn to reason based on proxy contexts and can be applied to long contexts without any grounding. Models trained

[5]We tested our models on 309 instances in Financial Reports and 119 in Academic Papers, omitting instances in the domain of Legal Cases which are all in Chinese.

[6]We follow the evaluation prompt and code from https://github.com/MozerWang/Loong.

| Model | Stage 1 (RVLR) | Stage 2 (SFT) | Proxy↑ | Full↑ |
|---|---|---|---|---|
| Qwen3-4B-Instruct | ✓ | ✓ | 88.5 | 46.5 |
| | ✓ | ✗ | 91.5 | 29.0 |
| | ✗ | ✓ | 77.5 | 46.3 |
| Gemma3-4B-IT | ✓ | ✓ | 69.8 | 43.7 |
| | ✓ | ✗ | 88.5 | 8.0 |
| | ✗ | ✓ | 65.2 | 37.3 |

Table 8: Ablations on ProxyCoT-RL using Qwen3-4B-Instruct and Gemma3-4B-IT with SciTrek. We report results in terms of exact match (%) on proxy and full contexts. There are two stages in training: reinforcement learning with verifiable rewards (RLVR) in Stage 1, and supervised fine-tuning (SFT) in Stage 2.

| | Type | Tokens | EM↑ |
|---|---|---|---|
| SciTrek | Random sentences from full context | 1,102 | 3.4 |
| | Titles, authors and references of the context articles | 20,344 | 24.6 |
| | Descriptions of the structured metadata (provided by SciTrek) | 659 | 91.5 |
| HotPotQA | Random sentences from full context | 385 | 15.8 |
| | Sentences semantically related to the question | 364 | 35.1 |
| | Ground-truth evidence sentences (provided by HotPotQA) | 301 | 92.2 |

Table 9: Performance of Qwen3-4B-Instruct after RLVR with different types of proxy contexts on SciTrek and HotPotQA in terms of EM (%).

with SFT learn the task by observing long contexts, questions, and reasoning traces with answers.

Our results show model performance on full long contexts is best when *both* training stages are employed (see column Full in Table 8). When only reinforcement learning takes place (Stage 1), we observe that performance on proxy contexts is superior to the other two strategies (i.e., Stage 2 or Stage 1 + Stage 2). We also find the first stage of training gives Gemma3-4B-IT a larger performance boost in long contexts (37.3 → 43.7) than Qwen3-4B-Instruct (46.3 → 46.5). As Gemma3-4B-IT has worse zero-shot performance than Qwen3-4B-Instruct (see Table 4), we conjecture that models with poorer zero-shot capabilities at long contexts may benefit more from RL on proxy contexts.

### 6.2 Alternative Proxy Contexts

Proxy contexts in our experiments were obtained from annotations provided in SciTrek and HotpotQA (see examples in Table 2). In this section, we experiment with automatically created proxies under the assumption that high-quality annotations might be scarce. We examine different proxies for full long contexts based on information retrieval. For SciTrek, we simply use article titles, authors, and reference sections of input articles as the proxy context, since the questions (and their answers) focus on information contained therein. For HotpotQA, we use sentences semantically related to the question as the proxy context, assuming that unrelated sentences introduce noise into the answer generation process. We first obtained related sentences via lexical search with BM25 (Robertson and Zaragoza, 2009) and then via semantic search based on sentence embeddings (*text-embedding-3-small* from OpenAI). Finally, we retrieved random sentences as proxy contexts for both datasets.

As shown in Table 9, proxy contexts based on annotations are of much higher quality compared to automatically sourced ones. Perhaps unsurprisingly, proxies based on randomly selected sentences carry no useful information for answering the question. For HotpotQA, semantically related sentences based on information retrieval are better than random ones but still a poor substitute for annotation-based proxies. For SciTrek, even though titles, authors and references are the only relevant pieces of information for answering the question, their unstructured nature makes learning difficult. Structured metadata is far more advantageous for learning targeted reasoning traces than pure text (despite the two formats being equally sufficient to answer the question). This analysis shows that better designed proxy contexts enable ProxyCoT to improve long-context reasoning.

## 7 Conclusion

We introduced ProxyCoT, a two-stage training framework for long-context reasoning that leverages *proxy contexts*, i.e., shorter inputs that preserve the information needed to solve a task. ProxyCoT is motivated by a consistent performance gap between full long contexts and their proxy counterparts, despite requiring the same underlying reasoning. Across model families and benchmarks, we show that ProxyCoT improves long-context question answering while reducing inference-time budget. More broadly, ProxyCoT offers a practical route to strengthening long-context reasoning without relying on teacher-generated traces over full long contexts, providing an efficient alternative in settings where training and supervising long-context models remain challenging.

## Limitations

While learning via proxy contexts offers a promising direction for improving long-context reasoning, our work has several limitations:

- Our approach assumes access to proxy contexts that contain sufficient evidence to answer each question. Constructing such proxies can be non-trivial for some real-world tasks and domains. In future work, we plan to explore methods for automatically constructing effective proxy contexts.
- Some of our tasks can also be addressed with workflow-based solutions, such as retrieval-augmented generation, which retrieves relevant evidence and conditions generation on the retrieved subset. Studying such systems-level approaches is outside the scope of this paper; our focus is on improving the inherent long-context capabilities of language models.
- Although our framework generalizes across multiple model families and datasets, our experiments are limited to English due to dataset availability and computational constraints. Extending to other languages and domains is an important direction for future work, contingent on obtaining suitable proxy contexts.

## References


Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Moa: Mixture of sparse attention for automatic large language model compression. *Preprint*, arXiv:2406.14909.

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

Gemma Team. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *Preprint*, arXiv:2506.04178.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14852–14882. Association for Computational Linguistics.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

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

Huiqiang Jiang, YUCHENG LI, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman H. Khan, and Fahad Shahbaz Khan. 2025. LLM post-training: A deep dive into reasoning large language models. *CoRR*, abs/2502.21321.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving

with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2665–2679. Association for Computational Linguistics.

Miao Li, Alexander Gurung, Irina Saparina, and Mirella Lapata. 2025a. Who gets cited most? benchmarking long-context language models on scientific articles. *CoRR*, abs/2509.21028.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and Xian Li. 2025b. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. *Preprint*, arXiv:2507.01921.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, and 18 others. 2025. A comprehensive survey on long context language modeling. *CoRR*, abs/2503.17407.

Llama Team. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *Technical Report*.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, and 82 others. 2025. Magistral. *Preprint*, arXiv:2506.10910.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty quantification in retrieval augmented question answering. *Trans. Mach. Learn. Res.*, 2025.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5627–5646. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025c. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

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, and 16 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

## A Long Context Construction for HotpotQA

For the long contexts of HotpotQA, we retained the original questions and answers from Yang et al. (2018) and constructed long contexts using the procedure described by Li et al. (2025a). We used the distractor setting and selected only bridge questions at the hard difficulty level. Bridge questions require chain reasoning, where the model must first identify a bridge entity connecting two paragraphs before completing the second reasoning step. We further filtered for examples with at least 3 supporting facts, i.e., sentences that crowd workers annotated as necessary for answering the question. Our training and validation sets were sampled from the original train split, while our test set was sampled from the original validation split, resulting in 4136 training examples, 400 validation examples, and 600 test examples.

We first retrieved the full texts of the Wikipedia articles containing the supporting facts, as well as the articles corresponding to the retrieved paragraphs in the distractor setting. We used the preprocessed Wikipedia dump released with HotpotQA for retrieval of all articles. We then expanded the context to 128K tokens by adding related Wikipedia articles: we gathered the links mentioned in each article and followed them to retrieve additional articles. We limited the maximum depth to two hops from the original Wikipedia articles. Finally, we concatenated all articles to form the full long context.

## B Evaluation Metrics with Model-as-Judge for HotpotQA

To compare predicted and ground-truth answers on HotpotQA, we use the prompt from Sun et al. (2024), who found 98% agreement between human judgments and LLM-based scores. We use GPT-5-mini as a judge (`gpt-5-mini-2025-08-07`, `https://platform.openai.com/docs/models/gpt-5-mini`). The prompt template based on few-shot prompting obtained from Sun et al. (2024) is shown in Figure 3.

**The Prompt template for HotpotQA Evaluation**

You need to check whether the prediction of a question-answering system to a question is correct.
You should make the judgment based on a list of ground truth answers provided to you.
Your response should be "correct" if the prediction is correct or "incorrect" if the prediction is wrong.

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: W Shakespeare
Correctness: correct

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: Roma Gill and W Shakespeare
Correctness: correct

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: Roma Shakespeare
Correctness: incorrect

Question: What country is Maharashtra Metro Rail Corporation Limited located in?
Ground truth: ["India"]
Prediction: Maharashtra
Correctness: incorrect

Question: What's the job of Song Kang-ho in Parasite (2019)?
Ground truth: ["actor"]
Prediction: He plays the role of Kim Ki-taek, the patriarch of the Kim family.
Correctness: correct

Question: Which era did Michael Oakeshott belong to?
Ground truth: ["20th-century philosophy"]
Prediction: 20th century.
Correctness: correct

Question: Edward Tise (known for Full Metal Jacket (1987)) is in what department?
Ground truth: ["sound department"]
Prediction: 2nd Infantry Division, United States Army
Correctness: incorrect

Question: What wine region is Finger Lakes AVA a part of?
Ground truth: ["New York wine"]
Prediction: Finger Lakes AVA
Correctness: incorrect

Question: {QUESTION}
Ground truth: {GROUND_TRUTH}
Prediction: {PREDICTION}
Correctness:

Figure 3: The prompt template for HotpotQA evaluation with GPT5-mini as the judge.