Chunks as Arms: Multi-Armed Bandit-Guided Sampling for Long-Context LLM Preference Optimization

Anonymous ACL submission

Abstract

Long-context modeling is critical for a wide range of real-world tasks, including long-context question answering, summarization, and complex reasoning tasks. Recent studies have explored fine-tuning Large Language Models (LLMs) with synthetic data to enhance their long-context capabilities. However, the effectiveness of such approaches is often limited by the low diversity and factual inconsistencies in the generated data. To address these challenges, we propose LongMab, a novel framework that leverages a Multi-Armed Bandit (MAB) rollout strategy to identify the most informative chunks from the given long context for sampling high-quality and diverse responses and constructing preference data pairs for Direct Preference Optimization (DPO) training. Specifically, we treat context chunks as arms of MAB, select chunks based on their expected reward scores to input into LLMs to generate responses, and iteratively update these scores based on reward feedback. Both exploration and exploitation during the rollout process enable the LLM to focus on the most relevant context segments, thereby generating and collecting high-quality and diverse responses. Experimental results on both Llama and Qwen show the effectiveness of LongMab by achieving more than a 4% improvement on long-context reasoning benchmarks. All data and code will be released via GitHub.

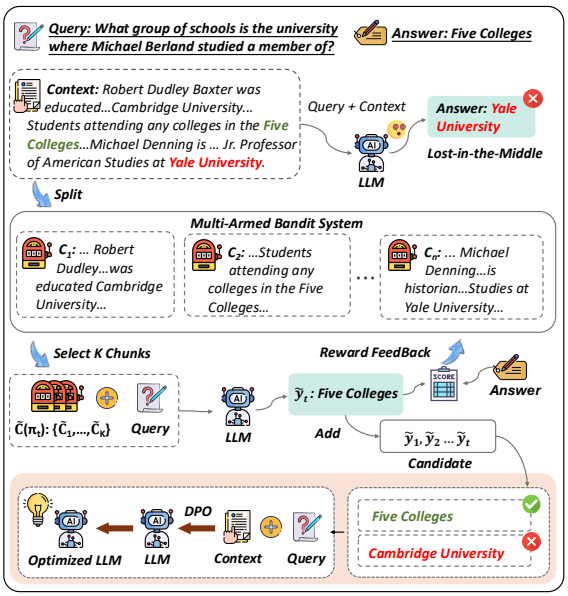


Figure 1: Illustration of the LongMab framework.

1 Introduction

Recent advancements in Large Language Models (LLMs), particularly the expansion of their context window (Dubey et al., 2024; Yang et al., 2024), have enabled their application in a variety of long-context tasks (Yang et al., 2025; Wang et al., 2022a; Zhao et al., 2024; Liu et al., 2024). Despite these developments, LLMs continue to suffer from the "lost-in-the-middle" problem (Liu et al., 2023; He et al., 2024), where LLMs tend to overemphasize the beginning and end of a long input while neglecting critical information in the middle. To address this challenge, recent studies have focused on constructing high-quality Supervised Fine-Tuning (SFT) datasets specifically designed to extend the context window to long-context scenarios (Chen et al., 2023; Bai et al., 2024a; Chen et al., 2024). Although these methods achieve certain improvements, they often induce overfitting to training-specific signals (Li et al., 2025), which in turn leads to catastrophic forgetting of general capabilities (Luo et al., 2023).

An alternative line of research investigates Direct Preference Optimization (DPO) (Rafailov et al., 2023; Sun et al., 2025), which enhances the long-context understanding capability of LLMs through preference-pair training. To this end, some works (Li et al., 2024a; Zhang et al., 2024a) employ full-context sampling strategies to construct preference pairs for DPO training. However, these

methods struggle to guarantee response quality, as LLMs remain highly susceptible to noise introduced by the lengthy context (Shi et al., 2023; Xu et al., 2025). To improve the quality of preference pair construction, recent studies (Tang et al., 2024) introduce chunk-aware sampling strategies. These methods divide the input context into multiple chunks and construct preference pairs based on various chunk combinations: relevant chunks are used to generate positive responses, while other chunks containing noise are treated as disturbances to encourage more diverse outputs. Despite these advantages, such approaches often overlook the supportive relationships among different chunks, which limits the ability of LLMs to comprehensively identify and integrate informative content across the entire context.

In this paper, we propose **M**ulti-**a**rmed **b**andit-guided sampling for **Long**-context LLM Preference Optimization (LongMab). As illustrated in Figure 1, the core idea is to treat divided context chunks as the arms of a Multi-Armed Bandit (MAB) and adopt a chunk-aware sampling strategy to encourage LLMs to construct higher-quality preference pairs for DPO training. During bandit rollouts, LongMab applies the Upper Confidence Bound (UCB) algorithm to select a subset of chunks according to their estimated rewards, thereby balancing exploration and exploitation. At each step, the selected chunks are fed into the LLM to produce a response, which is then evaluated to obtain a reward. This reward updates the expected values of the corresponding chunks, guiding subsequent selection. Through this iterative process, LongMab progressively prioritizes more informative chunk combinations, ultimately yielding diverse and high-quality responses.

The experimental results demonstrate the effectiveness of LongMab, which consistently outperforms existing SFT- and DPO-based baselines across multiple long-context tasks, achieving average gains exceeding 4%. Further analyses show that the UCB-guided sampling strategy effectively balances exploration and exploitation, enabling the model to explore diverse chunk combinations while progressively refining the selection toward evidence-rich and complementary information. This balance ultimately leads to higher quality and greater diversity in the sampled responses, allowing LongMab to construct effective and more informative preference pairs that significantly facilitate the DPO training process.

2 Related Work

Numerous studies have sought to improve how Large Language Models (LLMs) utilize long contexts (Hsieh et al., 2024; Levy et al., 2024; Wang et al., 2025a). Within this area, data-centric methods (Chen et al., 2023; Zhang et al., 2024a; Chen et al., 2025a), which focus on training better models through superior data engineering—have become the mainstream approach due to their excellent performance. These strategies generally fall into two primary categories, which mainly differ in how they construct and leverage training signals.

The first of these categories, Supervised Fine-Tuning (SFT), has been a foundational approach in this domain (Chen et al., 2023; Xiong et al., 2023; An et al., 2024; Xu et al., 2024; Li et al., 2024b). Early efforts relied on human annotations to create long-context question-answer pairs for training (Chen et al., 2023; Xu et al., 2024). However, the significant cost and scalability challenges of this manual process prompted a shift towards automated data synthesis. This is now commonly achieved by leveraging powerful LLMs to generate QA pairs from long documents, often via the self-instruct technique (Wang et al., 2022b; Bai et al., 2024a). To further refine data quality, the most recent works have augmented this pipeline by employing LLMs as judges to filter synthetic samples, thus ensuring a more reliable training set (Chen et al., 2024; Zhu et al., 2025).

As a counterpart to SFT, the other major approach leverages preference-based signals, with recent work focusing on Reinforcement Learning (RL) and particularly Direct Preference Optimization (DPO) (Rafailov et al., 2023; Tang et al., 2024; Li et al., 2024a). A primary challenge in this area is sourcing high-quality preference pairs from long contexts (Zhang et al., 2024a). One line of work addresses this by developing sophisticated reward models, using techniques like LLM-as-a-Judge or self-consistency to generate reliable preference labels (Zhang et al., 2024a; Li et al., 2024a). While these methods are effective, they can be hampered by contextual noise in long documents, which leads to suboptimal response generation and limits training efficacy (Shi et al., 2023). A complementary approach is heuristic sampling, in which positive responses are generated from query-relevant chunks and negative responses from irrelevant chunks or the full context (Tang et al., 2024; Chen et al., 2025a; Sun et al., 2025). How-

ever, this relevance-based method does not explore different chunk combinations for response generation, limiting its ability to identify the most informative evidence from the full context. To address this limitation, LongMab designs a multi-armed-bandit-guided sampling strategy that dynamically estimates the utility of chunk combinations during rollouts, constructing more informative preference pairs for DPO training.

3 Methodology

As illustrated in Figure 2, this section introduces LongMab, a novel framework designed to enhance the long-context reasoning ability of language models by leveraging the Multi-Armed Bandit (MAB) paradigm. We begin by describing how a MAB rollout process is used to optimize response sampling for more effective preference learning (§ 3.1). We then explain how the Upper Confidence Bound (UCB) algorithm is integrated to improve response quality by balancing the exploration-exploitation trade-off during the sampling process (§ 3.2).

3.1 Optimizing Long-Context LLMs via Multi-Armed Bandit-Guided Sampling

In long-context reasoning, a generation model $\mathcal{M}$ is tasked with generating an answer y for a given question q by utilizing information from a long context C:

$$y = \mathcal{M}(C, q). \tag{1}$$

To enhance the model's long-context reasoning, we follow recent studies (Zhang et al., 2024a) and employ DPO (Rafailov et al., 2023) to fine-tune the model on a preference dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(C,q,y^+,y^-)\sim\mathcal{D}}[\log \sigma(\beta \log \frac{\mathcal{M}(y^+ \mid C, q)}{\mathcal{M}^{\text{ref}}(y^+ \mid C, q)} - \beta \log \frac{\mathcal{M}(y^- \mid C, q)}{\mathcal{M}^{\text{ref}}(y^- \mid C, q)})], \tag{2}$$

where β is a hyperparameter, $\mathcal{M}^{\text{ref}}$ is a fixed reference model and each instance $(C, q, y^+, y^-) \in \mathcal{D}$ contains a context-question pair (C, q), along with a preferred (y^+) and dispreferred (y^-) response.

Chunk-Aware Response Sampling. To generate high-quality preference data for more effective training, we introduce a chunk-aware response sampling framework.

Specifically, we first divide the long context C into n equal-length chunks, denoted as $\mathcal{C}_{\text{chunk}} = \{C_1, C_2, \ldots, C_n\}$. Instead of generating a response from the entire, noisy long context C, our method samples from a subset of K chunks, $\tilde{\mathcal{C}}(\pi) = \{\tilde{C}_1, \ldots, \tilde{C}_K\}$, selected by a policy π. The response is then generated as follows:

$$\tilde{y} = \mathcal{M}(\tilde{\mathcal{C}}(\pi), q). \tag{3}$$

This sampling process is repeated for T iterations. In each step, the selection policy π evolves to explore different chunk combinations, with the goal of identifying the most informative subsets. This procedure yields a diverse set of candidate responses, $\tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_T\}$, which are then scored using the following reward function $r(\cdot)$ to construct preference pairs:

$$r(\tilde{y}) = (\text{SubEM}(\tilde{y}, y^*) + \text{F1}(\text{Ans}(\tilde{y}), y^*))/2, \tag{4}$$

where y^* is the ground-truth answer and SubEM/F1 are reasoning quality metrics. The winning response (y^+) and losing response (y^-) are the candidates from $\tilde{Y}$ with the highest and lowest reward scores, respectively, which completes the preference tuple (C, q, y^+, y^-) for tuning.

Optimize Chunk Selection via Multi-Armed Bandit Rollouts. To further improve the chunk-aware response sampling process, we design a chunk selection strategy π based on MAB rollouts to adaptively explore useful chunks.

Specifically, we model the chunk selection process as an MAB problem, where each chunk $C_i \in \mathcal{C}_{\text{chunk}}$ is treated as an individual arm. The process unfolds over T rollout steps. In each step $t \in \{1, \ldots, T\}$, the MAB policy π_t selects a subset of K chunks with the highest estimated expected rewards, $\tilde{\mathcal{C}}(\pi_t)$. This subset is then used along with the question q to prompt the LLM, generating a single response $\tilde{y}_t$. After T rollouts, these responses are collected to form the final candidate set, $\tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_T\}$.

$$\tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_T\}, \quad \text{where} \quad \tilde{y}_t = \mathcal{M}(\tilde{\mathcal{C}}(\pi_t), q), \tag{5}$$

where $\tilde{y}_t$ denotes the response generated by the LLM at rollout step t, and $\tilde{Y}$ are used to construct preference pairs for training. In the rollout process, the expected reward of a chunk $\tilde{C}_i \in \tilde{\mathcal{C}}(\pi_t)$ is determined by the quality of the response $\tilde{y}_t$. If the selected chunks $\tilde{C}_i$ lead the LLM to generate high-quality responses, they will receive higher expected rewards. Consequently, the MAB strategy is more likely to select more informative chuncks that win a higher expected reward in the next round, thereby continuously enhancing the quality of the generated responses.

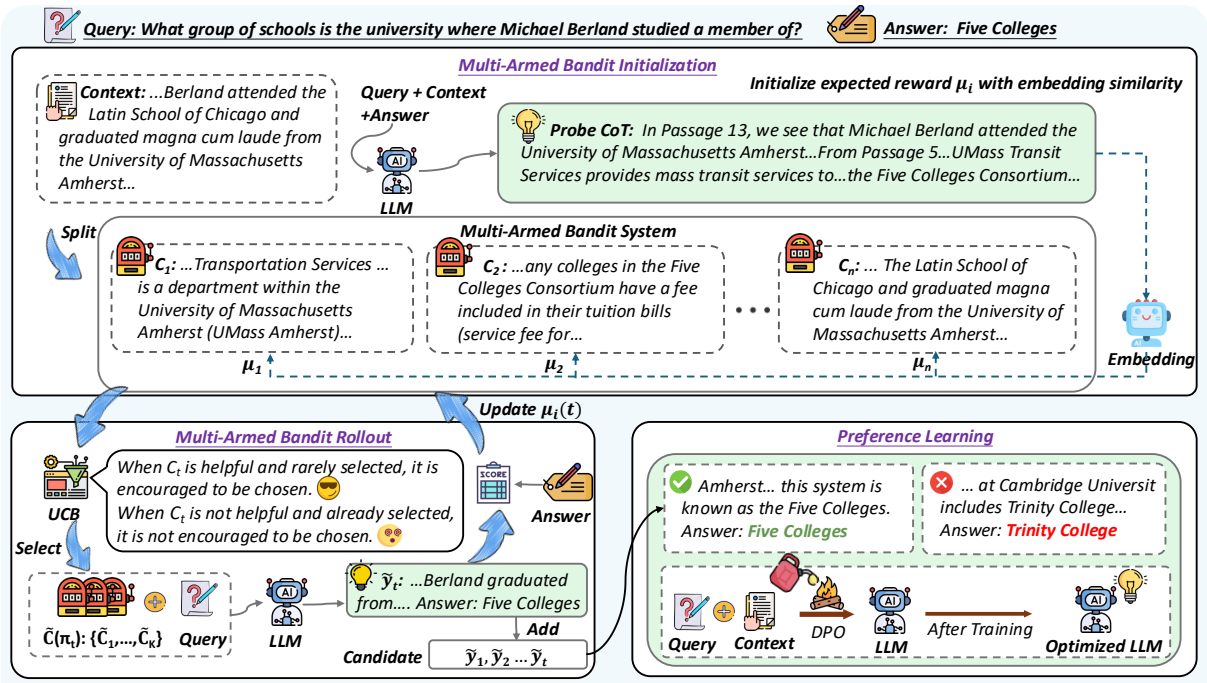


Figure 2: Illustration of multi-armed bandit-guided sampling for long-context LLM optimization (LongMab). LongMab collects preference pairs during the rollout process for DPO training.

3.2 Adaptive Chunk Prioritization through Multi-Armed Bandit Rollouts

The MAB rollout process must contend with the classic exploration–exploitation dilemma (Auer et al., 2002). Exploring a broader set of chunk combinations increases the chance of uncovering highly rewarding evidence but comes at a substantial computational cost. On the other hand, limited exploration may overlook informative chunks that are crucial for generating high-quality responses. Thus, it is essential to strike a careful balance between exploration (sampling under-visited chunks) and exploitation (favoring chunks with high expected rewards).

To solve this decision-making problem, we use the Upper Confidence Bound (UCB) algorithm (Komiyama et al., 2024) to guide chunk selection at each step. Specifically, at the t-th rollout step, the UCB score for chunk C_i is computed as:

$$\mathrm{UCB}_t(C_i) = \mu_i(t) + \alpha \cdot \sqrt{2 \ln t / (n_i(t) + \epsilon)}, \quad (6)$$

where $\mu_i(t)$ is the expected reward of chunk C_i (initialized via a probe-based strategy discussed later), $n_i(t)$ is its selection count, and t is the current rollout step. The hyperparameter α serves to balance exploration and exploitation, while the constant ϵ prevents division by zero. After scoring all chunks in $\mathcal{C}_{\text{chunk}}$, we form the subset for the current rollout, $\tilde{\mathcal{C}}(\pi_t)$, by selecting the top-$K$ chunks with the highest UCB scores:

$$\tilde{\mathcal{C}}(\pi_t) = \mathrm{TopK}_{C_i \in \mathcal{C}_{\text{chunk}}} \mathrm{UCB}_t(C_i). \quad (7)$$

The UCB score of each chunk is updated during the rollout process based on the current expected reward of the chunk. Then, we will provide a detailed explanation of how to initialize the expected reward for each chunk and the update process of each chunk's UCB score.

Multi-Armed Bandit Initialization via Evidence Probing. To alleviate the cold-start problem in multi-armed bandit rollouts (Wang et al., 2025b), we adopt a probe-based initialization strategy (Chen et al., 2025b) that assigns informed expected rewards $\mu_i(1)$ to each chunk C_i prior to the selection process in rollout step $t = 1$.

Specifically, we prompt the LLM $\mathcal{M}$ to generate a faithful reasoning trace, y_{Probe}, that explicitly identifies the evidence in the long context C required to derive the ground-truth answer, y^*:

$$y_{\text{Probe}} = \mathcal{M}(\mathrm{Instruct}_{\text{extract}}(q, \mathcal{C}, y^*)), \quad (8)$$

where $\mathrm{Instruct}_{\text{extract}}$ represents the instruction for extracting evidence from $\mathcal{C}$. We then score each chunk C_i against the probe trace y_{Probe} using cosine similarity in their embedding space:

$$s_i = \cos(\mathrm{Emb}(y_{\text{Probe}}), \mathrm{Emb}(C_i)). \quad (9)$$

These similarity scores s_i then become the initial expected rewards $\mu_i(1)$. Consequently, the initial policy π_1 selects the top-K chunks with the highest scores, forming the subset $\tilde{\mathcal{C}}(\pi_1)$. This strategy provides the MAB policy with a strong semantic prior, improving early-stage efficiency and reducing blind exploration.

UCB Score Update with MAB Rollouts. In each rollout step t of the MAB rollout process, we need to update the corresponding UCB scores of the chunks based on their expected rewards. During the t-th sampling step, we prompt the LLM $\mathcal{M}$ to answer the question q based on the selected chunks $\tilde{\mathcal{C}}(\pi_t)$:

$$\tilde{y}_t = \mathcal{M}(\tilde{\mathcal{C}}(\pi_t), q), \tag{10}$$

We then evaluate the utility of the selected chunks $\tilde{\mathcal{C}}(\pi_t)$ by computing a reward score $r(\tilde{y}_t)$ for the generated response $\tilde{y}_t$. Next, we update the UCB statistics based on the reward $r(\tilde{y}_t)$. For each chunk $C_i \in \mathcal{C}_{\text{chunk}}$, we update its selection count at the end of rollout step t:

$$n_i(t+1) = \begin{cases} n_i(t) + 1 & \text{if } C_i \in \tilde{\mathcal{C}}(\pi_t), \\ n_i(t) & \text{otherwise,} \end{cases} \tag{11}$$

and update its expected reward μ_i using an incremental average:

$$\mu_i(t+1) = \begin{cases} \frac{1}{t}\left(\mu_i(t) \cdot (t-1) + r(\tilde{y}_t)\right) & \text{if } C_i \in \tilde{\mathcal{C}}(\pi_t), \\ \mu_i(t) & \text{otherwise.} \end{cases} \tag{12}$$

The updated values $n_i(t+1)$ and $\mu_i(t+1)$ are subsequently used to compute UCB scores for the next rollout step $t+1$ using Eq. 6, enabling the bandit policy to continuously refine its estimation of chunk utility based on observed response quality.

4 Experimental Methodology

In this section, we describe the datasets, baselines, evaluation metrics, and implementation details.

Datasets. In our experiments, we use the MuSiQue training dataset (Trivedi et al., 2022) to sample responses and construct preference data pairs for DPO training. To better simulate long-context scenarios, we follow the methodology of previous work (Li et al., 2024a; Zhu et al., 2025), augmenting the context with randomly sampled Wikipedia documents to extend its length to 8k-16k tokens. For evaluation, we adopt five long-context QA tasks drawn from two widely used benchmarks: LongBench (Bai et al., 2024b) and InfiniteBench (Zhang et al., 2024b). Our evaluation suite includes four datasets from LongBench, including MuSiQue (Trivedi et al., 2022), 2WikiMultihopQA (Ho et al., 2020), MultiFieldQA-En (Bai et al., 2024b), and NarrativeQA (Kočiskỳ et al., 2018), along with the En.QA task from InfiniteBench. Detailed statistics for these test sets are provided in Appendix A.4.

Baselines. For a comprehensive evaluation of LongMab, we benchmark its performance against three distinct classes of baselines: (1) vanilla LLMs, (2) SFT-based methods, and (3) DPO-based methods.

Our SFT baselines include two established models, LongAlpaca (Chen et al., 2023) and LongAlign (Bai et al., 2024a), which are both fine-tuned on synthetic long-context QA data. We also introduce LongMab-SFT, an in-house baseline trained exclusively on the "chosen" responses from the preference dataset constructed for LongMab. For the DPO comparison, we select four state-of-the-art models: LongReward (Zhang et al., 2024a), SeaLong (Li et al., 2024a), LongFaith (Yang et al., 2025), and Logo (Tang et al., 2024). These models employ diverse strategies for generating preference pairs. For instance, LongReward and SeaLong rely on an LLM-as-a-Judge and self-consistency, respectively, to create reliable labels. LongFaith, in contrast, leverages attributed prompts to elicit positive responses from relevant passages while generating negative ones from the full, ungrounded document. Finally, Logo adopts a chunk-based approach, sampling positive responses from relevant document chunks and constructing negative ones by incrementally adding irrelevant chunks.

Evaluation Metrics. Following previous work (Li et al., 2024a; Yang et al., 2025), we employ the two complementary rule-based metrics for evaluation. Specifically, the substring exact match (SubEM) score checks whether the golden answer appears as a substring in the model's output. While widely used, it can be hacked by generating overly long responses that increase the chance of including the golden answer (Yang et al., 2025). To provide a more robust evaluation, we additionally report F1 score, which captures token-level overlap between the prediction and the golden answer by computing the harmonic mean of precision and recall.

Implementation Details. In line with previous studies (Li et al., 2024a; Yang et al., 2025), we conduct experiments using Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024) as backbone models. For training, we set the learning rate to 2×10^{-5}

Model	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
Llama-3.1-8B-Instruct												
Vanilla LLM	33.50	36.56	69.50	65.45	18.66	44.34	18.00	26.61	26.49	32.19	33.23	41.03
LongAlpaca	34.00	38.36	69.50	63.64	16.66	40.67	16.50	26.88	27.35	30.58	32.80	40.03
LongAlign	36.50	37.62	69.00	59.35	22.00	48.49	18.50	26.38	23.93	24.85	33.99	39.34
LongMab-SFT	35.00	39.60	69.00	65.45	21.33	45.03	15.50	24.98	23.93	27.33	32.95	40.48
LongReward	40.50	41.41	68.50	63.23	20.00	44.70	16.50	25.92	26.21	24.20	34.34	39.89
SeaLong	37.50	40.52	68.00	67.07	19.33	44.12	16.00	26.45	27.06	32.06	33.58	42.04
Logo	43.50	46.58	63.00	61.74	19.00	44.39	18.50	28.05	26.67	32.69	34.13	42.69
LongFaith	44.00	49.23	75.50	**72.80**	20.66	48.10	9.50	23.26	24.30	22.35	34.79	43.15
LongMab	**50.00**	**52.15**	**76.00**	68.60	**26.00**	**51.26**	**20.00**	**28.61**	**32.76**	**36.55**	**40.95**	**47.43**
Qwen-2.5-7B-Instruct												
Vanilla LLM	33.50	30.52	58.00	50.14	28.00	45.12	15.00	18.29	25.64	22.38	32.03	33.29
LongAlpaca	32.50	33.90	55.50	50.75	24.66	45.88	17.50	19.72	29.34	23.82	31.90	34.81
LongAlign	28.50	31.09	52.00	52.02	23.33	**49.76**	15.50	21.98	25.07	25.61	28.88	36.09
LongMab-SFT	36.50	39.32	59.50	58.34	21.33	44.06	15.50	22.17	26.29	28.01	31.82	38.38
LongReward	37.50	33.37	62.00	50.75	**29.33**	44.89	15.50	17.67	28.20	21.77	34.51	33.69
SeaLong	43.00	22.09	67.00	36.39	27.33	41.08	**18.00**	16.27	28.20	16.95	36.71	26.56
Logo	41.00	36.48	63.00	57.20	24.67	49.18	15.50	18.28	28.00	28.47	34.43	37.92
LongFaith	**48.50**	**43.38**	66.00	53.93	24.67	38.08	12.00	17.03	23.07	18.55	34.85	34.19
LongMab	44.00	43.25	**67.50**	**62.97**	25.33	48.07	**18.00**	**25.14**	**30.48**	**31.88**	**37.06**	**42.26**

Table 1: Overall performance of Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct on different long-context understanding tasks. The **best** and second best results are highlighted.

and train each model for 2 epochs. To ensure training efficiency, we leverage Low-Rank Adaptation (LoRA) (Hu et al., 2022) and the LLaMA Factory framework. During the sampling phase, we configure the maximum number of rollout steps (T) to 30. In each step, we select $K = 4$ context chunks, with each chunk comprising 1,500 tokens. More experimental details are provided in Appendix A.3.

5 Experiment Results

In this section, we first present the overall performance of LongMab, followed by ablation studies to evaluate the contribution of its components. Subsequently, we analyze the selected chunks and generated responses by LongMab.

5.1 Overall Performance

Table 1 presents the overall performance of LongMab against baseline methods across various long-context understanding tasks.

Our experimental results demonstrate the effectiveness of LongMab-PO in enhancing the long-context understanding capabilities of LLMs by consistently outperforms all baselines. Notably, although trained exclusively on the MuSiQue dataset, LongMab-PO exhibits strong generalization ability, maintaining consistent gains across multiple out-of-domain tasks. Among different optimization strategies, SFT-based methods, such as LongAlpaca (Chen et al., 2023) and LongAlign (Bai et al., 2024a), show identical performance compared with the Vanilla LLM. This suggests that simply overfitting to ground-truth answers is insufficient for enabling LLMs to effectively identify and extract salient information from long contexts. In contrast, DPO-based methods generally outperform SFT-based ones in improving long-context understanding. Furthermore, LongMab-PO achieves an improvement of over 4% compared to other DPO-based approaches, which can be attributed to its Multi-Armed Bandit–guided mechanism. This mechanism facilitates the generation of higher-quality and more informative preference data, thereby enhancing the capbility of LLMs to fully leverage information within extended contexts during the DPO training stage.

5.2 Ablation Studies

In this section, we first evaluate the role of different components involved in chunk sampling of our LongMab, and then investigate the impact of the number of selected chunks (K) during sampling.

As shown in Table 2, we analyze the contribution of each component in LongMab. First, LongMab (w/o Init.) removes the probability initialization step by setting all initial chunk rewards to zero. To further examine the impact of the UCB score in chunk sampling, we design three variants: LongMab (w/o UCB), LongMab (Rejected w/o UCB), and LongMab (Chosen w/o UCB). In LongMab (w/o UCB), chunks are sampled randomly to construct preference pairs for DPO training, without considering the UCB score. Similarly, LongMab (Rejected w/o UCB) and LongMab (Chosen w/o UCB) randomly select chunks to obtain the rejected and chosen responses, respectively.

Model	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
Llama-3.1-8B-Instruct												
Direct-PO	46.50	50.82	62.00	58.74	21.33	47.12	19.50	31.45	30.19	**37.31**	35.90	45.09
LongMab (w/o Init.)	43.50	49.59	71.50	**68.66**	23.33	46.28	17.00	29.39	29.71	33.46	37.01	45.48
LongMab (w/o UCB)	42.00	47.62	69.00	65.28	20.67	47.00	19.50	27.95	27.71	33.11	35.78	44.19
LongMab	**50.00**	**52.15**	**76.00**	68.60	**26.00**	**51.26**	20.00	28.61	**32.76**	36.55	**40.95**	**47.43**
Rejected w/o UCB	43.50	45.16	73.50	63.63	24.67	48.05	18.50	26.66	29.14	31.15	37.86	42.93
Chosen w/o UCB	42.50	48.84	66.00	66.07	16.67	44.32	**21.00**	**32.57**	28.57	36.84	34.95	45.73
Qwen-2.5-7B-Instruct												
Direct-PO	38.50	41.17	59.00	55.46	21.33	45.24	**19.00**	**25.21**	27.43	30.25	33.05	39.47
LongMab (w/o Init.)	**44.50**	44.06	65.00	56.57	25.00	47.88	18.50	21.53	**31.05**	29.70	36.81	39.95
LongMab (w/o UCB)	41.00	41.29	62.50	54.53	**25.33**	47.01	16.00	21.69	28.86	29.30	34.74	38.76
LongMab	44.00	43.25	**67.50**	**62.97**	**25.33**	**48.07**	18.00	25.14	30.48	**31.88**	**37.06**	**42.26**
Rejected w/o UCB	42.50	**45.33**	60.00	56.50	24.00	45.93	18.50	24.10	27.43	29.47	34.49	40.27
Chosen w/o UCB	41.50	43.39	62.50	57.05	21.33	43.45	17.50	21.65	29.43	30.12	34.45	39.13

Table 2: Ablation study results on various long-context understanding tasks. The results demonstrate the effectivess of UCB-guided sampling strategy and probe-based probability initialization.

K	MuSiQue		MFQA-En		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1
1	44.00	46.81	21.33	49.09	38.07	44.85
2	48.00	47.55	**26.00**	50.29	39.53	44.81
3	47.50	48.05	23.33	49.00	38.71	44.89
4	**50.00**	**52.15**	**26.00**	**51.26**	**40.95**	**47.43**
5	47.00	48.65	22.66	49.90	38.30	45.40

Table 3: Impact of the number of selected chunks (K) in LongMab. We present results for two representative datasets and the average (Avg.) performance across all five datasets. Comprehensive results for all datasets are provided in Appendix A.5.

Compared with LongMab (w/o Init.), the full LongMab achieves over a 2% improvement in F1 score, demonstrating the effectiveness of the evidence probe in providing a more informative prior probability to initialize the expected reward. We then explore the role of the UCB score in chunk sampling. The LongMab (w/o UCB) model exhibits performance comparable to the Direct-PO baseline, indicating that merely splitting long contexts into chunks does not enhance DPO training. Moreover, removing UCB-based sampling when selecting chosen or rejected responses leads to performance degradation, confirming the importance of UCB in guiding the selection of informative chunks for preference pair construction. Notably, LongMab (Rejected w/o UCB) consistently outperforms LongMab (w/o UCB), further illustrating that the UCB score facilitates the selection of higher-quality chosen responses, which benefits DPO optimization.

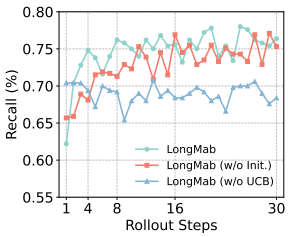
Next, we analyze the model's sensitivity to the number of selected chunks (K), a key hyperparameter in our MAB-guided sampling framework. We vary K from 1 to 5 and evaluate the model's performance under each setting. As shown in Table 3, the results reveal a clear trend: performance improves as K increases, peaking at $K = 4$. This suggests that selecting too few chunks provides insufficient evidence for generating high-quality responses. However, performance slightly declines when $K = 5$, indicating that $K = 4$ strikes an optimal balance between evidence coverage and noise control. Selecting too many chunks tends to introduce irrelevant or noisy information, thereby reducing the effectiveness in constructing preferece pairs for DPO training.

5.3 Mechanism Behind LongMab in Enabling Long-Context LLMs

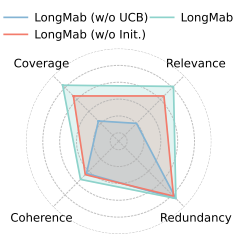
This section provides a series of in-depth analyses to uncover the factors behind the superior performance of LongMab in enabling long-context LLMs. We compare two baseline variants: LongMab (w/o UCB) and LongMab (w/o Init.).

Characteristics of LongMab-Selected Chunks. As illustrated in Figure 3, we conduct experiments to investigate the effectiveness of the UCB-guided sampling strategy for chunk selection.

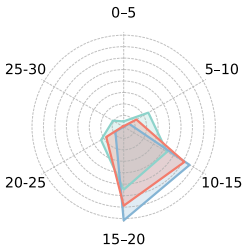
We first evaluate the quality of the selected chunks. To this end, we monitor the ground-truth recall throughout the rollout process. As shown in Figure 3(a), the recall of LongMab steadily increases and then stabilizes, whereas the recall of the baseline LongMab (w/o UCB) remains flat. This indicates that the UCB mechanism effectively guides the selection process toward more informative content. To further assess quality, we employ GLM-4-Plus (Du et al., 2021) as a judge to rate the final chunk sets along four dimensions: coverage, relevance, redundancy, and coherence. As shown in Figure 3(b), LongMab consistently outperforms all baselines across all dimensions. Notably, LongMab achieves higher scores in both coverage and rele-



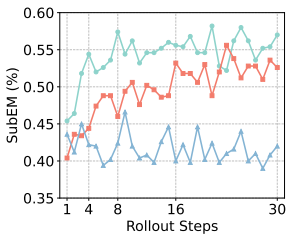
(a) Ground truth recall.



(b) GLM score.



(c) Chunk frequency.



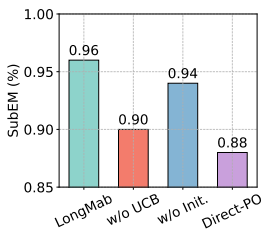
(d) SubEM of responses.

Figure 3: Characteristics of chunk sets obtained by different methods. Figure 3(a) shows the ground-truth recall during the rollout process. Figure 3(b) reports GLM-4-Plus scores for the chunk sets. Figure 3(c) illustrates the distribution of chunk selection frequencies, grouping chunks by their selection counts across all rollouts. Finally, Figure 3(d) presents the trend of SubEM scores over the rollout process.

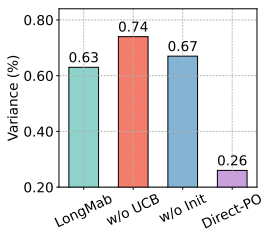
vance, underscoring its capability to form informative and well-structured chunk combinations.

Next, we analyze the selection mechanism to understand how UCB enhances chunk quality. To verify this, we examine the distribution of chunk selection frequencies shown in Figure 3(c). The baseline model exhibits a narrow and uniform distribution, suggesting indiscriminate sampling. In contrast, LongMab produces a broader and more polarized distribution, indicating that it simultaneously explores a diverse range of chunks while prioritizing high-value ones and suppressing less relevant options. The effectiveness of LongMab in chunk selection is further validated by the steadily increasing SubEM scores during rollout (Figure 3(d)), confirming that its improved selection directly lead to higher-quality responses.

Preference Pairs Analysis. We next examine whether this advantage translates into constructing more effective preference data. Specifically, we aim to assess both the quality of the chosen responses and the diversity of all generated responses. As shown in Figure 4, we compare LongMab with three baselines: LongMab (w/o UCB), LongMab (w/o Init.), and Direct-PO.



(a) Average chosen SubEM.



(b) Variance of responses.

Figure 4: Analysis of responses generated by different methods. Figure 4(a) shows the average SubEM score of the selected responses, while Figure 4(b) depicts the variance of pairwise similarities among all responses in the sampled response set.

We first analyze the quality of the chosen responses. Figure 4(a) reports the SubEM scores of the selected responses from different models. The results show that LongMab achieves over a 6% improvement compared to both LongMab (w/o UCB) and Direct-PO, demonstrating its effectiveness in generating higher-quality preferred responses. Next, we evaluate the diversity of all sampled responses, as shown in Figure 4(b). We use the MiniCPM-Embedding model (Hu et al., 2024) to encode all generated responses and compute the average variance of pairwise similarities. LongMab yields substantially higher variance scores than Direct-PO, while maintaining a comparable level of diversity to the random sampling variant LongMab (w/o UCB). These results suggest that LongMab not only improves response quality but also preserves diversity, enabling the construction of more informative preference pairs for DPO training.

6 Conclusion

This paper introduces LongMab, an innovative framework that applies a Multi-Armed Bandit (MAB) approach to optimize long-context language models. Specifically, LongMab leverages an iterative MAB rollout process to identify and select optimal combinations of context chunks, enabling the sampling of higher-quality and more diverse responses. These responses are then used to construct preference data for DPO training. As a result, the aligned model exhibits a superior ability to pinpoint key information within long contexts, leading to more accurate reasoning. Our experiments show that LongMab substantially outperforms strong baseline methods across numerous long-context understanding tasks.

Limitation

While LongMab demonstrates substantial improvements in long-context reasoning and alignment tasks, several limitations remain. First, although the UCB-guided sampling strategy effectively balances exploration and exploitation during chunk selection, its computational efficiency in extremely long contexts warrants additional optimization, particularly when a large number of rollout steps is used. Next, our experiments are conducted primarily on models around the 7B parameter scale, and the generalization of LongMab to larger LLMs remains to be verified. Finally, due to computational constraints, the synthesized instruction data used in our experiments are limited to lengths of 8K–16K tokens, and exploring the effectiveness of LongMab on even longer synthetic contexts remains an open question for future work.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024b. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Guanzheng Chen, Xin Li, Michael Qizhe Shieh, and Lidong Bing. 2025a. Longpo: Long context self-evolution of large language models through short-to-long preference optimization. *arXiv preprint arXiv:2502.13922*.

Hao Chen, Yukun Yan, Sen Mei, Wanxiang Che, Zhenghao Liu, Qi Shi, Xinze Li, Yuchun Fan, Pengcheng Huang, Qiushi Xiong, and 1 others. 2025b. Clueanchor: Clue-anchored knowledge reasoning exploration and optimization for retrieval-augmented generation. *arXiv preprint arXiv:2505.24388*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Wanxiang Che, Hang Yan, Kai Chen, and Dahua Lin. 2024. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. *arXiv preprint arXiv:2409.01893*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

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

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

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? *arXiv preprint arXiv:2404.06654*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Junpei Komiyama, Edouard Fouché, and Junya Honda. 2024. Finite-time analysis of globally nonstationary multi-armed bandits. *Journal of Machine Learning Research*, 25(112):1–56.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024a. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, and 1 others. 2025. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. In *The Thirteenth International Conference on Learning Representations*.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024b. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Xinyu Liu, Runsong Zhao, Pengcheng Huang, Chunyang Xiao, Bei Li, Jingang Wang, Tong Xiao, and Jingbo Zhu. 2024. Forgetting curve: A reliable method for evaluating memorization capability for long-context models. *arXiv preprint arXiv:2410.04727*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Huashan Sun, Shengyi Liao, Yansen Han, Yu Bai, Yang Gao, Cheng Fu, Weizhou Shen, Fanqi Wan, Ming Yan, Ji Zhang, and 1 others. 2025. Solopo: Unlocking long-context capabilities in llms via short-to-long preference optimization. *arXiv preprint arXiv:2505.11166*.

Zecheng Tang, Zechen Sun, Juntao Li, Qiaoming Zhu, and Min Zhang. 2024. Logo–long context alignment via efficient preference optimization. *arXiv preprint arXiv:2410.18533*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel Bowman. 2022a. Squality: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156.

Meiyun Wang, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025a. Lost in the distance: Large language models struggle to capture long-distance relational knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4536–4544.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. 2025b. Clusterucb: Efficient gradient-based data selection for targeted fine-tuning of llms. *arXiv preprint arXiv:2506.10288*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, and 1 others. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

Zhen Xu, Shang Zhu, Jue Wang, Junlin Wang, Ben Athiwaratkun, Chi Wang, James Zou, and Ce Zhang. 2025. When does divide and conquer work for long context llm? a noise decomposition framework. *arXiv preprint arXiv:2506.16411*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. *arXiv preprint arXiv:2502.12583*.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng,

and Juanzi Li. 2024a. Longreward: Improving long-context large language models with ai feedback. *arXiv preprint arXiv:2410.21252*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and 1 others. 2024b. Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Runsong Zhao, Xin Liu, Xinyu Liu, Pengcheng Huang, Chunyang Xiao, Tong Xiao, and Jingbo Zhu. 2024. Position ids matter: An enhanced position layout for efficient context compression in large language models. *arXiv preprint arXiv:2409.14364*.

Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. 2025. Chain-of-thought matters: improving long-context language models with reasoning path supervision. *arXiv preprint arXiv:2502.20790*.

A Appendix

A.1 Licence

This section summarizes the licenses of the datasets used in our experiments.

All of these datasets under their respective licenses and agreements allow for academic use: MuSiQue (CC-BY-4.0 license); 2WikiMultihopQA, MultiFieldQA-En, NarrativeQA, LongBench (Apache 2.0 license); En.QA, InfiniteBench (MIT license).

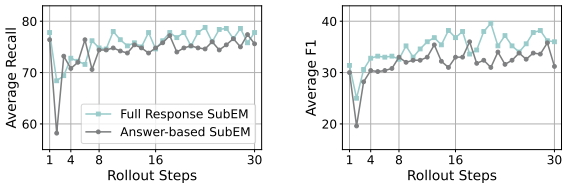
A.2 Impact of Reward Design on Multi-Armed Bandit Rollout Process

In this section, we investigate the impact of different reward designs in the multi-armed bandit rollout process on the quality of sampled responses. Specifically, we compare two different reward calculation strategies: the full response subem and the answer-based SubEM strategy. The first strategy computes the SubEM score using the entire generated response, while the second extracts the answer string from the response and then calculates the SubEM score. In both cases, the final reward of sampled response is computed by averaging the subem score with the F1 score of the extracted answer. We randomly sampled 500 examples for this experiment.

As shown in Figure 5, we present the trend of ground truth chunk recall and F1 score of generated responses during the rollout process. The full response SubEM strategy consistently outperforms the answer-based SubEM strategy on the recall and F1 score. This phenomenon indicates that overly strict reward formulation limits the LLM's exploration of diverse chunk combinations, thereby restricting the quality of sampled responses. To further evaluate the effectiveness of these two reward calculation strategies, we train LLMs using responses sampled under each strategy and evaluate them on downstream tasks. As shown in Table 5, LLMs trained with the answer-based SubEM strategy consistently underperform compared to those trained with the full response SubEM strategy across multiple datasets. These results further illustrate that a decrease in sampling quality will hinder the effectiveness of DPO training.

A.3 Additional Experimental Details

In this section, we introduce additional experimental details in our experiments. In our experiments, all models use the VLLM framework during the inference stage, with the same inference settings and random seed. This ensures that all models' evaluation results in the paper are reproducible. Furthermore, during the training process, we use a unified default random seed, typically set to 42. Besides, the prompt used in the LongMab is shown in Figure 6. All experiments were conducted on NVIDIA A100 GPUs with 80GB of memory.



(a) Ground truth recall. (b) Average f1 score.

Figure 5: Impact of different reward design to rollout process. The results are based on Llama-3.1-8B-Instruct.

Dataset	#Tokens	#Samples
MuSiQue	15.5k	200
2WikiMultihopQA	7.1k	200
MultiFieldQA-En	6.9k	150
NarrativeQA	29.8k	200
En.QA	192.6k	351

Table 4: The statistics of testsets from two benchmarks.

A.4 Statistics of Testsets

Table 4 shows the details of testsets from LongBench and InfiniteBench.

A.5 Additional Experiments

To evaluate the robustness of our framework, we conduct experiments with 4, 8, 16, and 30 rollout steps. At each setting, we construct the corresponding preference pairs for training. As shown in table 6, performance consistently improves across almost all datasets as the number of rollouts increases, indicating that our method is capable to progressively sample higher-quality preference pairs throughout the rollout process. All experiments are implemented on Llama-3.1-8B-Intruct, with the selected chunks K set to 4.

In addition, we conduct a top-k analysis to examine how the number of candidate chunks considered at each selection step affects response quality. As shown in Table 7, performance initially improves with increasing k, reaching a peak at $k = 4$, and then gradually declines, suggesting that a moderate number of candidates strikes the best balance between selection diversity and relevance, thereby

Instruction for Evaluation

[Instruction]: You are provided with a long document and a complex logical reasoning question. Read the document and follow my instructions to process it.
Document: {context}
Question: {question}
Instructions:
#####
1. Provide a reasoning process: You should first understand the complex problem and make a plan to solve it, then carry out the plan and solve the problem step-by-step and finally deduce the answer. You could perform reasoning with reflecting, verifying, and revising when encountering uncertain or contradictory information.
2. Provide an answer: Based on your reasoning process, give a short answer to the question. Your answer should be concise and do not include any reasoning process.
#####
Your output should follow this format:
Reasoning: reasoning process here.
Answer: answer here

Instruction for Probe-based Evidence Extraction

[Instruction]: You are provided with a long document, a complex logical reasoning question, and the correct answer. Your task is to read the document and perform step-by-step reasoning and finally reach the correct answer.
Instructions:
1.Each reasoning step should explicitly refer to the document.
2.End your reasoning with `The answer is` followed by the correct answer.
Document: {context}
Question: {question}
Answer: {answer}

Figure 6: Prompt template used in LongMab.

maximizing the quality of constructed preference pairs.

To validate the generality of LongMab, we conduct experiments on Qwen-3-8B with the same rollout steps and chunk selection size as before. As shown in Table 8, LongMab consistently outperforms baselines on long-context reasoning tasks, demonstrating that its UCB-guided chunk selection and preference pair construction effectively generalize across model architectures.

A.6 Case Study

In this section, we randomly select a case from the MuSiQue evaluation set to demonstrate the effectiveness of LongMab in capturing key information from long contexts and mitigating the "lost-in-the-middle" problem.

As shown in Table 9, we present a two-hop question in the MuSiQue dataset. To answer it, the LLM must first identify which province the Lago District belongs to, and then determine which provinces border it based on the context. LongFaith-PO is misled by noisy information near the two ends of the long context and fails to recognize the key evidence appearing in the middle of the long context (specifically, around the 45% mark of the context), leading to an incorrect answer. In contrast, LongMab enhances the LLM's ability to identify and integrate evidence scattered across different parts of the context, enabling more accurate multi-hop reasoning and ultimately yielding the correct answer.

Strategy	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
Llama-3.1-8B-Instruct												
Full Response	**50.00**	**52.15**	**76.00**	68.60	**26.00**	**51.26**	**20.00**	**28.61**	**32.76**	**36.55**	**40.95**	**47.43**
Answer-based	45.50	48.58	72.50	**69.43**	20.00	47.39	19.00	28.60	28.57	31.00	37.11	45.00
Qwen-2.5-7B-Instruct												
Full Response	**44.00**	43.25	**67.50**	**62.97**	**25.33**	**48.07**	18.00	25.14	**30.48**	**31.88**	**37.06**	**42.26**
Answer-based	**44.00**	**44.42**	63.00	56.85	20.66	46.10	**20.50**	**25.36**	29.62	30.73	35.56	40.69

Table 5: Performance of models trained with preference data created under different reward strategies.

Rollout Steps (T)	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
VallinaLLM	33.50	36.56	69.50	65.45	18.66	44.34	18.00	26.61	26.49	32.19	33.23	41.03
T=4	37.50	45.17	70.50	67.84	24.00	50.56	19.50	31.04	28.57	37.61	36.01	46.44
T=8	45.50	47.75	75.00	67.95	23.33	49.49	20.00	27.52	29.71	33.40	38.71	45.22
T=16	48.50	51.88	75.50	66.32	24.00	50.84	19.00	27.66	30.00	32.91	39.40	45.92
T=30	50.00	52.15	76.00	68.60	26.00	51.26	20.00	28.61	32.76	36.55	40.95	47.43

Table 6: Scaling the rollout step of LongMab. Performance consistently increases across all testsets.

Model	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
$K = 1$	44.00	46.81	**76.50**	64.81	21.33	49.09	18.50	28.40	30.00	<u>35.16</u>	38.07	44.85
$K = 2$	<u>48.00</u>	47.55	73.00	61.73	**26.00**	<u>50.29</u>	<u>19.50</u>	**29.81**	<u>31.14</u>	34.68	<u>39.53</u>	44.81
$K = 3$	47.50	48.05	75.50	67.61	<u>23.33</u>	49.00	17.00	25.71	30.20	34.06	38.71	44.89
$K = 4$	**50.00**	**52.15**	<u>76.00</u>	**68.60**	**26.00**	**51.26**	**20.00**	<u>28.61</u>	**32.76**	**36.55**	**40.95**	**47.43**
$K = 5$	47.00	<u>48.65</u>	73.50	<u>68.58</u>	22.66	49.90	19.00	26.94	29.36	32.93	38.30	<u>45.40</u>

Table 7: Impact of the number of selected chunks in LongMab. K denotes the number of selected chunks. The **best** and <u>second best</u> results are highlighted.

Model	MuSiQue		2WikiMQA		MFQA-En		NarrativeQA		En.QA		Avg.	
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
Qwen3-8B												
Vanilla LLM	41.50	44.24	77.00	72.40	23.33	47.07	19.00	26.26	30.29	31.35	38.22	44.26
LongAlpaca	36.50	38.06	71.00	68.76	23.33	47.73	15.00	22.31	25.43	27.69	34.25	40.91
LongAlign	29.00	28.76	62.00	54.34	27.33	50.29	15.00	22.89	24.86	26.86	31.64	36.63
LongMab-SFT	41.50	45.62	75.50	73.16	22.67	47.99	17.00	25.86	29.43	32.36	37.22	45.00
LongReward	40.50	41.94	76.00	70.08	24.00	45.48	17.00	23.13	28.00	29.01	37.10	41.93
SeaLong	42.50	44.11	77.00	73.38	22.67	48.56	18.00	25.15	32.86	33.54	38.61	44.95
Logo	41.00	44.98	75.50	71.34	23.33	47.32	17.00	24.48	30.29	30.75	37.42	43.77
LongFaith	41.00	46.35	66.00	66.79	17.33	38.01	11.50	18.31	24.29	28.08	32.02	39.51
LongMab	48.00	51.02	77.50	74.32	23.33	47.57	18.50	27.34	31.43	33.84	39.75	46.82

Table 8: Overall performance of Qwen3-8B on different long-context understanding tasks. The **best** and <u>second best</u> results are highlighted.

Query: What province shares a border with the province where Lago District is located?
Ground Truth: Cabo Delgado Province
Document: . . .
Passage 7: Niassa Reserve
. . . Niassa Reserve is a nature reserve in Cabo Delgado Province and Niassa Province, Mozambique. Covering over 42,000 square kilometres (10,000,000 acres), it is the largest protected area in the country...
Passage 12: Lago District
. . . Lago District is a district of Niassa Province in north-western Mozambique. The principal town is Metangula. Lago District is bordered to the west by Lake Nyasa, on the south by Lichinga, on the east by Sanga District. Lago shares its northern boundary with Tanzania...

LongFaith:

To determine which province shares a border with the province where Lago District is located, we need to analyze the information provided in the documents. First, we identify the location of Lago District. According to the document, Lago District is located in Niassa Province in northwestern Mozambique [12]. Next, we look for provinces that share a border with Niassa Province. The document states that Lago District is bordered to the west by Lake Nyasa, which implies that it shares a border with the province that includes Lake Nyasa. Lake Nyasa is located in Malawi, which is a neighboring country to Mozambique. Therefore, the province that shares a border with Niassa Province is Malawi. Thus, the answer to the question is Malawi. The answer is Malawi.

LongMab:

Reasoning: To find the province that shares a border with the province where Lago District is located, we need to identify the location of Lago District first. According to Passage 12, Lago District is a district of Niassa Province in north-western Mozambique. Next, we need to determine which provinces share a border with Niassa Province. Since Lago District is located in north-western Mozambique, it is likely to share borders with provinces in the neighboring countries or regions. In Passage 7, it is mentioned that Niassa Reserve is located in Cabo Delgado Province and Niassa Province, Mozambique. This indicates that Niassa Province shares a border with Cabo Delgado Province. Based on this information, we can conclude that the province that shares a border with the province where Lago District is located is likely to be Cabo Delgado Province. Answer: Cabo Delgado Province.

Table 9: Case study. We randomly sample an example from the MuSiQue dataset to analyze model behavior. Different colors are used to represent noise information or key evidence. Specifically, green indicates critical evidence supporting the ground truth, while pink represents distracting information.