LONGPO: LONG CONTEXT SELF-EVOLUTION OF LARGE LANGUAGE MODELS THROUGH SHORT-TO-LONG PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable capabilities through pretraining and alignment. However, superior short-context LLMs may underperform in long-context scenarios due to insufficient long-context alignment. This alignment process remains challenging due to the impracticality of human annotation for extended contexts and the difficulty in balancing short- and long-context performance. To address these challenges, we introduce LongPO, that enables short-context LLMs to self-evolve to excel on long-context tasks by internally transfer short-context capabilities. LongPO harnesses LLMs to learn from self-generated short-to-long preference data, comprising paired responses generated for identical instructions with long-context inputs and their compressed short-context counterparts, respectively. This preference reveals capabilities and potentials of LLMs cultivated during short-context alignment that may be diminished in under-aligned long-context scenarios. Additionally, LongPO incorporates a short-to-long KL constraint to mitigate short-context performance decline during long-context alignment. When applied to Mistral-7B-Instruct-v0.2 from 128K to 256K context length, LongPO fully retaining short-context performance and largely outperforms naive SFT and DPO in both long- and short-context tasks. Specifically, LongPO-trained models can achieve results on long-context benchmarks comparable to, or even surpassing, those of superior LLMs (e.g., GPT-4-128K) that involve extensive long-context annotation and larger parameter scales.

1 INTRODUCTION

Recent advancements in Large Language Models (LLMs) have revealed remarkable capabilities through extensive pretraining and subsequent alignment with human intentions. The alignment process, including methods such as Supervised Fine-Tuning (SFT) (Wei et al., 2022), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020), has effectively unleashed the potential of LLMs acquired during pretraining to achieve desired behaviors.

Although off-the-shelf alignment methods have made significant strides in short-context settings, their application to long-context situations remains challenging (Bai et al., 2024). First, the scarcity of high-quality, long-context annotated data poses a significant hurdle. Human annotation becomes impractical and less-reliable as context length increases (Dubey et al., 2024), while synthetic data generation using advanced LLMs lacks scalability and remains resource-intensive. Moreover, simply concatenating existing short-context datasets has been shown to yield unsatisfactory long-context performance (Liu et al., 2024b). Second, long-context alignment methods grapple with the balance between preserving short-context proficiency and cultivating long-context capabilities (Liu et al., 2024b). For instance, the LLaMA-3.1 series incorporate merely 0.1% long-context data with over 99% short-context data during alignment to maintain the short-context performance (Liu et al., 2024b). This limited exposure to natural long-context data may result in insufficient alignment, potentially blocking the intrinsic long-context capabilities in LLMs.

The challenges of long-context alignment suggest that the full potential of LLMs may remain untapped for long-context tasks. As illustrated in Figure 1, even superior models such as GPT-4,

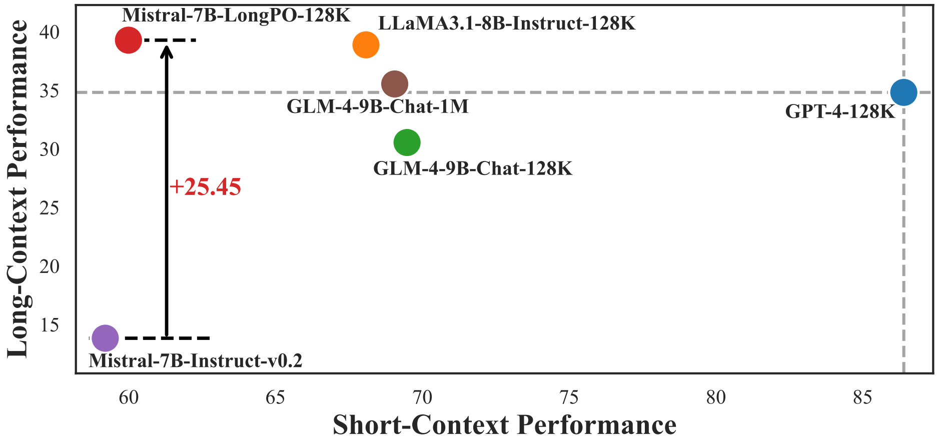


Figure 1: The comparison of long-context (InfiniteBench) and short-context (MMLU) performance among GPT-4-128K and smaller LLMs.

which excel in short-context tasks, unexpectedly underperform in long-context scenarios. Interestingly, despite the much stronger short-context capabilities, GPT-4 is still inferior to LLaMA3.1-8B on long-context tasks. This disparity underscores the need for more effective long-context alignment methods to fully unleash the intrinsic power of LLMs across variable context lengths.

In this work, we posit that the capabilities, deeply ingrained during short-context pretraining and alignment, can be effectively transferred to longer contexts without external guidance. To this end, we introduce Short-to-**Long** **P**reference **O**ptimization (**LongPO**), to steer long-context alignment by injecting internal short-context preferences into long-context scenarios. Specifically, we propose to construct the preference data pairs by prompting the short-context LLM (e.g., Mistral-Instruct) with two inputs: (1) a long input comprising an instruction over a long document and, (2) a short input with the identical instruction over the relevant shortened chunk within the same document. We then designate the responses to short and long inputs as chosen and rejected responses, respectively. The short-to-long preference, i.e., the discrepancies between each paired response, reveal the capabilities and potentials cultivated during short-context alignment that may be diminished in under-aligned long-context scenarios. In order to bring forward the established capabilities, LongPO is utilized to optimize the model towards short-to-long preferences using DPO-style objectives upon long contexts. Furthermore, to maintain the short-context performance, we incorporate a *short-to-long constraint* in LongPO by applying Kullback-Leibler (KL) divergence between the response distributions to short and long inputs, respectively. This constraint, inspired by the KL constraint in RLHF (Ouyang et al., 2022; Stiennon et al., 2020), guides the policy model to minimize the deviation from its short-context output distribution when giving the long context during training. We found that this straightforward constraint largely enhances the retention of short-context performance after the long-context alignment.

We apply LongPO to Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and iteratively extend its context length from 32K to 256K, with the self-generated short-to-long preference data only. The experimental results demonstrate that LongPO, as a long-context alignment method, surpasses naive SFT and DPO by large margins (over 10 points) in both long- and short-context tasks. Notably, LongPO fully retains the performance of short-context LLMs after long-context alignment, whereas SFT and DPO yield substantial performance degradation (10~20 points on most tasks). In terms of long-context performance, LongPO largely improves the Mistral-7B-Instruct-v0.2 by 25.45 points on InfiniteBench. Specifically, as depicted in Figure 1, the resulting model is comparable with superior long-context LLMs at various scales (e.g., Mistral-7B-LongPO-128K of 39.27 vs. GPT-4-128K of 34.81 on InfiniteBench), despite the latter often involving extensive continual training on hundreds of billions of tokens (Dubey et al., 2024) or labor-intensive long-context data annotation (Zeng et al., 2024). These findings underscore the efficacy of our proposed method in addressing the challenges of long-context alignment while simultaneously preserving short-context capabilities, offering a more efficient and balanced approach to the development of long-context LLMs.

2 PRELIMINARIES

In this section, we introduce two key methods for aligning language models with human preferences: Reinforcement Learning from Human Feedback (RLHF, §2.1) and Direct Preference Optimization (DPO, §2.2).

2.1 RLHF

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020) aims to optimize the policy model π_θ to maximize rewards while maintaining proximity to a reference policy π_{ref}. Formally, the objective is

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\big[r_\phi(x,y)\big] - \beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y\mid x)\parallel\pi_{\text{ref}}(y\mid x)\big], \tag{1}$$

where r_ϕ is the reward model that has been trained on ranked responses to reflect human preference, β is a hyper-parameter controlling the deviation from reference policy, and $\mathbb{D}_{\text{KL}}$ denotes the Kullback-Leibler divergence. Typically, both π_θ and π_{ref} are initialized with identical model.

2.2 DPO

Considering the instability and difficulty of RLHF training, DPO (Rafailov et al., 2023) offers an alternative approach by reparameterizing the reward function r that incorporates the optimal policy:

$$r(x,y) = \beta\log\frac{\pi_\theta(y\mid x)}{\pi_{\text{ref}}(y\mid x)} + \beta\log Z(x), \tag{2}$$

where $Z(x)$ is the partition function. DPO assumes access to preference data $\mathcal{D}$, which consists of paired responses (y_w, y_l) to an instruction x. Specifically, the y_w and y_l represent the preferred (winning) and dispreferred (losing) responses, respectively, based on human preference. Inspired by the Bradley-Terry (BT) theory that models the preference distribution p^* by

$$p^*(y_w \succ y_l\mid x) = \sigma(r(x,y_w) - r(x,y_l)), \tag{3}$$

where σ is the sigmoid function. DPO derives the preference optimization objective for the policy model π_θ as

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\sigma(r_\theta(x,y_w) - r_\theta(x,y_l))\right]\\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right)\right].\end{aligned} \tag{4}$$

3 LONGPO: SHORT-TO-LONG PREFERENCE OPTIMIZATION

Motivated by the challenges of data annotation and performance balance during long-context alignment, we introduce the Short-to-**Long** **P**reference **O**ptimization (**LongPO**), to effectively empowers a short-context LLM self-evolve to a long-context counterpart while preserving its original short-context capabilities. The foundation of LongPO lies in the transfer of capabilities deeply ingrained during short-context alignment to long-context scenarios by learning from short-to-long preference (§3.1). Additionally, LongPO incorporates a short-to-long constraint based on the KL divergence between short- and long-context models during training, to maintain the short-context performance in a simple yet effective way (§3.2). In §3.3, we present the details of curating short-to-long preference data without external guidance and self-evolving long context training process using LongPO.

3.1 LEARNING FROM SHORT-TO-LONG PREFERENCE

As outlined in §2, aligning LLMs with human preference typically relies on datasets comprising ranked responses to identical prompts or instructions. However, in long-context scenarios, constructing such datasets becomes impractical due to the extensive effort required for annotation. To circumvent the external data annotation, we leverage the *short-to-long preference* to internally transfer capabilities well-established in the short-context alignment of LLMs to long-context counterpart.

Concretely, we assume access solely to a short-context LLM π_{S} that has been well aligned. Given a long input $x_{\text{L}} = [C_{\text{L}}; I_{\text{L}}]$ where C_{L} is the long context and I_{L} is the instruction, we can acquire the response $y_{\text{L}} \sim \pi_{\text{S}}(y \mid x_{\text{L}})$ by conditioning on the entire context. Due to the limitations of π_{S} in handling long contexts, y_{L} is likely to be of lower quality.

We then hypothesize an ideal extractor $\mathcal{F}$ that can rigorously identify and extract all essential information C_{S} within C_{L} relevant to addressing I_{L}:

$$C_{\text{S}} = \mathcal{F}(C_{\text{L}}, I_{\text{L}}). \tag{5}$$

By querying the instruction I_{L} based on C_{S}, we obtain a new answer $y_{\text{S}} \sim \pi_{\text{S}}(y \mid x_{\text{S}})$, where $x_{\text{S}} = [C_{\text{S}}; I_{\text{L}}]$. As C_{S} is a shortened context for I_{L}, the well-aligned short-context model π_{S} should be capable of producing a high-quality answer that aligns with human preferences.

Intuitively, y_{S} can serve as a high-quality answer even when giving the whole long context, as its conditioned context is self-contained for instruction I_{L}. Hence, we definite the short-to-long preference distribution p^{SL} based on Bradley-Terry (BT) model following Eq. (3):

$$p^{\text{SL}}(y_{\text{S}} \succ y_{\text{L}} \mid x_{\text{L}}) = \sigma(r(x_{\text{L}}, y_{\text{S}}) - r(x_{\text{L}}, y_{\text{L}})). \tag{6}$$

We now steer a policy model π_θ (initialized with π_{S}) to follow the preference distribution p^{SL}, forming the LongPO objective:

$$\mathcal{L}_{\text{LongPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x_{\text{S}}, x_{\text{L}}, y_{\text{S}}, y_{\text{L}}) \sim \mathcal{D}^{\text{SL}}} \left[\sigma(r_\theta(x_{\text{L}}, y_{\text{S}}) - r_\theta(x_{\text{L}}, y_{\text{L}}))\right], \tag{7}$$

where $\mathcal{D}^{\text{SL}}$ is the short-to-long preference data consisting of quadruples $(x_{\text{S}}, x_{\text{L}}, y_{\text{S}}, y_{\text{L}})$. This objective encourages the policy model to consistently accommodate the well-aligned short-context preference while deviating the under-aligned long-context preference. Therefore, LongPO internally transfers preferences from short to long contexts without requiring external supervision, effectively addressing the challenge of long-context data annotation.

3.2 SHORT-TO-LONG CONSTRAINT

Long-context alignment often leads to an imbalance between long- and short-context performance. While this issue can be mitigated by carefully calibrating the scale and mixing proportion of long and short data across various context lengths, such an approach is resource-intensive and time-consuming. Moreover, an excessive incorporation of short-context data may inadvertently lead to insufficient long-context alignment. In LongPO, we recognize that the degradation in short-context performance during long-context alignment may be attributed to an improper (or missing) constraint in current alignment methods.

Specifically, the RLHF and DPO objectives (implicitly) include a KL divergence term, $\beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y \mid x) \,\|\, \pi_{\text{ref}}(y \mid x)\big]$, which serves as a constraint to prevent excessive deviation from the reference model in Eq. (1). For a long input x_{L}, this constraint $\mathcal{C}$ is expressed as:

$$\mathcal{C} = \beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y \mid x_{\text{L}}) \,\|\, \pi_{\text{ref}}(y \mid x_{\text{L}})\big]. \tag{8}$$

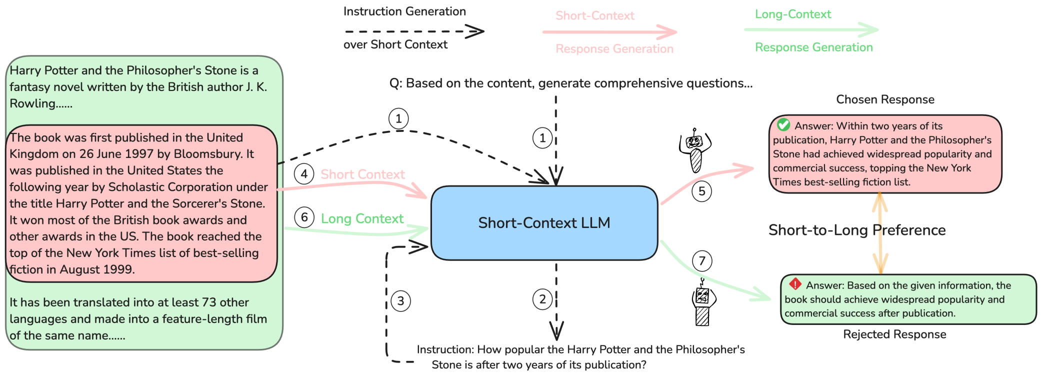
However, the reference model is typically the short-context model π_{S} itself, which is not adept at handling long contexts. This results in a problematic reference distribution $\pi_{\text{ref}}(y \mid x_{\text{L}})$, leading to undesired deviation from the short-context model distribution.

To address this issue, we propose a *short-to-long constraint* leveraging the quadruples introduced in Eq. (7). Recall that x_{S} contains all the essential information from x_{L} required to generate a satisfactory response, π_{S} can serve as a proficient reference model conditioned on x_{S}. While for an ideal reference model π^*_{ref} capable of handling context lengths from short to long, we should have:

$$\mathbb{D}_{\text{KL}}\big[\pi^*_{\text{ref}}(y \mid x_{\text{L}}) \,\|\, \pi^*_{\text{ref}}(y \mid x_{\text{S}})\big] = \mathbb{D}_{\text{KL}}\big[\pi^*_{\text{ref}}(y \mid x_{\text{S}}) \,\|\, \pi_{\text{S}}(y \mid x_{\text{S}})\big] = 0, \tag{9}$$

namely $\pi^*_{\text{ref}}(y \mid x_{\text{L}})$ and $\pi_{\text{S}}(y \mid x_{\text{S}})$ are identical distribution following Gibbs' inequality. We hence derive an adjusted short-to-long constraint between short-context reference model and "long-context" policy model given contexts of different lengths:

$$\mathcal{C}' = \beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y \mid x_{\text{L}}) \,\|\, \pi_{\text{S}}(y \mid x_{\text{S}})\big]. \tag{10}$$



Harry Potter and the Philosopher's Stone is a fantasy novel written by the British author J. K. Rowling......

The book was first published in the United Kingdom on 26 June 1997 by Bloomsbury. It was published in the United States the following year by Scholastic Corporation under the title Harry Potter and the Sorcerer's Stone. It won most of the British book awards and other awards in the US. The book reached the top of the New York Times list of best-selling fiction in August 1999.

It has been translated into at least 73 other languages and made into a feature-length film of the same name......

Q: Based on the content, generate comprehensive questions...

Instruction: How popular the Harry Potter and the Philosopher's Stone is after two years of its publication?

✓ Answer: Within two years of its publication, Harry Potter and the Philosopher's Stone had achieved widespread popularity and commercial success, topping the New York Times best-selling fiction list.

! Answer: Based on the given information, the book should achieve widespread popularity and commercial success after publication.

Figure 2: The procedure of generating short-to-long preference data from step 1 to 7.

This refined constraint ensures that the policy model π_θ operating on long contexts does not deviate significantly from the short-context model π_S when provided with the essential information. By enforcing this constraint, we aim to preserve the short-context performance during long-context alignment, thereby addressing the imbalance issue in a more principled manner.

By incorporating the short-to-long constraint in Eq. (2), we have a refined reward function for long input x_L (following derivation in Appx. §A.1):

$$r_\theta^{\mathrm{LongPO}}(x_\mathrm{L}, y) = \beta \log \frac{\pi_\theta(y \mid x_\mathrm{L})}{\pi_\mathrm{S}(y \mid x_\mathrm{S})} + \beta \log Z(x_\mathrm{L}, x_\mathrm{S}), \tag{11}$$

where x_S is extracted from x_L as illustrated in Eq. (5). Hence we access the LongPO objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{LongPO}}(\pi_\theta; \pi_\mathrm{S}) &= -\mathbb{E}_{(x_\mathrm{S}, x_\mathrm{L}, y_\mathrm{S}, y_\mathrm{L}) \sim \mathcal{D}^{\mathrm{SL}}} \left[\sigma\left(r_\theta^{\mathrm{LongPO}}(x_\mathrm{L}, y_\mathrm{S}) - r_\theta^{\mathrm{LongPO}}(x_\mathrm{L}, y_\mathrm{L})\right) \right] \\ &= -\mathbb{E}_{(x_\mathrm{S}, x_\mathrm{L}, y_\mathrm{S}, y_\mathrm{L}) \sim \mathcal{D}^{\mathrm{SL}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_\mathrm{S} \mid x_\mathrm{L})}{\pi_\mathrm{S}(y_\mathrm{S} \mid x_\mathrm{S})} - \beta \log \frac{\pi_\theta(y_\mathrm{L} \mid x_\mathrm{L})}{\pi_\mathrm{S}(y_\mathrm{L} \mid x_\mathrm{S})} \right) \right].\end{aligned} \tag{12}$$

3.3 SELF-EVOLVING

Initialization. LongPO relies solely on access to a well-aligned short-context LLM, i.e., π_S, in conjunction with a long-context plain corpus. Note that the long-context corpus need not be meticulously crafted, as it can be sampled and extracted from existing pretraining corpora of LLMs.

Construction of short-to-long preference data. The construction of short-to-long preference data $\mathcal{D}^{\mathrm{SL}}$ introduced in §3.1 assumes an impractical extractor capable of retrieving essential information from long contexts for each instruction. To satisfy this hypothesis, we reversely prompt π_S to generate instructions for shortened chunks within long documents. This ensures that the short context information is self-contained for instructions. Concretely, our data construction process involves two steps as displayed in Figure 2:

1. **Instruction Generation.** For each long document C_L, we randomly sample a shortened chunk C_S and prompt the π_S to generate an instruction via the Self-Instruct (Wang et al.). To ensure the diversity of instructions, the model is prompted to generate an instruction pool first and then we randomly sample an instruction I_L from this pool.
2. **Response Generation.** Using the generated instruction I_L, we prompt π_S to produce two responses: a chosen response $y_\mathrm{S} \sim \pi_\mathrm{S}(y \mid x_\mathrm{S})$ based on the short context x_S, and a rejected response $y_\mathrm{L} \sim \pi_\mathrm{S}(y \mid x_\mathrm{L})$ derived from the long context x_L.

Iterative self-evolving training with LongPO. LongPO employs an iterative process to extend LLM context length. Initially, a short-context LLM π_S generates short-to-long preference data for documents of length L^1. The resulting model after LongPO training, now capable of handling L^1

contexts, then serves as the new "short-context LLM" for the next iteration, generating data for an extended length L^2. This process repeats, progressively increasing context length capacity.

As there are multiple short chunks $C_{\mathrm{S}} = \{C_{\mathrm{S}}^i\}_{i=1}^n$ within a long document C_{L}, we collect the instruction-response triples $(I_{\mathrm{L}}^i, y_{\mathrm{S}}^i, y_{\mathrm{L}}^i)$ for each chunk within identical long document, to form a multi-turn dataset $\hat{\mathcal{D}}^{\mathrm{SL}}$. We then aggregate the probabilities across all turns to produce a multi-turn LongPO objective:

$$\mathcal{L}_{\mathrm{LongPO}}^{\mathrm{MT}}(\pi_\theta; \pi_{\mathrm{S}}) = -\mathbb{E}_{(x_{\mathrm{S}}, x_{\mathrm{L}}, y_{\mathrm{S}}, y_{\mathrm{L}}) \sim \hat{\mathcal{D}}^{\mathrm{SL}}} \left[\log \sigma \left(\beta \log \frac{\sum_{i=1}^n \pi_\theta(y_{\mathrm{S}}^i \mid x_{\mathrm{L}}^i)}{\sum_{i=1}^n \pi_{\mathrm{S}}(y_{\mathrm{S}}^i \mid C_{\mathrm{S}}^i)} - \beta \log \frac{\sum_{i=1}^n \pi_\theta(y_{\mathrm{L}}^i \mid x_{\mathrm{L}}^i)}{\sum_{i=1}^n \pi_{\mathrm{S}}(y_{\mathrm{L}}^i \mid C_{\mathrm{S}}^i)} \right) \right], \tag{13}$$

where $x_{\mathrm{S}} = \{[C_{\mathrm{S}}^i; I_{\mathrm{L}}^i]\}_{i=1}^n, x_{\mathrm{L}} = \{[C_{\mathrm{L}}; I_{\mathrm{L}}^i]\}_{i=1}^n, y_{\mathrm{S}} = \{y_{\mathrm{S}}^i\}_{i=1}^n$, and $y_{\mathrm{L}} = \{y_{\mathrm{L}}^i\}_{i=1}^n$. LLMs trained with LongPO do not necessarily involve continual training before, which may lead to instability when processing long contexts. To address this issue and stabilize the training process, we incorporate a continual training objective following Pang et al. (2024b). Specifically, we add the negative log-likelihood (NLL) loss over entire long chosen sequences $S_{\mathrm{L}} = [x_{\mathrm{L}}; \{I_{\mathrm{L}}^i; y_{\mathrm{S}}^i\}_{i=1}^n]$ to LongPO objective. Thus, our final training objective is:

$$\mathcal{L}_\theta = \lambda \cdot \mathcal{L}_{\mathrm{LongPO}}^{\mathrm{MT}}(\pi_\theta; \pi_{\mathrm{S}}) + \mathcal{L}_{\mathrm{NLL}}(\pi_\theta; S_{\mathrm{L}}) = \lambda \cdot \mathcal{L}_{\mathrm{LongPO}}^{\mathrm{MT}}(\pi_\theta; \pi_{\mathrm{S}}) + \frac{\pi_\theta(S_{\mathrm{L}})}{|S_{\mathrm{L}}|}. \tag{14}$$

4 EXPERIMENTAL SETUP

4.1 TRAINING SETUP

Data Curation Details. We curate the short-to-long preference data based on a long-context corpus sampled from the Book and ArXiv subsets of Long-Data-Collection[1], and the GitHub subset of RedPajama (Computer, 2023). For a specific target length (e.g., 128K tokens), we filter the corpus to include only documents that are shorter than this length but longer than 64K tokens. This process yields a corpus of 45K documents of 128K tokens and 22K documents of 256K tokens. Each long document is then segmented into chunks of up to 32K tokens, with a maximum of 4 randomly-sampled chunks retained per document. For instruction generation, we prompt short-context models to generate 4 instructions per document, from which we randomly select one for further use.

Training Details. We extend the context length of Mistral-7B-Instruct-v0.2 using our LongPO on short-to-long preference data specifically generated by model itself. The training process involves two iterations: (1) In the first iteration, we use Mistral-7B-Instruct-v0.2 to generate data with a length of 128K and extend the context length to 128K; (2) In the second iteration, we utilize the resulting model from first iteration to generate data with a length of 256K and further extend the context length to 256K. We leverage Deepspeed-Ulysses (Jacobs et al., 2023) for sequence parallelism and employ Flash Attention (Dao et al., 2022; Dao, 2023) for efficient computation. All models are optimized using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 5e-7. We set the margin β in Eq. (13) to 0.1 and the weighting factor λ in Eq. (14) to 0.01. The batch size is set as 8.

4.2 EVALUATION BENCHMARKS

We assess both the long- and short-context capabilities of our models against baselines. The long-context evaluation utilizes the following benchmarks:

- **InfiniteBench** (Zhang et al.). We evaluate all models on three tasks in this benchmark: summarization (En.Sum), long-book question answering (En.QA), and multi-choice question-answering (En.MC). The evaluation length is beyond 100K.
- **RULER** (Hsieh et al., 2024). This benchmark comprises four types of synthetic tasks across variable sequence lengths (4K to 128K): Needle-in-a-haystack (NIAH) retrieval, Multi-hop Tracing with Variable Tracking (VT), Aggregation, and Question Answering (QA). We exclude the Aggregation tasks, which involve word frequency counting within the context, since they present challenges in word counting beyond mere long-context capabilities that current LLMs still struggle in.

---

[1] `https://huggingface.co/datasets/togethercomputer/Long-Data-Collections`

- **LongBench-Chat** (Bai et al., 2024). This benchmark assesses instruction-following abilities over long contexts (10K to 100K tokens), employing GPT-4-128K as an impartial judge to evaluate model-generated responses. We filter out the English samples for fair comparison across different models.

For short-context evaluation, we employ MMLU (Hendrycks et al., 2021), ARC-C (Clark et al., 2018), Hellaswag (Zellers et al., 2019) and Winogrande (Sakaguchi et al., 2019) for assessing the general language understanding and reasoning capabilities, and MT-Bench (Zheng et al., 2023) for assessing instruction-following capability.

4.3 BASELINES

We train our LongPO on Mistral-7B-Instruct-v0.2, comparing them against a range of powerful LLMs including GPT-4-128K, Qwen2-72B-Instruct (Yang et al., 2024), LLaMA-3.1-70B, LLaMA-3.1-8B, GLM-4-9B-Chat, GLM-4-9B-Chat-1M, LWM-Text-Chat-1M (Liu et al., 2024b), and Yarn-Mistral-7b-128k (Peng et al., 2023). Additionally, we establish baselines using Mistral-7B-Instruct-v0.2 trained with conventional SFT and DPO on the same dataset used for LongPO.

For short-context evaluation, we primarily compare the performance of naive LLMs against their counterparts post-trained with SFT, DPO, and LongPO on our synthetic data. To provide a more comprehensive comparison, we also include two series of open-source long-context language models: GLM-4-9B-Chat versus GLM-4-9B-Chat-1M, and LWM-Text-Chat-128k versus LWM-Text-Chat-1M. This allows us to assess the effectiveness of our LongPO to maintain the short-context performance during long-context alignment, comparing with baselines utilizing various strategies.

5 RESULTS AND ANALYSES

In this section, we demonstrate the exceptional effectiveness of LongPO through two types of comparisons: (1) comparison with naive SFT and DPO trained on identical models and datasets; (2) comparison with SOTA long-context LLMs. Both comparisons are conducted on both long-context and short-context benchmarks.

5.1 COMPARISON WITH SFT AND DPO

We first compare LongPO with conventional SFT and DPO using identical LLM (Mistral-7B-Instruct-v0.2). All models are trained on equivalent self-generated datasets, as detailed in §4.1. Given the inability of SFT to leverage preference data, we apply it to the instructions paired with chosen responses.

LongPO exhibits superior performance over SFT and DPO. The experimental results, illustrated in Table 1, reveal consistent and substantial performance gains (10 to 20+ points) of LongPO over SFT and DPO across a diverse range of long-context tasks. Crucially, as depicted in Figure 3, LongPO maintains robust short-context performance compared with original short-context LLMs (59.99 vs 59.15 on MMLU), whereas SFT and DPO exhibit notable degradation in short-context scenarios after long-context alignment process.

The performance disparity between LongPO and SFT can be attributed to the explicit integration of short-to-long preference in LongPO, which is either absent or merely implicit in the chosen responses utilized by SFT. While both LongPO and DPO leverage the proposed short-to-long preference data, the pivotal difference lies in the short-to-long constraint introduced in §3.2. The marked performance gaps between LongPO and DPO, observed across both long- and short-context tasks, highlight the effectiveness of the proposed constraint for successfully mitigating the problematic limitations in DPO and retaining the short-context performance during long-context training. More ablations are detailed in §5.3.

5.2 COMPARISON WITH SOTA LONG-CONTEXT LLMS

To further substantiate the efficacy of LongPO, we conducted an extensive comparison between our LongPO-trained Mistral-7B and leading long-context LLMs across varying model scales.

Table 1: Long-Context Performance of our LongPO compared with baselines. Higher is better for all metrics. Results marked with $\flat$ are evaluated by ourselves, while other results of baselines are sourced from the original benchmarks.

Model	Train/Claimed	InfiniteBench				RULER				LongBench-
	Length	En.Sum	En.QA	En.MC	AVG.	NIAH	VT	QA	AVG.	Chat (EN)
GPT-4-128K	128K	14.73	22.44	67.25	34.81	95.4	99.9	70.3	88.53	8.40
Qwen2-72B	128K	24.32$^\flat$	7.03$^\flat$	72.05$^\flat$	34.47$^\flat$	88.6	95.7	66.7	83.67	7.72$^\flat$
LLaMA 3.1-70B	128K	33.55$^\flat$	36.08$^\flat$	69.00$^\flat$	46.21$^\flat$	96.1	93.2	67.8	85.7	6.67$^\flat$
LLaMA 3.1-8B	128K	28.06$^\flat$	30.47$^\flat$	58.08$^\flat$	38.87$^\flat$	97.93	91.4	64.7	84.68	6.22$^\flat$
GLM-4-9B	128K	14.84$^\flat$	9.51$^\flat$	67.25$^\flat$	30.53$^\flat$	96.51$^\flat$	97.3$^\flat$	64.8$^\flat$0	86.20$^\flat$	5.67$^\flat$
GLM-4-9B-1M	1M	28.3	9.7	68.6	35.53	98.2	99.4	69.4	89.0	5.03$^\flat$
LWM-7B-1M	1M	4.33$^\flat$	0.0$^\flat$	3.06$^\flat$	2.46$^\flat$	87.20	57.5	56.4	67.03	1.25$^\flat$
YaRN-Mistral-7B	128K	9.09	9.55	27.95	15.53	63.4	36.1	25.9	41.8	-
Mistral-7B	32K	22.13	4.93	14.41	13.82	72.60	74.40	52.2	66.4	4.10
- SFT	128K	23.44	13.45	53.21	30.03	88.73	79.64	51.08	73.15	4.25
- DPO	128K	15.21	10.34	48.14	25.56	74.25	72.36	50.24	65.62	4.08
- LongPO (iter1)	128K	27.05	23.51	67.25	39.27	96.88	96.49	64.81	86.06	5.42
- LongPO (iter2)	256K	28.16	24.43	66.35	39.65	96.80	97.0	64.87	86.22	5.48
- LongPO (iter3)	512K	29.10	27.85	66.67	41.21	97.28	97.48	64.92	86.56	5.80
Qwen 2.5-7B	32K	22.89	6.08	52.4	27.12	83.82	81.31	53.14	72.76	5.80
- LongPO (iter1)	128K	32.06	17.32	72.05	40.48	96.08	92.24	64.4	84.24	5.75

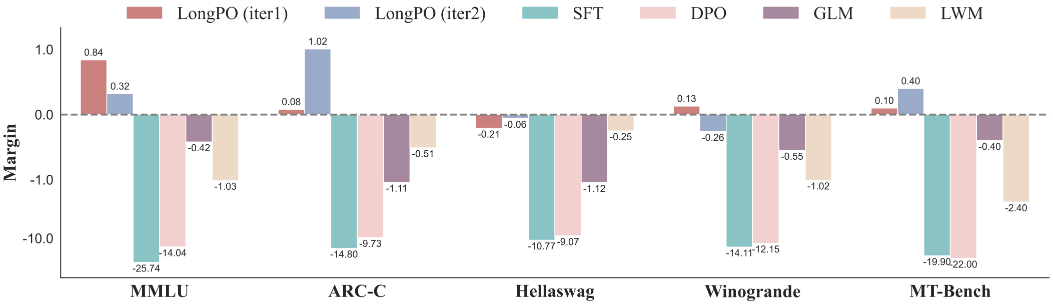


Figure 3: The margins of the short-context performance of LongPO and baselines relative to correspond base model. GLM and LWM refer to the margins of GLM-9B-1M and LWM-7B-1M over GLM-9B-128K and LWM-7B-128K, respectively. MT-Bench metrics ($\in$[0, 10]) are linearly scaled to [0, 100] for better comparability across tasks.

LongPO demonstrates exceptional competitiveness at similar scale. As detailed in Table 1, LongPO demonstrates formidable competitiveness in terms of models at similar scale. For example, Mistral-7B-LongPO significantly outperforms some established long-context models, including LWM-7B and YaRN-Mistral, across all long-context tasks in InfiniteBench and RULER. Remarkably, Mistral-7B-LongPO-128K surpass GLM-4-9B (39.27 vs. 30.53 on InfiniteBench and 86.06 vs. 86.20 on RULER), although the latter is training on manually annotated long-context data spanning up to 128K sequence length. Moreover, GLM-4-9B-1M, an extension of GLM-4-9B trained on contexts up to 1M tokens, demonstrates slightly superior performance than LongPO on the RULER benchmark. However, these performance gains come at the costs of *degenerated short-context performance* (0.41 on MMLU) and *long-context instruction-following capability* (0.64 on LongBench-Chat (EN)) as illustrated in Figure 3. Notably, our models still outperform GLM-4-9B-1M on InfiniBench even trained with substantially shorter sequences.

This striking outcome underscores the exceptional efficiency of LongPO in transferring and amplifying performance from short to long contexts through the use of self-generated data, thereby circumventing the need for extensive manual annotation and mitigating the trade-offs typically associated with extended context training.

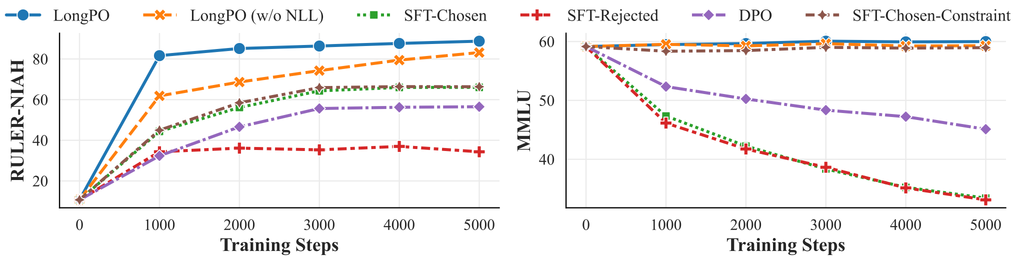


Figure 4: Long- and short-context performance comparison among LongPO, SFT on chosen responses (**SFT-Chosen**), SFT on rejected responses (**SFT-Rejected**), DPO, and SFT on chosen responses with short-to-long constraint (**SFT-Chosen-Constraint**).

Long-context annotation is not sufficient. The superiority of our approach is particularly evident in the En.QA task within InfiniteBench, which involves complex free-form question answering over extensive book-length contexts. In this challenging task, our models surpass both GLM-4-9B and GLM-4-9B-1M by substantial margins (10+ points). The inherent difficulty of such task, which poses challenges even for human annotators, highlights the limitations of relying solely on manually annotated long-context data. By effectively transferring short-context capabilities to long-context scenarios, LongPO demonstrates superior scalability and efficacy across diverse and intricate tasks.

Dominant LLMs Yet to Conquer Long-Context Scenarios When benchmarked against leading models such as GPT-4-128K, our LongPO-trained models still exhibit comparable or even superior long-context performance (e.g., Mistral-7B-LongPO-128K of 39.27 vs. GPT-4-128K of 34.81 on InfiniteBench), despite being based on significantly smaller Mistral-7B. This observation reveals that even the most advanced LLMs have not yet achieved the same level of dominance in long-context scenarios as they have in short-context tasks. This performance gap can be attributed primarily to the scarcity of high-quality, large-scale long-context training data. The dearth of such data is particularly impactful for larger LLMs, given the established scaling laws in language model training. This finding underscores the potential of LongPO as a pivotal approach in advancing long-context capabilities, offering a pathway to enhanced performance without the requirement for externally annotated long-context datasets.

5.3 ABLATION STUDIES

We conduct comprehensive ablation studies to investigate the efficacy of components in LongPO:

Effectiveness of short-to-long preference. The core of LongPO is learning the short-to-long preference between chosen and rejected responses given short and long contexts, respectively. To evaluate this component's effectiveness, we compare LongPO with two baseline methods: SFT on chosen responses (SFT-Chosen) and on rejected responses (SFT-Rejected). SFT-Chosen implicitly incorporates short-context preference, while SFT-Rejected entirely omits it. As illustrated in Figure 4, LongPO consistently outperforms both SFT variants in long-context performance (RULER-NIAH) throughout the training process. This substantial improvement underscores the efficacy of our short-to-long preference approach in enhancing long-context capabilities.

Effectiveness of short-to-long constraint. To assess the impact of our short-to-long constraint, we compare LongPO with DPO upon short-to-long preference that removes this constraint. As evident in Figure 4, the unconstrained DPO demonstrates markedly inferior performance throughout the training process, both in long- and short-context tasks. Notably, short-context capabilities degrade rapidly in DPO during the initial training. Conversely, when we apply our short-to-long constraint to naive SFT without explicit short-to-long preference, the model maintains short-context performance on par with the original LLMs, even after long-context alignment. These results demonstrate the crucial role of our short-to-long constraint in preserving short-context capabilities while improving long-context performance.

Impact of NLL loss. We investigate the effect of incorporating a negative log-likelihood (NLL) loss over long context input and chosen response in Eq. (14) during LongPO training. As shown in Figure 4, removing the NLL loss significantly degrades the long-context performance of LongPO across the training procedure. Specifically, the convergence of training for long-context performance becomes slower. This demonstrates the crucial role of NLL loss in enhancing long-context capabilities without resorting to continual training on long data.

6 RELATED WORK

Alignment of LLMs. Aligning Large Language Models (LLMs) with human preferences and values has been crucial to unlocking their full potential from large-scale pretraining. The typical alignment process begins with Supervised Finetuning (SFT) on annotated instruction-response pairs. This is followed by Reinforcement Learning from Human Feedback (RLHF), which aligns LLMs more closely with human intentions through reward model training and policy optimization (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Stiennon et al., 2020). To streamline RLHF training, Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants (Ethayarajh et al., 2024; Azar et al., 2023; Pang et al., 2024a; Hong et al., 2024; Meng et al., 2024) have been proposed, eliminating the need for explicit reward model training by learning preferences directly from human-ranked response pairs. While these alignment methods have shown significant success, they heavily rely on human-annotated data. This reliance becomes problematic for long-context data, where human annotation is both challenging and potentially less reliable.

Long-context extending of LLMs. Extending the context length of LLMs has been approached through various methods. Some techniques involve scaling the rotary position embedding (Su et al., 2022) followed by continual training on a small corpus of long documents (Chen et al., 2023b; Peng et al., 2023; Rozière et al., 2023; Chen et al., 2023a). Alternative approaches, such as those proposed by Jin et al. (2024); An et al. (2024), introduce hierarchical or chunked attention mechanisms to extend context length without additional training. However, these methods often involve limitations in practical applications. Recent advancements include the work of Dubey et al. (2024), who proposed continual pretraining on a massive long-context corpus (800B tokens) and incorporating a small fraction (0.1%) of long-context data during SFT to enhance long-context capabilities. Zeng et al. (2024) utilizes human-annotated long-context data for SFT and DPO to align long-context LLMs. Despite their effectiveness, these methods require either extensive training or human annotation of long-context data, making them prohibitively expensive and lack scalability.

Self-Evolving LLMs. Recent works (Yuan et al., 2024; Liu et al., 2024a; Li et al., 2024) have unveiled the remarkable capability of Large Language Models (LLMs) to evolve from relatively weak to significantly stronger performance through self-augmented data. Yuan et al. (2024); Liu et al. (2024a) leverage iterative training on model-generated responses, ranked by LLM-as-a-Judge (Zheng et al., 2023) prompting, to enhance model itself. Li et al. (2024) introduces the instruction backtranslation to produce self-augmenting data that further enhances model capabilities. Our work first extends the self-evolution property to the context length, to develop long-context LLMs without relying on external annotations.

7 CONCLUSION AND DISCUSSION

In this work, we propose LongPO, a novel long-context alignment method that enables LLMs to effectively transfer their short-context capabilities to long-context scenarios. Our approach addresses key challenges in long-context alignment by leveraging intrinsic model knowledge, eliminating the need for external long-context annotated data. LongPO is built on short-to-long preference data, comprising paired responses for the same instruction given a long context and relevant shortened chunk, respectively. By steering the policy model to learn from the discrepancies within these paired responses, LongPO facilitates the transfer of established capabilities from short to long contexts. In addition, LongPO incorporates a short-to-long constraint using KL divergence, that effectively preserve short-context performance during training. Experimental results demonstrate that LongPO significantly improves long-context performance across various tasks, outperforming existing alignment methods and even surpassing more sophisticated models. Importantly, this improvement is achieved without sacrificing short-context proficiency. The success of LongPO highlights the potential of leveraging internal model knowledge for alignment tasks, opening new avenues for efficient adaptation of LLMs to diverse context lengths.

REFERENCES

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models, 2024.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *ArXiv*, abs/2310.12036, 2023. URL `https://api.semanticscholar.org/CorpusID:264288854`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. 2022. URL `https://arxiv.org/abs/2212.08073`.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. 2024. URL `https://arxiv.org/abs/2401.18058`.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard (2023-2024). `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`, 2023.

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Li Bing. Clex: Continuous length extrapolation for large language models. *ArXiv*, abs/2310.16450, 2023a. URL `https://api.semanticscholar.org/CorpusID:264451707`.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. 2023b. URL `https://arxiv.org/abs/2306.15595`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL `https://api.semanticscholar.org/CorpusID:3922816`.

Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023. URL `https://arxiv.org/abs/2307.08691`.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. URL `https://arxiv.org/abs/2205.14135`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *ArXiv*, abs/2402.01306, 2024. URL `https://api.semanticscholar.org/CorpusID:267406810`.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL `https://zenodo.org/records/12608602`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *ArXiv*, abs/2403.07691, 2024. URL `https://api.semanticscholar.org/CorpusID:268363309`.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Sam Adé Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *ArXiv*, abs/2309.14509, 2023. URL `https://api.semanticscholar.org/CorpusID:262826014`.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL `https://api.semanticscholar.org/CorpusID:263830494`.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation, 2024. URL `https://arxiv.org/abs/2308.06259`.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, and Lijie Wen. Direct large language model alignment through self-rewarding contrastive prompt distillation, 2024a. URL `https://arxiv.org/abs/2402.11907`.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024b. URL `https://arxiv.org/abs/2402.08268`.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *ArXiv*, abs/2405.14734, 2024. URL `https://api.semanticscholar.org/CorpusID:269983560`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL `https://api.semanticscholar.org/CorpusID:246426909`.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *ArXiv*, abs/2404.19733, 2024a. URL `https://api.semanticscholar.org/CorpusID:269457506`.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024b. URL `https://arxiv.org/abs/2404.19733`.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. 2023. URL `https://arxiv.org/abs/2309.00071`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. 2023. URL `https://arxiv.org/abs/2308.12950`.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 2022. URL `https://arxiv.org/abs/2104.09864`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. URL `https://aclanthology.org/2023.acl-long.754`.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=gEZrGCozdqR`.

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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *ArXiv*, abs/2401.10020, 2024. URL `https://api.semanticscholar.org/CorpusID:267035293`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL `https://api.semanticscholar.org/CorpusID:159041722`.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Ming yue Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiaoyu Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yi An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhenyi Yang, Zhengxiao Du, Zhen-Ping Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793, 2024. URL `https://api.semanticscholar.org/CorpusID:270562306`.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. URL `https://aclanthology.org/2024.acl-long.814`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL `https://api.semanticscholar.org/CorpusID:259129398`.

A MATHEMATICAL DERIVATIONS

A.1 DERIVING THE LONGPO OBJECTIVE

In this section, we will derive the reward function of our LongPO objective in Eq. (11) by incorporating short-to-long constraint in Eq. (10). Starting from RLHF objective in Eq. (1) with short-to-long constraint, we have

$$\max_{\pi} \mathbb{E}_{x_{\mathrm{L}}\sim\mathcal{D}, y\sim\pi}\big[r(x_{\mathrm{L}}, y)\big] - \beta\mathbb{D}_{\mathrm{KL}}\big[\pi(y|x_{\mathrm{L}})||\pi_{\mathrm{S}}(y|x_{\mathrm{S}})\big] \tag{15}$$

Following the DPO derivation process (Rafailov et al., 2023), we have:

$$\begin{aligned}
&\max_{\pi} \mathbb{E}_{(x_{\mathrm{L}},x_{\mathrm{S}})\sim\hat{\mathcal{D}}^{\mathrm{SL}}, y\sim\pi(y|x_{\mathrm{L}})}\big[r(x_{\mathrm{L}}, y)\big] - \beta\mathbb{D}_{\mathrm{KL}}\big[\pi(y|x_{\mathrm{L}}) \parallel \pi_{\mathrm{S}}(y|x_{\mathrm{S}})\big] \\
&= \max_{\pi} \mathbb{E}_{(x_{\mathrm{L}},x_{\mathrm{S}})\sim\hat{\mathcal{D}}^{\mathrm{SL}}} \mathbb{E}_{y\sim\pi(y|x_{\mathrm{L}})}\left[r(x_{\mathrm{L}}, y) - \beta\log\frac{\pi(y|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y|x_{\mathrm{S}})}\right] \\
&= \min_{\pi} \mathbb{E}_{(x_{\mathrm{L}},x_{\mathrm{S}})\sim\hat{\mathcal{D}}^{\mathrm{SL}}} \mathbb{E}_{y\sim\pi(y|x_{\mathrm{L}})}\left[\log\frac{\pi(y|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y|x_{\mathrm{S}})} - \frac{1}{\beta}r(x_{\mathrm{L}}, y)\right] \\
&= \min_{\pi} \mathbb{E}_{(x_{\mathrm{L}},x_{\mathrm{S}})\sim\hat{\mathcal{D}}^{\mathrm{SL}}} \mathbb{E}_{y\sim\pi(y|x_{\mathrm{L}})}\left[\log\frac{\pi(y|x_{\mathrm{L}})}{\frac{1}{Z(x_{\mathrm{L}},x_{\mathrm{S}})}\pi_{\mathrm{S}}(y|x_{\mathrm{S}})\exp\left(\frac{1}{\beta}r(x_{\mathrm{L}}, y)\right)} - \log Z(x_{\mathrm{L}}, x_{\mathrm{S}})\right],
\end{aligned} \tag{16}$$

where we have partition function:

$$Z(x_{\mathrm{L}}, x_{\mathrm{S}}) = \sum_{y}\pi_{\mathrm{S}}(y|x_{\mathrm{S}})\exp\left(\frac{1}{\beta}r(x_{\mathrm{L}}, y)\right).$$

The partition function is only related to x_{L}, x_{S}, and original short-context LLM π_{S}. Hence we have the optimal solution following Rafailov et al. (2023):

$$\pi^*(y|x_{\mathrm{L}}) = \frac{1}{Z(x_{\mathrm{L}}, x_{\mathrm{S}})}\pi_{\mathrm{S}}(y|x_{\mathrm{S}})\exp\left(\frac{1}{\beta}r(x_{\mathrm{L}}, y)\right). \tag{17}$$

The optimal reward function would be derived:

$$r^*(x_{\mathrm{L}}, y) = \beta\log\frac{\pi^*(y|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y|x_{\mathrm{S}})} + \beta\log Z(x_{\mathrm{L}}, x_{\mathrm{S}}). \tag{18}$$

We thus have:

$$\begin{aligned}
p^*(y_1 \succ y_2|x_{\mathrm{L}}) &= \frac{\exp\left(r^*(x_{\mathrm{L}}, y_1)\right)}{\exp\left(r^*(x_{\mathrm{L}}, y_1)\right) + \exp\left(r^*(x_{\mathrm{L}}, y_2)\right)} \\
&= \frac{\exp\left(\beta\log\frac{\pi^*(y_1|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_1|x_{\mathrm{S}})} + \beta\log Z(x_{\mathrm{L}}, x_{\mathrm{S}})\right)}{\exp\left(\beta\log\frac{\pi^*(y_1|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_1|x_{\mathrm{S}})} + \beta\log Z(x_{\mathrm{L}}, x_{\mathrm{S}})\right) + \exp\left(\beta\log\frac{\pi^*(y_2|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_2|x_{\mathrm{S}})} + \beta\log Z(x_{\mathrm{L}}, x_{\mathrm{S}})\right)} \\
&= \frac{1}{1 + \exp\left(\beta\log\frac{\pi^*(y_2|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_2|x_{\mathrm{S}})} - \beta\log\frac{\pi^*(y_1|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_1|x_{\mathrm{S}})}\right)} \\
&= \sigma\left(\beta\log\frac{\pi^*(y_1|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_1|x_{\mathrm{S}})} - \beta\log\frac{\pi^*(y_2|x_{\mathrm{L}})}{\pi_{\mathrm{S}}(y_2|x_{\mathrm{S}})}\right).
\end{aligned}$$

By optimizing the π_θ towards the optimal policy π^*, we finally access the objective of LongPO in Eq. (12).

B EXPERIMENTAL DETAILS

B.1 DATA CONSTRUCTION DETAILS

We prompt the Mistral-7B-Instruct-v0.2 to generate instructions with decode parameters of temperature $T = 0.7$ and $p = 0.9$. The prompt of Self-Instruct to generate an instruction pool is shown

> Based on the content presented above, generate 4 comprehensive English questions that test a reader's comprehension, analytical skills, and ability to extract and interconnect key themes and ideas across the entire document.
>
> Each question should:
> 1. Encourage the reader to draw connections between different sections or concepts within the text.
> 2. Challenge the reader to not only recall information but also to synthesize and summarize the material in a coherent manner.
> 3. Be unique in its focus, avoiding repetition and ensuring a broad coverage of the document's content.
> 4. Stimulate critical thinking by requiring the application or evaluation of the text's information in broader contexts or hypothetical scenarios, if relevant.
> 5. Ensure the question is clear and unambiguous.
>
> Please directly give the questions without verbose illustration, and format the 4 questions numerically from “1:” to “4:”.

Figure 5: The prompt for generating instruction pool.

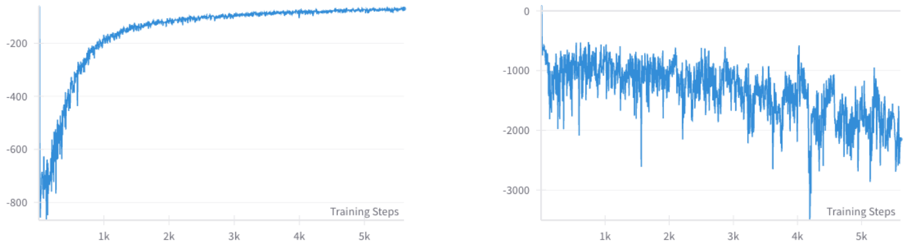
in Figure 5. For generating the corresponding responses, we directly concatenate the short or long context with corresponding instructions and adopt the greedy decoding to maintain the deterministic behaviour of LLMs. As shown in Figure 6, the model would tend to prefer the high-quality chosen response and deviate from the low-quality rejected response over long context, hence improve the long-context capabilities.

B.2 EVALUATION DETAILS

On long-context benchmarks InfiniteBench and RULER, we evaluate our models and all baselines following the settings in the original benchmarks. For short-context evaluation, we utilize the lm-evaluaton-harness framework (Gao et al., 2024) and following the evaluation settings in (Beeching et al., 2023): 5-shots for MMLU, 25-shots for ARC-C, 10-shots for Hellaswag, and 5-shots for Winogrande. We use GPT-4-Turbo-1106-Preview as the judge for MT-Bench and LongBench-Chat evaluation.

B.3 MORE TRAINING DETAILS

Leveraging the DeepSpeed-Ulysses sequence parallel framework, we train the Mistral-7B-LongPOmodel with a sequence length of 128K on an 8xA800 80GB, achieving a throughput of 4,401 tokens per second. For sequence lengths of 256K and 512K, the models are trained on a 16xA800 80GB, yielding throughputs of 4,120 tokens per second and 2,744 tokens per second, respectively. To facilitate a comparison with standard LLM alignment methods, we train Mistral-7B-Instruct-v0.2 using SFT and DPO utilizing the same short-to-long preference data of LongPO. For DPO training, we apply the same settings as LongPO outlined in §4.1, but excluding the short-to-long constraint of LongPO introduced in §3.2. Since SFT cannot utilize paired responses within preference data, we train it using only the chosen responses provided alongside long context inputs. The hyperparameters for SFT remain unchanged, except for an increase in the learning rate to 2e-5.



(a) The rewards for chosen response during training. (b) The rewards for rejected response during training.

Figure 6: The chosen and rejected rewards during the training of Mistral-7B-LongPO-128K.