GATEAU: Selecting Influential Samples for Long Context Alignment

Anonymous ACL submission

Abstract

Aligning large language models to handle instructions with extremely long contexts has yet to be fully investigated. Previous studies have attempted to scale up the available data volume by synthesizing long instruction-following samples, as constructing such a dataset tends to be challenging for annotators. However, a lack of a well-defined strategy for ensuring data quality may introduce low-quality samples and restrict the model's performance. Thus, we propose **GATEAU**, a novel framework to address the unique challenge of long context alignment by identifying the influential samples enriched with long-range dependency relations. Specifically, GATEAU measures the long-range dependencies from two essential aspects: the difficulty of generating target responses due to the long-range dependencies, and the difficulty of understanding long inputs due to such dependencies. Comprehensive experiments indicate that GATEAU effectively identifies influential samples and the model trained on these selected samples exhibits better instruction-following and long-context understanding capabilities.

1 Introduction

Large language models (LLMs) with large context windows (Du et al., 2022; Li et al., 2023; Chen et al., 2024b) have shown impressive capabilities in real-world tasks that involve extremely long contexts (Bai et al., 2023). Recent works to build long-context LLMs mainly focus on broadening context windows via position encoding extension and continual pre-training on a long corpus (Chen et al., 2023b; Peng et al., 2024; Xiong et al., 2024).

Despite these advancements, few studies consider the long context alignment of LLMs to leverage their capabilities in understanding lengthy inputs and following complex instructions. A primary obstacle lies in the difficulty of constructing a high-quality, long instruction-following dataset for supervised fine-tuning (SFT). Annotating long instruction-following data tends to be much more challenging than short ones, as it is non-trivial for annotators to understand an excessively long context and provide high-quality responses, e.g., annotators might be tasked with writing a summary for a document containing 64k words. Furthermore, modeling long-range dependencies is crucial for long-context tasks (Chen et al., 2024a; Wu et al., 2024), as such strong semantic dependencies benefit LLMs to understand lengthy inputs and generate high-quality responses. Thus, recent works (Li et al., 2023; Xiong et al., 2024) attempt to construct the long instruction-following dataset by concatenating short instruction-following samples. While these methods successfully increase sequence lengths, simply concatenating unrelated samples fails to effectively simulate the inherent long-range dependencies in authentic long samples. To address this issue, Yang (2023); Chen et al. (2024b); Bai et al. (2024) focus on synthesizing long instruction-following data. For instance, Bai et al. (2024) synthesizes 10k samples by employing Claude 2.1 (Anthropic., 2023), which supports a context window of 200k tokens, to get responses for the collected long documents.

However, when training on such synthetic samples with sufficiently lengthy contexts, LLMs still struggle to learn and model the long-range dependencies (Chen et al., 2024a). This is because indiscriminately increasing the quantity of data without a well-defined strategy for ensuring data quality can introduce low-quality samples lacking long-range dependency relations. Such samples may rely only on a few tokens before the instruction or may not require long inputs to get the target response. Thus, a critical question arises: ***How can we effectively select influential samples from a vast amount of synthetic long instruction-following data for long context alignment?***

Previous studies for selecting influential instruction data primarily focus on short samples (Li

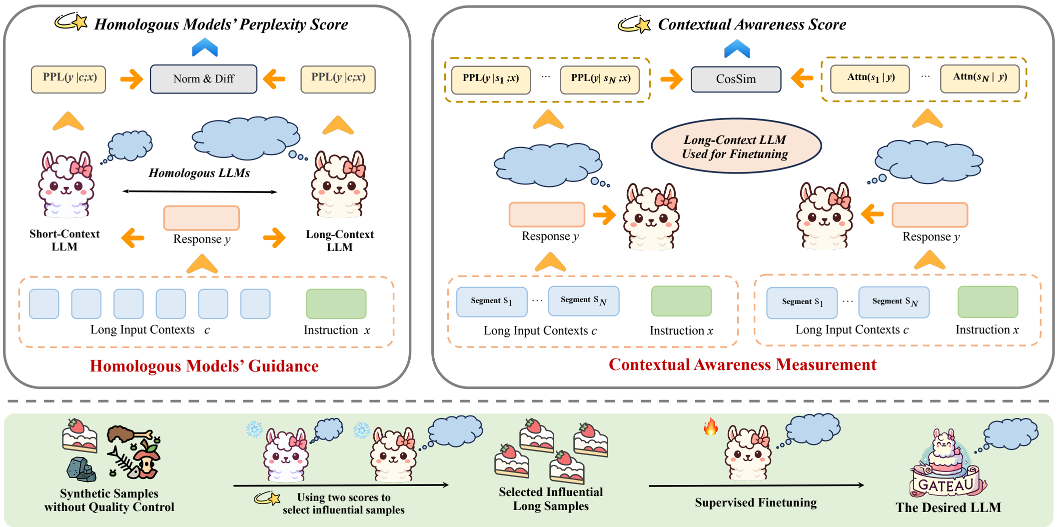

Figure 1: An overview of **GATEAU**. GATEAU first selects samples enriched with long-range dependency relations by using two proposed methods. Then it uses selected influential samples for training long-context LLMs.

et al., 2024b; Xia et al., 2024). Thus, these studies may not be effective for long context alignment as they ignore the unique challenge in long context alignment, i.e., how to select the samples enriched with meaningful long-range dependency relations. To address this challenge, we measure long-range dependencies from two essential aspects: the difficulty of generating target responses due to long-range dependencies, and the difficulty of understanding long inputs due to such dependencies. We introduce **GATEAU**, which consists of **Homologous Models' GuidAnce (HMG)** and **ConTExtual Awareness MeasUrement (CAM)**, to identify the influential long samples enriched with long-range dependency relations to achieve better long context alignment.

Specifically, HMG measures the difficulty of generating target responses due to long-range dependencies, by comparing perplexity scores of the given response between two homologous models (Yu et al., 2024) with different context windows (e.g., the perplexity scores from LLaMA-3-base-8k (Grattafiori et al., 2024) and LLaMA-3-base-64k (Lian, 2024)). The idea behind HMG is that the primary difference between homologous models with varying context windows lies in their different capabilities for modeling long-range dependencies. Thus, the disparity in the perplexity scores can be interpreted as reflecting the difficulty of generating the response caused by long-range dependencies. We further introduce CAM to measure the difficulty of understanding long input contexts due to long-range dependencies. We first calculate the importance score of different input segments concerning the given response and subsequently measure whether LLMs can pay more attention to more important segments. If LLM's attention focuses more on less important segments, it implies that it is challenging for the LLM to comprehend the long inputs correctly. Ultimately, we take the weighted sum of both scores from the two methods as the final criterion for ranking the data, selecting the most challenging samples as influential ones. When trained on these selected samples with rich long-range dependency relations, LLMs could effectively model the long-range dependencies and achieve better instruction-following performance.

We conduct extensive experiments to evaluate GATEAU, including long-context understanding benchmark (LongBench (Bai et al., 2023)) and instruction-following benchmarks (LongBench-Chat (Bai et al., 2024), MT-Bench (Zheng et al., 2023)). With GATEAU, significant performance boosts are observed, e.g., the model trained on just 10% selected samples of the dataset achieves better performance than the vanilla fine-tuning method.

2 Methodology

As shown in Figure 1, we propose **GATEAU** to select influential samples from a vast ocean of synthetic data instead of indiscriminately increasing the quantity of synthetic long instruction-following data (Chen et al., 2024b; Bai et al., 2024). Different from previous studies that only consider the short context scenarios (Li et al., 2024b; Xia et al., 2024), we attempt to address the unique challenge in long context alignment, i.e., modeling long-range dependencies. GATEAU consists of Homologous Models' Guidance and Contextual Awareness Measurement, which separately measure the difficulty of generating corresponding responses and understanding long input contexts due to the long-range dependencies. In this way, GATEAU can comprehensively and effectively measure the richness of long-range dependency relations in long samples.

2.1 Homologous Models' Guidance

Modeling long-range dependencies is essential for long context alignment (Chen et al., 2024a). However, there is still no effective metric to directly quantify the richness of long-range dependency relations in data, which hinders the selection of influential data. Therefore, in this section, we attempt to approximately assess the richness of long-range dependency relations by measuring the difficulty in generating corresponding responses due to the long-range dependencies. If LLMs find it harder to generate target responses due to long-range dependencies, it means the sample has more complex and meaningful long-range dependency relations. An intuitive approach is to use the perplexity score to measure the difficulty of generating corresponding responses (Cao et al., 2024; Li et al., 2024b), as the score evaluates the extent to which the LLM's output aligns with the corresponding correct answer. For a given long instruction-following sample $(c, x; y)$, the perplexity score of the given response y from LLMs θ is calculated as:

$$\mathrm{PPL}_{\theta}(y|c,x) = \mathrm{Exp}(-\frac{1}{|y|}\sum_{i=1}^{|y|}\log P(y_i|c,x,y_{<i};\theta)), \quad (1)$$

where c means long input contexts and x means the given instruction. A higher $\mathrm{PPL}_{\theta}(y|c,x)$ indicates the harder the response of this long instruction-following data for LLM to generate.

However, we argue that a higher $\mathrm{PPL}_{\theta}(y|x)$ does not mean the increased difficulty in generating target responses is due to long-range dependencies. A higher $\mathrm{PPL}_{\theta}(y|c,x)$ might be attributed to certain limited capabilities of LLMs, such as the limited instruction-following capability for the model without alignment, instead of handling the long-range dependency relations in this sample is more challenging for the LLM. Therefore, to minimize the influence of other factors, we propose **Homologous Models' Guidance (HMG)**. Specifically, we compare the perplexity scores of the response between two homologous models with different context windows to measure the difficulty due to the long-range dependencies. As homologous models (Yu et al., 2024) share the same pre-training stage and model architecture (e.g., LLaMA-3-base-8k and LLaMA-3-base-64k), the only difference lies in their capabilities to model long-range dependency due to the context windows extending stage. Based on this motivation, we introduce the homologous models' perplexity score $\mathrm{HMP}(c, x; y)$:

$$\begin{aligned}\mathrm{HMP}(c,x;y) = \mathrm{Norm}(\mathrm{PPL}_{\theta_A}(y|c,x)) \\ -\mathrm{Norm}(\mathrm{PPL}_{\theta_B}(y|c,x)). \quad (2)\end{aligned}$$

Model θ_A employs short context windows and θ_B is the model with long ones, e.g., LLaMA-3-base-8k θ_A and LLaMA-3-base-64k θ_B. We compute the difference in normalized perplexity scores between two homologous models with different context windows as the metric. We apply softmax normalization to each score to determine its respective ranking among the datasets, since perplexity scores of one sample from different models often can't be directly compared. By introducing a model θ_A with weaker long-range dependencies modeling capability but other similar capabilities learned during the pre-training stage, we mitigate the influence brought by lacking other capabilities compared to simply using the perplexity score as Eq. (1). Thus, the difference in perplexity scores is primarily attributed to the different abilities in modeling long-range dependencies between model θ_A and model θ_B. In other words, Eq. (2) reflects the difficulty of generating the corresponding response caused by long-range dependencies. We use the drop from $\mathrm{PPL}_{\theta_\mathrm{A}}$ to $\mathrm{PPL}_{\theta_\mathrm{B}}$ in Eq. (2) as model θ_A tends to produce a high perplexity score due to its weak ability to model long-range dependencies. Thus, a higher $\mathrm{HMP}(c, x; y)$ indicates more difficulties for LLM in response generation due to the long-range dependencies, i.e., more long-range dependency relations in this sample.

2.2 Contextual Awareness Measurement

Another challenge in long context alignment lies in enabling LLMs to understand and utilize extremely long inputs. Due to the long-range dependencies, it is hard for LLMs to utilize crucial information hidden in extremely long contexts, e.g., LLM's attention may focus on irrelevant content. Thus, we introduce **Contextual Awareness Measurement (CAM)** to evaluate whether LLMs' attention is appropriately focused on important segments within the lengthy inputs. We attempt to evaluate the importance score of each segment and calculate the LLM's attention weights on each one, getting the **Contextual Awareness Score (CAS)** via computing their similarity. For a given data $(c, x; y)$, we divide the input contexts c into N segments $[s_1, s_2, s_3, ..., s_N]$ of equal length L. For segment s_i, we first compute the designed importance score $\mathrm{IS}_\theta(s_i)$ to measure the significance of the segment in the response generation for LLM θ:

$$\mathrm{IS}_\theta(s_i) = \mathrm{Norm}(\mathrm{Exp}(-\frac{1}{|y|}\sum_{j=1}^{|y|}\log P(y_i|s_i, x, y_{<j}; \theta))). \tag{3}$$

We only keep the segment s_i as the input to calculate the perplexity score of generating the response y, indicating the difficulty of generating response y based on segment s_i. We apply softmax normalization to each score to determine its respective ranking among the segments $\{s_i\}_{i=1}^{N}$ as shown in Eq. (3). The higher $\mathrm{IS}_\theta(s_i)$ suggests a greater difficulty for LLM θ to generate the response based on segment s_i, implying that it is less important.

Once the importance scores of different segments are calculated, we then utilize the attention weights (i.e., the value of $\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})$) in the multi-head attention mechanism (Vaswani et al., 2017) to measure how the LLM utilizes these segments. We use the averaged attention weights of tokens $[t_1, ..., t_L]$ in segments s_i as the score $\mathrm{Attn}_\theta(s_i)$, which takes the form:

$$\mathrm{Attn}_\theta(s_i) = \mathrm{Norm}(\frac{1}{L}\sum_{j=1}^{L}\mathrm{Attn}_\theta(t_j|y; \theta)), \tag{4}$$

where $\mathrm{Attn}_\theta(t_j|y;\theta)$ means the attention weights averaged across the tokens in targeted response y to the token t_j in segment s_i. Meanwhile, we harness the attention weights averaged across different decoder layers and attention heads to thoroughly model how the LLM utilizes the long input contexts during the response generation (Hsieh et al., 2024). We apply softmax normalization to each score $\frac{1}{L}\sum_{j=1}^{L}\mathrm{Attn}_\theta(t_j|y;\theta)$ to determine its respective ranking among the segments $\{s_i\}_{i=1}^{N}$ to yield the score $\mathrm{Attn}_\theta(s_i)$. In so doing, we can calculate the attention weights between the response and segments, indicating how segments are utilized during the response generation.

Finally, we can measure the difficulty of understanding the long input contexts due to the long-range dependencies. For a given long instruction-following sample, we compute the CAS by resorting to the cosine similarity between importance scores $[\mathrm{IS}_\theta(s_1), ..., \mathrm{IS}_\theta(s_N)]$ and attention weights $[\mathrm{Attn}_\theta(s_1), ..., \mathrm{Attn}_\theta(s_N)]$, as follows:

$$\begin{aligned}\mathrm{CAS}(c, x; y) =& \mathrm{CosSim}([\mathrm{IS}_\theta(s_1), \ldots, \mathrm{IS}_\theta(s_N)], \\ & [\mathrm{Attn}_\theta(s_1), \ldots, \mathrm{Attn}_\theta(s_N)]).\end{aligned} \tag{5}$$

By doing this, we can measure the difficulty of understanding the long input contexts by evaluating whether LLMs' attention is focused on important segments. The insight is that if the LLM's attention focuses more on less important segments, it suggests that the LLM struggles to accurately comprehend the given long input contexts. The higher $\mathrm{CAS}(c, x; y)$ indicates more difficulties in utilizing the long input contexts to generate corresponding responses due to the long-range dependencies, which also implies more long-range dependency relations in this sample.

2.3 Selecting and Training

We frame the final score by weighting two metrics of the sample $(c, x; y)$, then select the most challenging samples as the influential samples, i.e.,

$$\begin{aligned}\mathrm{Score}(c, x; y) =& \alpha * \mathrm{Norm}(\mathrm{HMP}(c, x; y)) \\ & + (1-\alpha) * \mathrm{Norm}(\mathrm{CAS}(c, x; y)),\end{aligned} \tag{6}$$

where α is a hyperparameter. We tap softmax normalization to the $\mathrm{HMP}(c, x; y)$ and $\mathrm{CAS}(c, x; y)$ across the whole dataset. Inspired by active learning (Li et al., 2024a), when trained on these challenging data with complex long-range dependency relations, LLMs could learn such dependencies and achieve better long context alignment.

Training LLMs with instruction-following data can teach LLMs to follow user instructions. Thus, we apply SFT on the selected data (e.g., selecting 10% samples of full datasets with top 10% scores according to Eq. (6)). Then, we train LLMs using the following objective function:

$$\mathcal{L}_\theta(c, x; y) = -\sum_{i=1}^{|y|}\log P(y_i|c, x, y_{<i}; \theta). \tag{7}$$

3 Experiment

3.1 Experimental Setup

Training Datasets. We use LongAlign (Bai et al., 2024) as the long instruction-following dataset, which contains 10,000 long samples. We apply GATEAU to the LongAlign dataset. Meanwhile, similar to Bai et al. (2024), to maintain the model's general capabilities and its proficiency in following short instructions, we utilize the ShareGPT dataset (Chiang et al., 2023) as the source of short instruction data in training data. To study the impact of mixing long and short instruction samples, we evaluate GATEAU in both **Real-world Settings** and **Limited Short Instruction Data Settings**. Real-world Settings indicate real-world users prioritize short instruction-following interactions (Chiang et al., 2023). Thus, we use the full ShareGPT dataset as short instruction-following data. We also explore scenarios where short instruction data is limited, utilizing only the first 10% of ShareGPT, namely Limited Short Instruction Data Settings. More details are shown in the Appendix A.

Training Details. For the fair comparisons with Bai et al. (2024), we use LLaMA-2-7B-base-4k (Touvron et al., 2023) and LLaMA-2-7B-base-64k (Bai et al., 2024) as homologous models to apply HMG. For CAM, we use LLaMA-2-7B-base-64k to calculate the CAS. We train the LLaMA-2-7B-base-64k based on selected samples as our final model GATEAU-LLaMA. We also find GATEAU can fit in other LLMs in the Appendix I.2, including ChatGLM-3 (Zeng et al., 2023; Bai et al., 2024) and LLaMA-3 series (Grattafiori et al., 2024; Lian, 2024). More details are shown in the Appendix A.

Baselines. We compare our method with multiple SFT data selection baselines. **Cherry Selection** (Li et al., 2024b) and **CaR** (Ge et al., 2024) are state-of-the-art methods to select the influential short instruction-following data. We also use the perplexity score from long-context LLM as guidance to select long instruction-following samples according to Eq. (1), namely **Perplexity Guidance**. More details can be found in the Appendix B.

Evaluation. To gauge the effectiveness of our method, we conduct extensive evaluations on different benchmarks. We use **LongBench-Chat** (Bai et al., 2024) to evaluate the models' ability to follow long instructions, which comprises open-ended questions of 10k-100k in length. We also employ a bilingual and multi-task benchmark **LongBench** (Bai et al., 2023) to evaluate the model's long-context understanding abilities. We conduct evaluations on three tasks following Bai et al. (2024), including Single-Doc QA, Multi-Doc QA, and Summarization. Meanwhile, as aligned models generally produce longer responses, rather than relying solely on the automatic metrics (e.g., ROUGE) to evaluate the results, we follow Bai et al. (2024) to employ GPT-4 to evaluate the model outputs based on their alignment with the ground-truth answers on LongBench. We use **MT-Bench** (Zheng et al., 2023) to measure the models' ability to follow short instructions via GPT-4 rating. To ensure the most stable evaluation results, we use GPT-4 to score twice and average these scores to obtain the final results. More details about evaluation (e.g., the rating prompts) can be found in the Appendix C.

Model	Real-world	Limited
LongBench-Chat		
w/o SFT	10.4	10.4
w/o Long SFT	37.4	36.2
Full - 100%	48.8	50.8
Perplexity Guidance - 10%	52.2	49.0
CaR - 10%	50.8	49.0
Cherry Selection - 10%	53.2	50.8
GATEAU-LLaMA - 10%	**55.4**	**58.0**
Perplexity Guidance - 30%	50.6	51.8
CaR - 30%	48.6	51.4
Cherry Selection - 30%	50.4	52.4
GATEAU-LLaMA - 30%	**57.8**	**55.2**
Perplexity Guidance - 50%	49.8	51.0
CaR - 50%	49.6	51.6
Cherry Selection - 50%	50.6	53.2
GATEAU-LLaMA - 50%	**56.8**	**59.0**
MT-Bench		
w/o SFT	34.6	34.6
w/o Long SFT	53.7	50.5
Full - 100%	54.3	47.7
Perplexity Guidance - 10%	56.1	50.9
CaR - 10%	54.9	49.9
Cherry Selection - 10%	56.8	47.6
GATEAU-LLaMA - 10%	**58.6**	**53.4**
Perplexity Guidance - 30%	55.0	50.2
CaR - 30%	54.3	48.6
Cherry Selection - 30%	54.3	45.8
GATEAU-LLaMA - 30%	**58.8**	**52.9**
Perplexity Guidance - 50%	55.9	49.2
CaR - 50%	54.7	51.2
Cherry Selection - 50%	56.3	49.6
GATEAU-LLaMA - 50%	**57.3**	**54.2**

Table 1: Results (%) on LongBench-Chat and MT-Bench in two different settings.

3.2 Main Results

GATEAU Improves Instruction-Following Capabilities for Both Short and Long Inputs. The results are shown in Table 1 for the LongBench-Chat and MT-Bench benchmarks in two settings. It shows that GATEAU can improve LLMs' capabilities in following both long and short instructions and generating high-quality responses. Compared to indiscriminately using the whole dataset, using the selected subset of the long instruction-following dataset (*GATEAU-LLaMA*) can significantly improve the instruction-following capabilities, e.g., increasing 9% in LongBench-Chat and 6.5% in MT-Bench. Meanwhile, the low performance of *w/o*

Model	Single-Doc QA					Multi-Doc QA					Summarization				
	1-1	1-2	1-3	1-4	Avg	2-1	2-2	2-3	2-4	Avg	3-1	3-2	3-3	3-4	Avg
Real-world Settings															
w/o SFT	33.8	38.0	41.1	34.8	36.9	41.3	37.2	33.3	42.0	38.5	39.2	20.2	37.1	30.9	31.9
w/o Long SFT	58.7	66.7	83.1	79.2	71.9	70.2	53.4	48.7	61.3	58.4	57.3	36.2	55.2	38.4	46.8
Full - 100%	62.8	69.0	83.1	81.3	74.1	71.5	54.8	51.3	66.2	61.0	58.7	39.8	57.6	41.2	49.3
Perplexity Guidance - 10%	62.0	68.8	86.4	85.6	75.7	73.5	59.7	52.1	68.2	63.4	67.6	41.3	**67.0**	44.9	55.2
CaR - 10%	60.3	69.0	86.0	84.8	75.0	69.1	58.3	52.3	68.5	62.1	64.1	41.4	60.3	42.1	52.0
Cherry Selection - 10%	60.8	67.2	86.7	84.3	74.8	71.3	57.8	51.0	69.0	62.3	61.3	40.0	64.8	41.5	51.9
GATEAU-LLaMA - 10%	**63.6**	**69.2**	**86.9**	**87.1**	**76.7**	**74.8**	**60.8**	**53.1**	**69.5**	**64.6**	**67.6**	**42.6**	66.2	**47.8**	**56.1**
Δ compared to Full - 100%	+0.8	+0.2	+3.8	+5.8	+2.7	+3.3	+6.0	+1.8	+3.3	+3.6	+8.9	+2.8	+8.6	+6.6	+6.7
Perplexity Guidance - 30%	62.8	67.3	86.2	82.6	74.7	72.3	59.3	50.8	67.8	62.6	62.3	41.7	64.8	42.7	52.9
CaR - 30%	61.3	67.3	86.4	85.3	75.1	68.3	58.3	53.2	66.8	61.7	64.6	39.7	60.7	41.2	51.6
Cherry Selection - 30%	62.0	66.8	87.1	84.3	75.1	74.3	59.3	52.7	68.7	63.8	62.3	40.5	64.6	44.4	53.0
GATEAU-LLaMA - 30%	**63.0**	**70.8**	**87.6**	**85.8**	**76.8**	**75.7**	**61.0**	**55.7**	**69.5**	**65.5**	**67.5**	**44.7**	**65.9**	**47.4**	**56.4**
Δ compared to Full - 100%	+0.2	+1.8	+4.5	+4.5	+2.8	+4.2	+6.2	+4.4	+3.3	+4.5	+8.8	+4.9	+8.3	+6.2	+7.1
Perplexity Guidance - 50%	63.1	68.1	87.8	82.1	75.3	74.2	59.2	52.5	69.2	63.8	64.7	41.1	65.7	42.1	53.4
CaR - 50%	60.0	66.3	85.6	84.2	74.0	70.7	55.8	54.3	68.2	62.3	64.4	41.1	60.8	40.3	51.7
Cherry Selection - 50%	62.8	65.5	86.2	82.8	74.3	72.2	56.8	52.7	67.8	62.4	64.6	39.4	64.1	42.1	52.6
GATEAU-LLaMA - 50%	**63.5**	**70.3**	**89.7**	**86.5**	**77.5**	**75.3**	**60.8**	**53.5**	**68.5**	**64.5**	**65.1**	**41.6**	**65.9**	**46.1**	**54.7**
Δ compared to Full - 100%	+0.7	+1.3	+6.6	+5.2	+3.5	+3.8	+6.0	+2.2	+2.3	+3.6	+6.4	+1.8	+8.3	+4.9	+5.4
Limited Short Instruction Data Settings															
w/o SFT	33.8	38.0	41.1	34.8	36.9	41.3	37.2	33.3	42.0	38.5	39.2	20.2	37.1	30.9	31.9
w/o Long SFT	62.3	70.8	88.5	82.7	76.1	72.8	60.6	51.8	67.3	63.1	64.7	41.1	61.4	41.6	52.2
Full - 100%	58.7	69.7	85.8	83.0	74.3	70.5	58.7	50.8	67.8	62.0	59.6	38.4	59.6	43.3	50.2
Perplexity Guidance - 10%	62.8	69.2	89.3	85.7	76.8	73.8	59.1	54.1	**71.1**	64.5	69.8	45.8	65.7	50.1	57.9
CaR - 10%	62.8	68.3	88.0	82.7	75.5	71.8	58.0	52.7	68.8	62.8	65.5	42.0	61.8	43.1	53.1
Cherry Selection - 10%	62.8	69.8	86.7	85.7	76.3	72.0	58.7	52.5	69.3	63.1	63.2	43.3	60.1	46.4	53.3
GATEAU-LLaMA - 10%	**64.8**	**74.7**	**89.8**	**86.5**	**79.0**	**75.2**	**61.2**	**54.6**	70.0	**65.3**	**71.1**	**47.3**	**67.0**	**54.2**	**59.9**
Δ compared to Full - 100%	+6.1	+5.0	+4.0	+3.5	+4.7	+4.7	+2.5	+3.8	+2.2	+3.3	+11.5	+8.9	+7.4	+10.9	+9.7
Perplexity Guidance - 30%	62.5	71.8	88.2	83.8	76.6	74.6	58.5	53.5	69.3	64.0	67.5	44.0	64.7	50.4	56.7
CaR - 30%	60.8	70.7	88.4	81.8	75.4	73.0	59.0	53.5	68.5	63.5	64.1	40.9	62.3	45.8	53.3
Cherry Selection - 30%	62.8	71.7	88.9	**87.5**	77.7	70.3	58.7	50.3	68.2	61.9	62.9	43.5	65.2	44.6	54.1
GATEAU-LLaMA - 30%	**64.8**	**73.0**	**89.3**	86.2	**78.3**	**74.7**	**61.0**	**54.2**	**69.8**	**64.9**	**70.8**	**46.0**	**66.4**	**51.4**	**58.7**
Δ compared to Full - 100%	+6.1	+3.3	+3.5	+3.2	+4.0	+4.2	+2.3	+3.4	+2.0	+3.0	+11.2	+7.6	+6.8	+8.1	+8.4
Perplexity Guidance - 50%	61.5	68.3	85.1	82.8	74.4	72.3	59.3	52.0	67.7	62.8	60.2	40.9	58.6	42.3	50.5
CaR - 50%	62.3	68.1	86.9	80.1	74.4	71.0	58.7	52.8	68.0	62.6	64.4	41.2	61.1	45.6	53.1
Cherry Selection - 50%	61.2	69.7	86.2	83.7	75.2	69.7	56.8	49.5	66.2	60.6	64.1	41.8	60.5	43.7	52.5
GATEAU-LLaMA - 50%	**63.7**	**71.8**	**87.1**	**84.7**	**76.8**	**74.0**	**60.0**	**53.8**	**69.0**	**64.2**	**66.1**	**43.9**	**62.4**	**46.4**	**54.7**
Δ compared to Full - 100%	+5.0	+2.1	+1.3	+1.7	+2.5	+3.5	+1.3	+3.0	+1.2	+2.3	+6.5	+5.5	+2.8	+3.1	+4.5

Table 2: GPT-4 evaluation results (%) on LongBench in Real-world Settings. We use the ID to represent the dataset in LongBench, e.g., 1-1 is the ID of the NarrativeQA dataset. More details can be found in the Appendix C.2. Automatic metrics evaluation results (%) are shown in Table 5.

Long SFT in LongBench-Chat indicates that using long SFT data is important for the performance in handling the instructions with long input contexts. The results also show that our method GATEAU achieves consistently better performance in varying ratios of used long instruction-following samples compared with other baselines, indicating the effectiveness of our method. Compared with baselines focusing on short SFT samples (*CaR* and *Cherry Selection*), GATEAU can identify samples enriched with long-range dependency relations more effectively and help LLMs to achieve better long context alignment. We also observe that the selection of long instruction-following samples aids in augmenting the instruction-following capabilities for short inputs. We conjecture that handling complex tasks (i.e., long input contexts) contributes to handling the easy ones (i.e., short input contexts).

GATEAU Enhances the Long-Context Understanding Capabilities. The results are shown in Table 2 and Table 5 (in the Appendix) for the LongBench benchmark. Our methods achieve consistent and remarkable performance gains in different settings and evaluation methods. We show the improved scores (Δ *compared to Full-100%*) compared to indiscriminately using the whole dataset (*Full-100%*), indicating that GATEAU helps LLM to better understand the long input contexts. We also find that the baselines focusing on the selection of short instruction-following data (*CaR* and *Cherry Selection*) hold inferior results, sometimes even worse than using the whole dataset (*Full-100%*). This can be attributed to these methods are not designed for long context alignment and understanding, thus failing to select the samples enriched with long-range dependency relations.

3.3 Analysis

Ablation Study. To evaluate the effectiveness of our proposed GATEAU, we also conduct the ab-

Model	LongBench			LongBench-Chat	MT-Bench		
	Single-Doc QA	Multi-Doc QA	Summarization	Avg	First-turn	Second-turn	Avg
			Real-world Settings				
GATEAU-LLaMA - 13B - 50%	**40.2**	**27.1**	**25.7**	**61.4**	**66.8**	**55.3**	**61.1**
-w/o Contextual Awareness Measurement	38.1	25.8	24.6	60.2	66.2	55.0	60.6
-w/o Homologous Models' Guidance	38.6	26.0	25.1	60.6	66.0	54.6	60.3
-w/o Data Selection (i.e., Full - 100%)	33.6	16.7	24.4	59.4	66.0	54.1	59.6
GATEAU-LLaMA - 7B - 50%	**38.9**	**25.8**	**25.5**	**56.8**	**64.1**	**50.4**	**57.3**
-w/o Contextual Awareness Measurement	38.4	24.3	25.1	53.2	61.7	51.5	56.6
-w/o Homologous Models' Guidance	38.6	24.5	24.9	52.8	63.1	49.3	56.3
-w/o Data Selection (i.e., Full - 100%)	36.1	22.3	23.8	48.8	60.0	48.7	54.3
			Limited Short Instruction Data Settings				
GATEAU-LLaMA - 13B - 50%	**32.1**	**19.1**	**25.3**	**62.6**	**66.0**	**51.5**	**58.8**
-w/o Contextual Awareness Measurement	31.4	18.4	24.7	59.6	64.2	50.3	57.3
-w/o Homologous Models' Guidance	30.8	18.6	25.0	60.4	63.6	50.6	57.1
-w/o Data Selection (i.e., Full - 100%)	30.4	17.8	24.5	54.2	61.0	49.8	55.4
GATEAU-LLaMA - 7B - 50%	**31.0**	**18.1**	**25.3**	**59.0**	**64.2**	**44.1**	**54.2**
-w/o Contextual Awareness Measurement	28.5	17.5	24.7	53.2	61.3	42.4	51.8
-w/o Homologous Models' Guidance	28.7	17.3	24.6	54.4	56.1	45.0	50.6
-w/o Data Selection (i.e., Full - 100%)	27.2	16.1	24.5	50.8	54.5	40.9	47.7

Table 3: Results (%) of ablation and scalability study. We show automatic metrics evaluation results on LongBench.

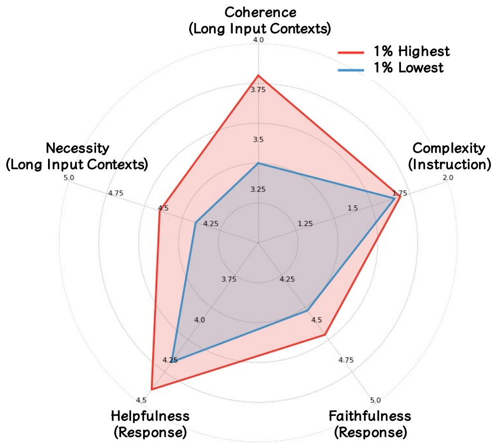

Figure 2: The comparison between samples with top 1% and least 1% scored by GATEAU.

lation study in Table 3. We can find that HMG and CAM can both enhance LLMs' instruction-following and long-context understanding capabilities. This indicates the effectiveness of GATEAU, and using the two proposed methods can further improve the performance as they separately measure the difficulty from two different perspectives.

Scalability Study. We explore whether GATEAU can fit in larger LLMs in Table 3. We apply our method on the Llama-2-13B-base series and fine-tune Llama-2-13B-base-64k (Bai et al., 2024) using the selected samples. Compared to the 7B-scale *GATEAU-LLaMA-7B*, the 13B *GATEAU-LLaMA-13B* shows consistent improvements on three benchmarks. This indicates that GATEAU scales effectively to larger-scale models.

General Characteristics of Selected Samples. We delve into whether the selected samples based on GATEAU align with known characteristics of high-quality data as shown in Figure 2. We select 100 samples with the 1% highest scores and 100 samples with the 1% lowest scores. Utilizing GPT-4, we evaluate each sample on five aspects: the coherence of long input contexts, the necessity of long input contexts, helpfulness of response, the faithfulness of response, and the complexity of instruction. A sample with a higher score tends to be more high-quality, especially the long input contexts and the response. The complexity of instruction, in particular, shows a mere improvement compared to other characteristics. We evaluate the whole dataset on this characteristic and find that all samples show consistently low scores, which may be due to the limitation of the synthetic dataset. More details are shown in the Appendix D.

Variation of Abilities under Different Context Lengths. Figure 3 shows the macro-average results (%) on data in length ranges of 0-4k, 4k-8k, and 8k+. We find that GATEAU improves the performance in long-context scenarios (i.e., 4k-8k and 8k+). Indiscriminately using the full long SFT dataset (*Full-100%*) even hinders the performance in long-context scenarios compared to solely using the short instruction-following dataset (*-w/o Long SFT*). This confirms the necessity of selecting influential samples and the effectiveness of GATEAU.

Human Evaluation. We conduct a human evaluation on the LongBench-Chat. We invite three participants (Ph.D. students or Master students) to compare the responses generated by the models. For each comparison, three options are given (Win, Tie, and Loss), and the final result is determined by majority voting. Figure 4 shows the effectiveness of our method, i.e., our trained models show consistent preference from participants. Details can be

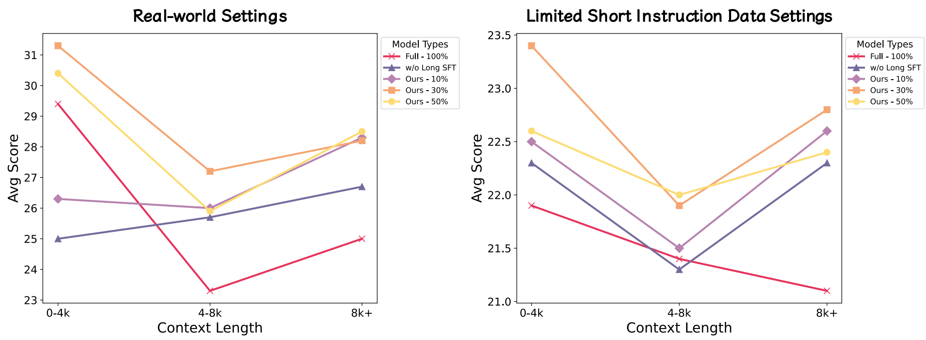

Figure 3: Automatic metrics evaluation results (%) under different context lengths on LongBench.

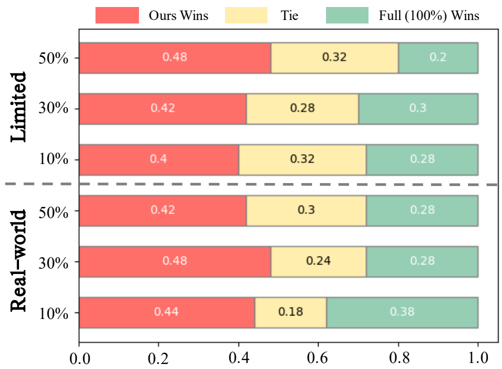

Figure 4: Human evaluation in two settings.

found in the Appendix E.

Needle in the Haystack Test. We conduct a "Needle in A HayStack" test in the Appendix F to show GATEAU can fully utilize the information.

Parameter Study and Case Study. We also conduct the parameter study and a practical case study in the Appendix G and Appendix H.

Discussion. We further discuss some possible concerns about GATEAU in the Appendix I. For example, we report the execution time of GATEAU.

4 Related Work

Long Context Alignment. Aligning the LLMs to follow user instructions has been extensively studied in short-context scenarios (Taori et al., 2023; Wang et al., 2023a,b). However, excessively long contexts present unique challenges for long context alignment. Li et al. (2023); Tworkowski et al. (2023); Xiong et al. (2024) construct the long SFT dataset by concatenating short SFT samples. Yet, simply concatenating unrelated sentences can not effectively simulate the long-range dependency relations for long-context tasks. Thus, Yang (2023); Chen et al. (2024b); Bai et al. (2024) construct long SFT data by collecting long-context materials as inputs and querying Claude to get the response. However, using these synthetic data without a clear strategy for ensuring data quality may introduce low-quality samples (e.g., samples without meaningful long-range dependency relations). Training LLMs on such low-quality samples can ultimately constrain their final performance.

Data Selection for Alignment. As Zhou et al. (2023) states *less is more for alignment*, many works attempt to select influential samples to empower the LLMs' instruction-following capabilities. Chen et al. (2023a); Liu et al. (2024) attempt to utilize the feedback from well-aligned closed-source LLMs to select samples. Cao et al. (2024); Li et al. (2024b); Ge et al. (2024); Xia et al. (2024) try to utilize the well-designed metrics (e.g., complexity) based on open-source LLMs to rank and select the samples. Meanwhile, Li et al. (2024c); Zhang et al. (2024) attempt to utilize the guidance from in-context learning. However, these methods only focus on selecting short SFT data, ignoring the unique challenge in long context alignment, i.e., selecting the samples enriched with meaningful long-range dependency relations.

5 Conclusion

In this study, we introduce **GATEAU**, a new novel framework designed to select influential samples for long context alignment. Different from previous studies, we attempt to address the unique challenge in long context alignment, i.e., modeling long-range dependencies. To measure the richness of long-range dependency relations in long SFT samples, GATEAU separately measures the difficulty of generating corresponding responses and understanding lengthy inputs due to the long-range dependencies. Trained on these selected influential samples, our model achieves better alignment. Extensive experiments consistently show the effectiveness of GATEAU compared to other methods.

Limitations

Although empirical experiments have confirmed the efficacy of the proposed GATEAU, three major limitations remain. Firstly, our proposed HMG requires two homologous models with different context windows, thus limiting the range of models we can use to conduct more experiments in our paper. However, in practical scenarios, training a powerful long-context LLM always involves homologous models with different context windows (though these models may not be open-sourced). This is because existing LLMs are often initially pre-trained on a large-scale corpus with smaller context windows due to device limitations, they then conduct continual pre-training to extend the window size. Therefore, our method still remains effective in real-world scenarios. Secondly, GATEAU is designed to improve overall performance in instruction-following and long-context understanding tasks. It is not suitable to improve the performance of LLMs in a targeted capability or task, e.g., mathematical questions. Lastly, The size of the context window is a critical factor, and it is often determined by the continual training stage of the open-source base models. If the data exceeds the context window of the base model, the effectiveness of GATEAU will be limited. However, with the advancement of open-source models, the current context window size is rapidly increasing (Yang et al., 2025). Thus, based on these open-sourced long-context LLMs, GATEAU can be further used to select longer samples.

References

Anthropic. 2023. Anthropic: Introducing claude 2.1.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2024. Instruction mining: Instruction data selection for tuning large language models. *Preprint*, arXiv:2307.06290.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. 2024a. Long context is not long at all: A prospector of long-dependency data for large language models. *Preprint*, arXiv:2405.17915.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *Preprint*, arXiv:2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024b. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Hao Yang, and Tong Xiao. 2024. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*.

Gkamradt. 2023. Llmtest_needleinahaystack.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen,

Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-

say, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. *Preprint*, arXiv:2406.16008.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source LLMs truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024a. A survey on deep active learning: Recent advances and new frontiers. *Preprint*, arXiv:2405.00334.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7595–7628, Mexico City, Mexico. Association for Computational Linguistics.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024c. One shot learning as instruction data prospector for large language models. *Preprint*, arXiv:2312.10302.

Wing Lian. 2024. Using pose to extend llama's context length from 8k to 64k.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

OpenAI. 2023. Openai: Gpt-4.

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

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling. *Preprint*, arXiv:2307.03170.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. 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)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Wenhao Wu, Yizhong Wang, Yao Fu, Xiang Yue, Dawei Zhu, and Sujian Li. 2024. Long context alignment with short instructions and synthesized positions. *Preprint*, arXiv:2405.03939.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663, Mexico City, Mexico. Association for Computational Linguistics.

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. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *Preprint*, arXiv:2501.15383.

Jianxin Yang. 2023. Longqlora: Efficient and effective method to extend context length of large language models. *Preprint*, arXiv:2311.04879.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Qi Zhang, Yiming Zhang, Haobo Wang, and Junbo Zhao. 2024. Recost: External knowledge guided data-efficient instruction tuning. *Preprint*, arXiv:2402.17355.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Appendix

This appendix is organized as follows.

- In Section A, we report the training details, e.g., training datasets and hyperparameters.
- In Section B, we go into detail about the baselines used in our experiments.
- In Section C, we show the details of evaluations, e.g., the introduction of the used benchmarks and evaluation prompts.
- In Section D, we list the details of the general characteristics of selected samples.
- In Section E, we show the implementation details of human evaluation.
- In Section F, we conduct a "Needle in A HayStack" experiment to test the ability to utilize information from different positions.
- In Section G, we conduct experiments to explore the impact of hyperparameters.
- In Section H, we come up with a practical case study to show the effectiveness of GATEAU.
- In Section I, we discuss some possible questions, including execution time (Sec. I.1), experiments in other LLMs (Sec. I.2), experiments in other long SFT datasets (Sec. I.3), the diversity of selected samples (Sec. I.4), further exploration of HMG (Sec. I.5), orthogonality with baselines (Sec. I.6), and mix-quality training strategy (Sec. I.7).
- In Section J, we show the difference between samples with high or low scores.

A Training

Training Datasets. LongAlign dataset (Bai et al., 2024) is developed by using collected long sequences from 9 sources and applying the Self-Instruct (Wang et al., 2023b) approach with long-context LLM Claude 2.1 (Anthropic., 2023). Though initially competitive, its dependence on Claude 2.1 synthesized data may lead to quality concerns. For the ShareGPT dataset (Chiang et al., 2023), we filter the sample with an empty response.

Training Details. LLaMA-2-7B-base-4k is an open-sourced LLM with a context window of 4k tokens. To extend context windows, Bai et al. (2024) proposes LLaMA-2-7B-base-64k by modifying the RoPE position encoding (Su et al., 2023) and applying continual training on data with lengths under 64k, for a total of 10 billion tokens. Meanwhile, for LLaMA-2-7B-base-4k, we expand the base frequency b of the RoPE position encoding by 200 times (from 10,000 to 2,000,000) to extend the context windows and avoid the model conducting extreme perplexity score (>1,000) in HMG. For CAM, we use LLaMA-2-7B-base-64k to calculate the score and use selected samples to train the LLaMA-2-7B-base-64k as our final model.

Devices and Hyperparameters. All models are trained with 8xA800 80G GPUs and DeepSpeed+ZeRO3+CPU offloading. We use BF16 in both our training and inference. The models can be trained with a maximum length of 64k tokens without GPU memory overflow. We set the maximum length of the training data to 64k, with any data exceeding this length being truncated from the right side. We keep the same maximum length in the HMG and CAM, but truncate from the left side to keep the original responses. We set the batch size to 8, with a gradient accumulation step of 12 for all the training methods. We train 2 epochs on the training data. We set the learning rate as 2e-5 and use AdamW (Loshchilov and Hutter, 2019) as our optimizer. The β_1 and β_2 in the AdamW optimizer are set to 0.9 and 0.95. Meanwhile, the length of segment L is set to 128 in CAM. Hyperparameter α in Eq. (6) is set to 0.7 in Limited Short Instruction Data Settings and 0.8 in Real-world Settings.

B Baselines

We will detail the baselines in our experiments.

w/o SFT. For w/o SFT, we directly utilize the base model without alignment to get the experiment results, i.e., the results of LLaMA-2-7B-base-64k.

w/o Long SFT. For baseline w/o Long SFT, we only use the short instruction data from the ShareGPT dataset to apply the supervised fine-tuning stage for alignment. The number of short instruction samples used from the ShareGPT dataset is determined by the different settings.

Full - 100%. For baseline Full - 100%, we use the full data of the LongAlign dataset, including 10k long instruction samples, to conduct the SFT for alignment. The number of short instruction samples used from the ShareGPT dataset is determined by the different settings.

Perplexity Guidance. We use the perplexity score from LLM as guidance to select long instruction-following samples according to Eq. (1). We select the long instruction-following samples with the highest perplexity scores as the most influential samples to train the model. Meanwhile, the number of short instruction samples used from ShareGPT is determined by the different settings.

CaR. This work (Ge et al., 2024) proposes a straightforward yet efficacious short instruction-following selection framework. This method first selects a subset that ensures the retention of a large number of high-quality instructions and then supplements a small number of high-quality instructions from each cluster to enhance the diversity of the data while preserving instruction quality. Specifically, this work first employs a small-scale trained reward model to get the score of the samples. Meanwhile, the cluster model is employed to cluster all candidate instruction pairs into k clusters Finally, all instruction pairs are sorted based on their scores, and the top n_1 pairs are selected; within each cluster, instruction pairs are sorted by score, and the top n_2 pairs are chosen. A high-quality sub-dataset with preserved diversity is then curated by duplicating $n_1 + k \times n_2$ pairs of instructions. We directly use the same reward model and hyperparameters to select long samples. Meanwhile, the number of short samples used from ShareGPT is determined by the different settings.

Cherry Selection. Li et al. (2024b) proposes a method for autonomously sifting through expansive open-source short instruction-following datasets to discover the most influential training samples. At the heart of this method is the hypothesis that during their preliminary training stages with carefully chosen instruction data, LLMs can develop an intrinsic capability to discern instructions. This foundational understanding equips them with the discernment to assess the quality of broader datasets, thus making it possible to estimate the instruction-following difficulty in a self-guided manner. To estimate the difficulty of a given example, this work proposes a novel metric called Instruction-Following Difficulty (IFD) score in which both models' capability to generate a response to a given instruction and the models' capability to generate a response directly are measured and compared. This method quantifies the challenge each sample presents to the model and utilizes selected data with standout IFD scores to hone the model. We apply this method to select the long instruction-following samples as the baseline. Meanwhile, the number of short instruction samples used from ShareGPT is determined by the different settings.

C Evaluations

C.1 LongBench-Chat

Evaluation Data. LongBench-Chat focuses on assessing LLMs' instruction-following capability under the long context. LongBench-Chat includes 50 long context real-world queries ranging from 10k to 100k in length. It covers diverse aspects of instruction-following abilities such as reasoning, coding, summarization, and multilingual translation over long contexts. It consists of 40 tasks in English and 10 in Chinese. GPT-4 (OpenAI, 2023) is employed to give a score on a scale of 10 to the machine-generated responses based on the annotated ground-truths. Bai et al. (2024) finds that with their proposed few-shot evaluation prompting, GPT-4's correlation with human annotations not only aligns but also surpasses the level of agreement among human annotators.

Evaluation Prompts. LongBench-Chat employs GPT-4 to score the model's response in 1-10 based on a given human-annotated reference answer and few-shot scoring examples for each question. We use the same prompt as LongBench-Chat to get GPT-4's evaluation shown in Figure 8.

C.2 LongBench

Evaluation Data. LongBench is the first bilingual, multitask benchmark tailored for long context understanding. LongBench includes different languages (Chinese and English) to provide a more comprehensive evaluation of the large models' bilingual capabilities in long-context understanding. Detailed statistics of the used dataset in LongBench can be found in Table 4.

Evaluation Prompts. We conduct GPT-4 evaluation for LongBench as Bai et al. (2024). As aligned models generally produce longer responses, rather than relying solely on the original automatic metrics (e.g., ROUGE) to evaluate the models' replies, we employ GPT-4 to assess the model outputs based on their alignment with the ground-truth answers on LongBench. For the first two QA tasks, the prompt for the GPT-4 evaluator is the same as Bai et al. (2024), shown in Figure 9. The prompt for GPT-4 evaluation on summarization tasks is the same as Bai et al. (2024), shown in Figure 10.

Automatic Metrics Evaluation Results We show the detailed automatic metric evaluation results on

Dataset	ID	Source	Avg len	Auto Metric	Language	#data
Single-Document QA						
NarrativeQA	1-1	Literature, Film	18,409	F1	English	200
Qasper	1-2	Science	3,619	F1	English	200
MultiFieldQA-en	1-3	Multi-field	4,559	F1	English	150
MultiFieldQA-zh	1-4	Multi-field	6,701	F1	Chinese	200
Multi-Document QA						
HotpotQA	2-1	Wikipedia	9,151	F1	English	200
2WikiMultihopQA	2-2	Wikipedia	4,887	F1	English	200
MuSiQue	2-3	Wikipedia	11,214	F1	English	200
DuReader	2-4	Baidu Search	15,768	Rouge-L	Chinese	200
Summarization						
GovReport	3-1	Government report	8,734	Rouge-L	English	200
QMSum	3-2	Meeting	10,614	Rouge-L	English	200
MultiNews	3-3	News	2,113	Rouge-L	English	200
VCSUM	3-4	Meeting	15,380	Rouge-L	Chinese	200

Table 4: An overview of the dataset statistics in LongBench. 'Source' denotes the origin of the context. 'Avg len' is computed using the number of words for the English datasets and the number of characters for the Chinese datasets.

LongBench in Table 5. Meanwhile, we can see that using 30% of the whole long instruction-following dataset (*GATEAU-LLaMA-30%*) can achieve the best performance of LongBench in two different settings. This is because of its ability to maintain an optimal balance between the volume and quality of the long instruction-following samples it utilizes, leading to the most desirable results.

C.3 MT-Bench

Evaluation Data. MT-Bench is a comprehensive benchmark comprising 80 multi-turn questions. It is designed to assess the ability to engage in multi-turn conversations and follow instructions. The benchmark covers common use cases and emphasizes challenging questions to effectively differentiate among models. It is meticulously designed to distinguish chatbots based on their fundamental capabilities, which include writing, roleplay, extraction, reasoning, mathematics, coding, knowledge in STEM fields, and knowledge in the humanities and social sciences. MT-Bench prompts large language models, such as GPT-4, to serve as judges and evaluate the quality of the models' responses. Zheng et al. (2023) conducted a series of experiments and found that LLM judges like GPT-4 can align impressively well with both controlled and crowd-sourced human preferences, achieving over 80% agreement. For each turn, GPT-4 assigns a score on a scale of 1 to 10. We then report the average score across all turns.

More Detailed Results. We show the detailed results of MT-Bench in Table 6.

C.4 GPT-4 Version

For all the evaluations using the GPT-4 (including LongBench-Chat, LongBench, MT-Bench, and Needle in the Haystack test), we used `GPT-4 API` in August 2024 to ensure that we keep the same as Bai et al. (2024). According to the documents from OpenAI, `GPT-4 API` points to `GPT-4-0613 API`.

D General Characteristics of Selected Samples from GATEAU

Utilizing GPT-4, we evaluate each sample on five aspects: the coherence of long input contexts, the necessity of long input contexts, helpfulness of response, the faithfulness of response, and the complexity of instruction. Different from the previous GPT-4 evaluation detailed in Appendix C.4, we use `GPT-4-Turbo API` (now points to `GPT-4-Turbo-2024-04-09`) as our evaluator, as this version of API has larger context window to conduct the more correct evaluation for our long input contexts. To ensure stable evaluation results, we use GPT-4 to score twice on 200 selected samples, and then average these scores to obtain the final results. The prompt for GPT-4 evaluation on different characteristics can be found in Figure 11, Figure 12, Figure 13, Figure 14, and Figure 15.

E Human Evaluation

During the human evaluation, the participants follow the principles in Figure 16 to make the decision.

Model	Single-Doc QA					Multi-Doc QA					Summarization				
	1-1	1-2	1-3	1-4	Avg	2-1	2-2	2-3	2-4	Avg	3-1	3-2	3-3	3-4	Avg
Real-world Settings															
w/o SFT	0.9	3.9	6.4	3.6	3.7	7.3	8.7	2.1	15.4	8.4	23.9	6.2	14.0	1.8	11.5
w/o Long SFT	16.8	29.1	45.8	48.7	35.1	27.8	17.6	11.4	25.3	20.5	27.4	23.3	27.8	14.3	23.2
Full - 100%	18.4	29.9	46.1	**49.9**	36.1	27.1	20.8	11.2	30.0	22.3	28.7	24.0	26.7	15.9	23.8
Perplexity Guidance - 10%	19.9	32.0	46.6	45.8	36.1	22.1	23.2	10.4	30.3	21.5	31.3	23.8	26.0	17.7	24.7
CaR - 10%	16.9	24.1	47.6	42.3	32.7	22.1	19.8	11.3	30.0	20.8	**31.9**	23.1	26.2	18.6	25.0
Cherry Selection - 10%	19.9	30.8	47.2	43.1	35.3	25.2	21.4	10.6	28.3	21.4	30.0	24.1	25.1	17.0	24.1
GATEAU-LLaMA - 10%	**23.5**	**34.2**	**49.6**	**54.5**	**40.5**	**28.7**	**25.0**	**12.1**	**30.5**	**24.0**	31.2	**24.7**	**26.9**	**18.9**	**25.4**
△ compared to Full - 100%	+5.1	+4.3	+3.5	+4.6	+4.4	+1.6	+4.2	+0.9	+0.5	+1.8	+2.5	+0.7	+0.2	+3.0	+1.6
Perplexity Guidance - 30%	21.1	33.6	46.1	46.7	36.9	23.4	21.0	10.1	30.1	21.2	30.2	24.7	26.4	18.9	25.1
CaR - 30%	18.0	24.4	46.9	45.0	33.6	25.4	20.8	14.4	29.4	22.5	30.1	24.8	26.5	18.2	24.9
Cherry Selection - 30%	20.5	33.1	48.0	51.0	38.2	26.7	20.4	13.5	29.1	22.4	30.4	24.1	26.9	17.7	24.8
GATEAU-LLaMA - 30%	**23.7**	**34.1**	**49.6**	**54.6**	**40.5**	**30.1**	**23.8**	**14.9**	**30.4**	**24.8**	**30.5**	**24.9**	**27.2**	**18.9**	**25.4**
△ compared to Full - 100%	+5.3	+4.2	+3.5	+4.7	+4.4	+3.0	+3.0	+3.7	+0.4	+2.5	+1.8	+0.9	+0.5	+3.0	+1.6
Perplexity Guidance - 50%	19.2	32.8	50.1	49.5	37.9	27.1	23.1	12.1	31.1	23.4	**31.5**	24.1	27.1	18.7	25.4
CaR - 50%	17.6	24.5	47.6	44.7	33.6	29.3	19.4	17.3	29.6	23.9	30.3	23.7	26.0	18.2	24.6
Cherry Selection - 50%	19.0	32.6	51.7	49.6	38.2	26.2	23.9	13.5	30.4	23.5	30.5	23.8	26.9	18.8	25.0
GATEAU-LLaMA - 50%	**20.2**	**33.4**	**52.1**	49.8	**38.9**	**30.7**	**25.2**	**15.0**	**32.5**	**25.8**	31.3	**24.6**	**27.1**	**18.8**	**25.5**
△ compared to Full - 100%	+1.8	+3.5	+6.0	-0.1	+2.8	+3.6	+4.4	+3.8	+2.5	+3.6	+2.6	+0.6	+0.4	+2.9	+1.6
Limited Short Instruction Data Settings															
w/o SFT	0.9	3.9	6.4	3.6	3.7	7.3	8.71	2.1	15.4	8.4	23.9	6.2	14.0	1.78	11.5
w/o Long SFT	13.8	19.2	38.3	37.1	27.1	15.2	14.7	8.2	25.7	16.0	29.4	24.4	25.0	19.3	24.5
Full - 100%	14.7	20.1	37.0	37.0	27.2	15.4	13.8	8.6	26.7	16.1	29.3	24.5	25.6	18.6	24.5
Perplexity Guidance - 10%	15.4	19.2	41.0	37.8	28.4	15.0	14.8	**8.5**	25.6	16.0	28.8	23.9	26.1	17.8	24.2
CaR - 10%	11.5	17.7	37.7	30.0	24.2	15.6	12.5	8.4	25.9	15.6	29.3	24.1	26.2	18.2	24.5
Cherry Selection - 10%	14.6	19.2	41.2	37.7	28.2	15.7	14.6	7.6	25.3	15.8	29.4	24.1	26.0	17.8	24.3
GATEAU-LLaMA - 10%	**17.1**	**20.7**	**43.4**	**38.3**	**29.9**	**19.9**	**18.5**	8.2	**26.8**	**18.4**	**29.6**	**24.3**	**26.3**	**18.3**	**24.6**
△ compared to Full - 100%	+2.4	+0.6	+6.4	+1.3	+2.7	+4.5	+4.7	-0.4	+0.1	+2.2	+0.3	-0.2	+0.7	-0.3	+0.1
Perplexity Guidance - 30%	15.3	**20.6**	42.3	38.2	29.1	17.4	15.9	8.6	27.5	17.4	28.3	24.3	25.7	19.0	24.3
CaR - 30%	13.6	18.3	41.0	30.5	25.9	16.7	15.8	9.4	27.0	17.2	28.8	24.3	25.3	18.4	24.2
Cherry Selection - 30%	15.9	19.5	42.3	**39.0**	29.2	17.3	16.3	9.3	26.2	17.3	29.2	**25.0**	26.1	18.2	24.6
GATEAU-LLaMA - 30%	**17.7**	20.4	**43.1**	38.6	**29.9**	**22.5**	**18.5**	**11.6**	**27.7**	**20.1**	**30.5**	24.3	**26.8**	**19.7**	**25.3**
△ compared to Full - 100%	+3.0	+0.3	+6.1	+1.6	+2.7	+7.1	+4.7	+3.0	+1.0	+4.0	+1.2	-0.2	+1.2	+1.1	+0.8
Perplexity Guidance - 50%	16.4	20.6	39.1	37.1	28.3	16.7	16.4	8.2	26.0	16.8	29.3	25.1	25.2	19.1	24.7
CaR - 50%	12.1	18.1	40.4	30.4	25.3	17.3	15.1	9.0	26.3	16.9	28.3	23.6	25.1	18.9	24.0
Cherry Selection - 50%	15.5	19.5	38.9	37.3	27.8	15.4	16.3	8.8	26.1	16.7	**30.6**	24.8	25.3	18.9	24.9
GATEAU-LLaMA - 50%	**18.5**	**22.5**	**43.9**	**39.1**	**31.0**	**17.9**	**16.7**	**9.6**	**28.0**	**18.1**	30.1	**25.3**	**26.6**	**19.4**	**25.3**
△ compared to Full - 100%	+3.8	+2.4	+6.9	+2.1	+3.8	+2.5	+2.9	+1.0	+1.3	+1.9	+0.8	+0.8	+0.9	+0.8	+0.8

Table 5: Automatic metrics evaluation results (%) on LongBench in two different settings. We use the ID to represent the dataset in LongBench, e.g., 1-1 is the ID of the NarrativeQA dataset.

If the final result can not be determined by majority voting, we will hold a discussion among the participants and vote on the result again.

F Needle in the Haystack Test

We conduct "Needle in A HayStack" experiments in Figure 5 to test the model's ability to utilize information from 10 different positions. This task asks for the model to retrieve a piece of fact (the 'needle') that is inserted in the middle (positioned at a specified depth percent) of a long context window (the 'haystack'). These results show that GATEAU can help LLMs to utilize information from different positions within long texts, resulting in a decrease in the model's retrieval error. Following the same original configuration as the original method (Gkamradt, 2023), we use "The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day." as the needle fact, and Paul Graham's essays as the long haystack context. We use the same prompt as Bai et al. (2024): "What is the best thing to do in San Francisco? Here is the most relevant sentence in the context:".

G Parameter Study

As shown in Figure 6, we conduct experiments to explore the impact of the important hyperparameter α in Eq. (6). We report the results of *GATEAU-LLaMA - 50%* on LongBench-Chat in two settings. Overall, although the choice of different α will have some impact on the LLM's performance, the performance will always be improved over the baseline *Full-100%*, i.e., using the whole training dataset without data selection. Meanwhile, we also find that using both the Homologous Model's Guidance and Contextual Awareness Measurement will further improve the performance than only using one of them. This is because the Homologous Model's Guidance and Contextual Awareness Measurement attempts to

Model	First-turn	Second-turn	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities
Real-world Settings										
w/o SFT	43.5	25.6	44.5	44.0	35.0	16.5	18.0	28.0	42.0	48.8
w/o Long SFT	60.0	47.4	73.8	72.0	44.0	22.0	25.5	42.5	63.0	86.5
Full - 100%	60.0	48.7	78.5	70.3	45.5	19.0	29.0	42.0	67.5	83.0
Perplexity Guidance - 10%	63.1	48.9	68.7	67.0	43.5	26.5	33.2	50.5	69.8	88.5
CaR - 10%	59.8	50.0	76.5	75.3	44.5	24.5	24.8	43.5	64.2	84.9
Cherry Selection - 10%	63.0	50.5	74.5	73.8	42.3	25.0	32.5	48.3	70.3	87.5
GATEAU-LLaMA - 10%	**63.1**	**54.1**	73.8	**79.2**	43.8	**26.5**	27.8	46.0	**77.0**	**94.8**
Perplexity Guidance - 30%	62.1	47.8	69.0	63.7	46.0	28.0	28.4	49.0	72.5	82.2
CaR - 30%	60.0	48.6	79.3	77.0	38.5	21.0	19.8	44.0	71.9	83.0
Cherry Selection - 30%	61.6	47.0	68.2	71.5	39.8	22.0	26.3	50.8	69.3	88.4
GATEAU-LLaMA - 30%	**64.1**	**50.4**	78.0	73.5	42.0	24.5	**29.5**	46.8	**73.8**	**92.1**
Perplexity Guidance - 50%	62.3	49.6	79.0	71.0	47.3	24.5	28.0	42.0	69.5	86.3
CaR - 50%	61.6	47.9	74.0	77.3	39.0	21.5	24.5	42.0	67.8	91.8
Cherry Selection - 50%	62.9	49.6	77.8	76.2	48.3	22.5	30.5	35.8	68.2	91.5
GATEAU-LLaMA - 50%	**64.1**	**50.4**	**78.0**	73.5	42.0	**24.5**	29.5	**46.8**	**73.8**	**92.1**
Limited Short Instruction Data Settings										
w/o SFT	43.5	25.6	44.5	44.0	35.0	16.5	18.0	28.0	42.0	48.8
w/o Long SFT	56.4	44.5	66.3	65.8	46.5	21.0	23.5	38.3	63.5	79.1
Full - 100%	54.5	40.9	65.8	56.0	35.5	21.0	23.5	34.0	67.5	78.3
Perplexity Guidance - 10%	61.9	39.5	73.8	61.8	39.3	27.5	29.1	47.1	58.5	72.3
CaR - 10%	59.3	40.3	66.5	64.3	49.3	21.5	26.3	28.8	62.0	80.5
Cherry Selection - 10%	53.0	42.3	56.8	72.3	39.5	17.0	26.5	34.8	59.3	75.3
GATEAU-LLaMA - 10%	**62.2**	**44.6**	69.9	67.5	39.8	24.0	27.5	**50.7**	**66.3**	**83.0**
Perplexity Guidance - 30%	58.9	41.4	69.4	68.0	37.0	28.5	28.9	47.8	57.8	64.8
CaR - 30%	52.8	44.3	67.0	66.5	37.3	25.0	24.8	28.5	68.5	71.0
Cherry Selection - 30%	54.8	36.6	67.5	57.5	34.0	19.5	20.4	35.5	63.5	69.7
GATEAU-LLaMA - 30%	**62.0**	43.7	62.0	65.7	**45.4**	27.5	**31.7**	41.7	**71.7**	**72.0**
Perplexity Guidance - 50%	57.6	40.9	59.5	74.5	41.0	25.0	26.0	37.3	55.3	75.3
CaR - 50%	58.3	44.1	70.0	67.2	43.3	25.5	30.5	28.5	71.5	73.5
Cherry Selection - 50%	57.7	41.4	70.0	63.2	37.5	18.3	26.3	43.9	61.1	76.5
GATEAU-LLaMA - 50%	**64.2**	**44.1**	61.5	67.0	**46.3**	**28.0**	**31.4**	**47.0**	65.8	**84.3**

Table 6: Detailed results (%) of MT-Bench.

measure the difficulty brought by the long-range dependencies from two different perspectives, i.e., separately measuring the difficulty of generating corresponding responses and understanding long input contexts due to the long-range dependencies. Meanwhile, we further explore the impact of the length of the segment L in CAM. We report the results of *GATEAU-LLaMA - 50%* on LongBench-Chat in Real-world Settings. As shown in Figure 7, different segment lengths affect the model's performance; however, as long as a reasonable length value is chosen, the fluctuations in model performance are not significant. Meanwhile, the performance will always be improved over using the whole long SFT dataset (namely *Full-100%*) and only using the HMG method (namely *-w/o CAM*), showing the effectiveness of our proposed CAM.

H OOD Case Study

As part of our research on aligning LLMs on long context, we further come up with a practical case study shown in Figure 17. We use an out-of-distribution (OOD) query, which has not been encountered in the long context SFT data. Specifically, we select the *Biden-Trump debate transcript* [1] from the 2024 election season as the OOD query, because this debate is organized subsequent to the collection of our used training datasets. We show the results generated by *GATEAU-LLaMA-30%* and *Full-100%* in Real-world settings. We highlight the sentences that can be easily misunderstood or contain factual errors (e.g., this debate is organized in 2024 instead of 2020). We can find that our method achieves better faithfulness and fluency.

I Discussion

I.1 Discussion about Execution Time and GPU Burdens

Execution Time. Based on the principle of making full use of GPU devices (e.g., using a multi-processing strategy), we list the execution time in Table 9. We can find that GATEAU introduces an acceptable offline time overhead compared to the supervised fine-tuning stage and improves the overall performance of long-context LLMs. Perplexity Guidance applies a single LLM to compute the score, thus, it achieves less execution time but

[1] https://edition.cnn.com/2024/06/27/politics/read-biden-trump-debate-rush-transcript/index.html

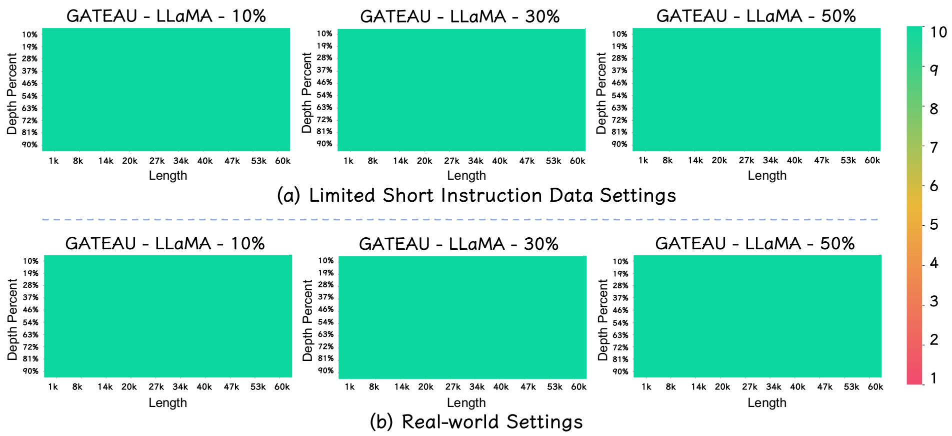

Figure 5: Needle in the Haystack test.

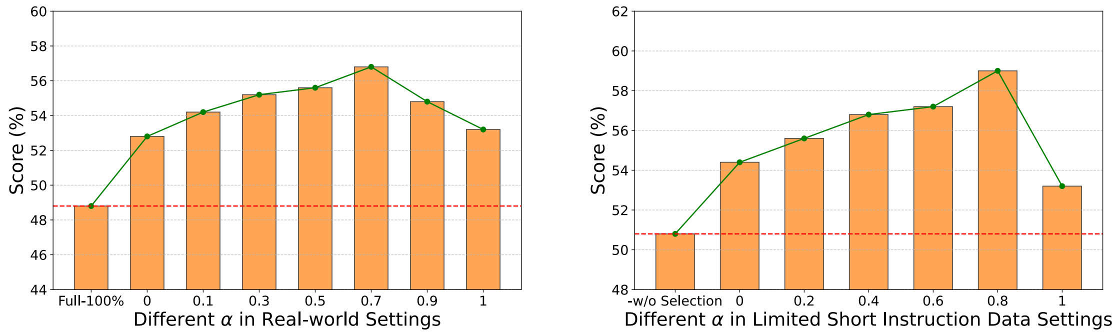

Figure 6: Results (%) on LongBench-Chat with different hyperparameter α in Eq. (6).

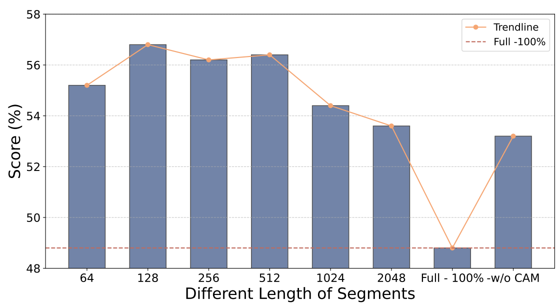

Figure 7: Results (%) on LongBench-Chat with different hyperparameter L in CAM.

worse performance in our experiments. Meanwhile, another strong baseline Cherry Selection introduces an additional training stage and computes the proposed Instruction-Following Difficulty (IFD) by applying the forward propagation twice on a single long SFT data, thus necessitating more execution time compared to our proposed HMG. Meanwhile, our CAM and HMG can process the data in parallel to further decrease the execution time, e.g., only 8 hours with 16xA800 80G GPUs. The experimental results of our proposed GATEAU demonstrate that the additional execution time is worthwhile.

GPU Burdens. GATEAU is designed to score long SFT data and then select the influential samples used for alignment. Thus, our method does not introduce the additional memory burden during the SFT stage and inference stage. For HMG, we compute perplexity scores generated from two models for a given SFT data in parallel and use the computed perplexity scores (cached in JSON files) to get the HMP score as shown in Eq. (2). Thus, HMG does not introduce additional GPU memory burden, only introducing acceptable additional execution time as shown in Table 9. The GPU memory requirements of CAM rise from the calculation of the attention scores for lengthy inputs, as well as the perplexity score computation. This process is equivalent to performing two forward passes over the dataset without updating gradients, thus it does not add an extra GPU memory burden.

I.2 Discussion about Whether GATEAU Can Fit in Other LLMs

We explore whether GATEAU can fit in other LLMs in Table 7. We further apply GATEAU on ChatGLM3-6B-base-8k (Zeng et al., 2023) and

Model	LongBench			LongBench-Chat	MT-Bench		
	Single-Doc QA	Multi-Doc QA	Summarization	Avg	First-turn	Second-turn	Avg
			Real-world Settings				
GATEAU-ChatGLM3 - 6B - 50%	**30.2**	**20.0**	**24.6**	**60.2**	**63.2**	**50.2**	**56.7**
-w/o Data Selection (i.e., Full - 100%)	26.6	16.9	23.4	55.9	59.2	47.6	53.4
GATEAU-LLaMA3 - 8B - 50%	**42.1**	**30.2**	**26.1**	**65.3**	**72.8**	**58.4**	**65.6**
-w/o Data Selection (i.e., Full - 100%)	35.2	24.3	25.6	54.4	67.2	54.2	60.7
			Limited Short Instruction Data Settings				
GATEAU-ChatGLM3 - 6B - 50%	**24.4**	**15.4**	**22.4**	**57.0**	**57.2**	**49.2**	**53.2**
-w/o Data Selection (i.e., Full - 100%)	20.2	13.2	21.2	50.4	55.4	45.2	50.3
GATEAU-LLaMA3 - 8B - 50%	**34.2**	**25.3**	**25.2**	**63.2**	**68.4**	**55.2**	**61.8**
-w/o Data Selection (i.e., Full - 100%)	30.2	23.2	24.7	55.6	62.3	50.2	56.3

Table 7: Results (%) of GATEAU-ChatGLM3 and GATEAU-LLaMA3 series. We show automatic metrics evaluation results on LongBench.

Model	LongBench			LongBench-Chat	MT-Bench		
	Single-Doc QA	Multi-Doc QA	Summarization	Avg	First-turn	Second-turn	Avg
			Real-world Settings				
GATEAU-LLaMA - 7B - 50%	**39.1**	**27.5**	**27.8**	**50.2**	**55.7**	**45.3**	**50.5**
-w/o Data Selection (i.e., Full - 100%)	37.5	24.5	26.9	45.6	52.5	42.1	47.3
			Limited Short Instruction Data Settings				
GATEAU-LLaMA - 7B - 50%	**32.5**	**19.2**	**26.4**	**54.2**	**50.0**	**42.8**	**46.4**
-w/o Data Selection (i.e., Full - 100%)	28.4	17.0	25.5	48.2	47.5	41.4	44.5

Table 8: Experiments to explore whether GATEAU can fit in other long SFT datasets. We use LongAlpaca as the long SFT dataset. We show automatic metrics evaluation results on LongBench.

Methods	Execution Time
Real-world Settings	
Training on the full dataset	176 GPU hours
Selecting long SFT data via HMG	64 GPU hours
Selecting long SFT data via CAM	48 GPU hours
Selecting long SFT data via Cherry Selection	80 GPU hours
Selecting long SFT data via Perplexity Guidance	32 GPU hours

Table 9: Execution time.

ChatGLM3-6B-base-64k (Bai et al., 2024), then fine-tune ChatGLM3-6B-base-64k using the selected samples. We also conduct the experiments on LLaMA3-8B-base-8k and LLaMA3-8B-base-64k, then fine-tune LLaMA3-8B-base-64k using the selected samples. We can find consistent improvements on three benchmarks compared to using the full long SFT dataset. This indicates that GATEAU effectively fits in other LLMs.

I.3 Discussion about Whether GATEAU Can Fit in Other Long SFT Datasets

Meanwhile, we explore whether GATEAU can fit in other long SFT datasets. Specifically, we implement our proposed GATEAU on the long SFT dataset LongAlpaca (Chen et al., 2024b), which contains 9,000 long SFT samples. As shown in Table 8, we can find that our method GATEAU achieves consistent improvements on three benchmarks, including long-context understanding benchmark and two instruction-following benchmarks, showing the GATEAU can generalize across different long SFT datasets.

I.4 Discussion about the Diversity of Selected Samples

In this section, we further explore the diversity of selected samples. We employ the cluster model as CaR (Ge et al., 2024) to cluster all candidate instruction pairs into k clusters. Specifically, we employ the k-Means algorithm and a sentence transformer model, which is used to map sentences to a 384-dimensional dense vector space. Subsequently, semantic features are PCA-reduced to retain 95% of dimensions. Finally, by setting the number of clusters as $k = \sqrt{n/2}$ for n long SFT samples, all 10k long SFT samples are clustered into 70 clusters. Finally, all samples are sorted based on their scores according to Eq. (6), and the top n_1 samples are selected. Within each cluster, samples are sorted by score from GATEAU, and the top n_2 pairs are chosen. We set n_2 to 1, which is the same as Ge et al. (2024). Finally, we can get $n_1 + k * n_2$ (i.e., $4300{+}70{*}1$) samples and use these selected data to train the model, namely *-w Diversity-preserved Selection*. We report the results of GATEAU-LLaMA - 50% on LongBench-Chat and MT-Bench. Shown in Table 11, we find that using *Diversity-preserved Selection* does not consistently improve the performance, showing our proposed GATEAU has implicitly ensured the diversity of selected long SFT

Model	LongBench			LongBench-Chat	MT-Bench		
	Single-Doc QA	Multi-Doc QA	Summarization	Avg	First-turn	Second-turn	Avg
Real-world Settings							
GATEAU-LLaMA - 7B - 50%	**38.9**	**25.8**	25.5	**56.8**	**64.1**	**50.4**	**57.3**
-w/o Extended Context Windows	38.1	25.4	**25.6**	55.8	63.7	50.6	57.1
-w/o Norm in Eq. (2)	37.5	24.1	25.3	56.2	64.1	50.4	57.3
Homologous Model's Guidance	38.4	24.3	25.1	53.2	61.7	51.5	56.6
Perplexity Guidance	37.9	23.4	25.4	49.8	62.3	49.6	55.9
Non-Homologous Model's Guidance	37.2	23.2	24.8	48.2	59.2	49.3	54.3
Limited Short Instruction Data Settings							
GATEAU-LLaMA - 7B - 50%	**31.0**	18.1	**25.3**	**59.0**	**64.2**	**44.1**	**54.2**
-w/o Extended Context Windows	29.2	**18.8**	25.2	57.6	60.2	44.0	52.1
-w/o Norm in Eq. (2)	29.7	18.7	24.9	55.2	62.0	40.1	51.1
Homologous Model's Guidance	28.5	17.5	24.7	53.2	61.3	42.4	51.8
Perplexity Guidance	28.3	16.8	24.7	51.0	57.6	40.9	49.2
Non-Homologous Model's Guidance	28.7	16.8	24.8	50.2	60.1	40.3	50.2

Table 10: Discussion about Homologous Model's Guidance.

Models	LongBench-Chat	MT-Bench
Real-world Settings		
GATEAU-LLaMA - 7B - 50%	**56.8**	57.3
-w/ Diversity-preserved Selection	56.2	**57.8**
Limited Short Instruction Data Settings		
GATEAU-LLaMA - 7B - 50%	59.0	**54.2**
-w/ Diversity-preserved Selection	**59.2**	53.4

Table 11: Experiments to explore the diversity of selected samples by GATEAU.

data. This is because HMG and CAM separately measure the difficulty of generating corresponding responses and understanding long input contexts due to the long-range dependencies, thus the final score derived from two different perspectives inherently ensures the diversity of selected long SFT data. Meanwhile, as shown in Table 6, GATEAU achieves better overall performance and more balanced performance in 8 different tasks, showing its effectiveness and diversity of selected samples.

I.5 Discussion about Homologous Model's Guidance

We further explore some key questions in the Homologous Model's Guidance.

Why Do We Need Homologous Models? Homologous Model's Guidance (HMG) aims to assess the degree of long-range dependencies required for the corresponding response generation by comparing the perplexity scores of the response between two homologous models with different context windows. The idea behind HMG is that the primary difference between homologous models with varying context windows lies in their different capabilities for modeling long-range dependencies instead of other capabilities. Thus, the disparity in the perplexity scores can be interpreted as reflecting the difference in the long-range dependencies modeling capabilities required to generate the given response. To evaluate the effectiveness of our idea, we replace *LLaMA-2-7B-base-4k* with *Qwen-2-7b-base-8k* (Yang et al., 2024) as model θ_A in Eq. (2), namely *Non-Homologous Model's Guidance*. As shown in Table 10, we find *Non-Homologous Model's Guidance* achieves worse performance than *Homologous Model's Guidance* in two designed settings. It shows that HMG can exclusively measure the richness of long-range dependency relations in long SFT samples. As non-homologous models have different pre-training phases and model architectures, the modified Eq. (2) can not effectively measure the degree of long-range dependencies required for response generation and introduce the influence brought by other different capabilities of non-homologous models, resulting in worse performance.

Why Do We Apply Normalization in Eq. (2) ? We apply softmax normalization to each score in Eq. (2) to determine its respective ranking among the datasets for two perplexity scores. This is because our early experiments observed that applying softmax normalization can further improve the performance shown in Table 10. This may due to the fact that some extremely noisy samples tend to have large perplexity scores, which in turn lead to unstable HMP scores if we do not apply normalization in Eq. (2). Training LLMs on these noisy samples further leads to poor results.

What Will Happen if We Do not Extend the Context Windows of LLaMA-2-4k? Our early experiments also explore what will happen if we do not extend the context windows of model θ_A in Eq. (2). As shown in Table 10, we are sur-

Models	Perplexity
LLaMA-2-7B-base-4k	3.72
LLaMA-2-7B-base-64k	2.61

Table 12: Perplexity from base models on LongAlign.

Models	LongBench-Chat	MT-Bench
Real-world Settings		
GATEAU-LLaMA - 7B - 50%	**56.8**	57.3
w/ ICL Alignment	56.2	**57.9**
Limited Short Instruction Data Settings		
GATEAU-LLaMA - 7B - 50%	59.0	**54.2**
w/ ICL Alignment	**59.4**	53.5

Table 13: Experiments to explore whether alignment via in-context learning helps HMG.

Methods	Data Overlap Ratio with GATEAU
Cherry LLM	12%
CaR	5%

Table 14: Data overlap of top 10% of long SFT data selected by baselines and our proposed GATEAU.

prised to find that *-w/o Extended Context Windows* also achieves competitive results in three benchmarks compared to *GATEAU-LLaMA*. Even the perplexity score $\mathrm{PPL}_{\theta_A}(y|c,x)$ from the model θ_A can be very large, e.g., the value of $\mathrm{PPL}_{\theta_A}(y|c,x)$ can be larger than 1000, the value after softmax normalization is still useful and applicable in the Homologous Models' Guidance. This interesting finding can be used to reduce the complexity of applying Homologous Models' Guidance and achieve competitive performance.

Is the Perplexity Score from the Base Model Really so High that It Can Not Accurately Measure the Difficulty? As the base model performs well on conditional generation tasks like continuation, it should be able to generate accurate perplexity scores on the response of instruction-following data, even though the model might not be able to produce high-quality responses correctly, because these two capabilities are not the same. We explore whether our long-context LLM would produce incorrect perplexity values in Table 12. We calculate the average perplexity value generated by LLaMA-2-7B-base-64k for the entire long SFT dataset LongAlign during the whole HMG process, which is 2.61. We further calculate the average perplexity value generated by LLaMA-2-7B-base-4k for the entire long SFT dataset LongAlign during the whole HMG process, which is 3.72. This is because we expand the base frequency of the RoPE position encoding by 200 times (from 10,000 to 2,000,000) to extend the context windows and avoid the model conducting extreme perplexity score (e.g., >1,000) in HMG. Thus, there is no issue of the perplexity from the base model being too high to accurately measure the difficulty.

Can the Perplexity Score Generated from the Base Model be Used as Guidance to Select Influential Samples? The perplexity of the responses computed with the base model is an intuitive metric, as it measures the difficulty of the data sample during the generation. As shown in Table 1, Table 2, Table 5, and Table 8, we find simply using high perplexity (namely Perplexity Guidance in our paper) can also improve the performance compared with using the whole long SFT dataset, indicating that the effectiveness of the perplexity score from the base model in selecting long SFT samples. Cherry Selection (Li et al., 2024b) also finds using the Instruction-Following Difficulty (a variant of perplexity score) computed with the base model also works in selecting SFT samples. According to these experiments, we believe that the perplexity generated from a base model can be used as positive guidance to select SFT samples. Therefore, the use of the perplexity score generated from the base model in our method makes sense when selecting long SFT data. Meanwhile, our method HMG is designed to minimize the influence of other factors (e.g., the limited instruction-following ability of a base model) and model the difficulty in modeling the long-range dependencies to construct the more effective guidance of long SFT data selection, and further improve overall performance. For CAM, utilizing perplexity scores to compute importance scores is also reasonable, and the experiments show improvement even when only using CAM.

We further conduct additional experiments to explore the effect of perplexity scores generated from the base model. In HMG, we use in-context learning technology to align the base model and use the perplexity score from the aligned model to select long SFT data. Specifically, we use the same 3 demonstration examples as URIAL (Lin et al., 2024). In this way, we can get models more aligned without updating the parameters. However, as shown in Table 13, using the aligned model via in-context learning does not consistently improve the final performance. This indicates that

Model	LongBench			LongBench-Chat	MT-Bench		
	Single-Doc QA	Multi-Doc QA	Summarization	Avg	First-turn	Second-turn	Avg
Real-world Settings							
GATEAU-LLaMA - 7B - 10%	**40.5**	**24.0**	**25.4**	**55.4**	**63.1**	**54.1**	**58.6**
-w/o Mixed-Quality Data	39.1	22.4	24.6	53.6	61.8	52.3	57.1
Limited Short Instruction Data Settings							
GATEAU-LLaMA - 7B - 10%	**29.9**	**18.4**	**24.6**	**58.0**	**62.2**	**44.6**	**53.4**
-w/o Mixed-Quality Data	28.7	17.5	24.2	55.6	61.2	42.3	51.8

Table 15: Discussion about training models on mixed-quality data.

using only base models in the HMG phase can also achieve good results. Therefore, HMG can effectively minimize the influence of other factors (e.g., the limited instruction-following ability of a base model) and model the difficulty in modeling the long-range dependencies. Meanwhile, from the real-world implementation viewpoint, directly using the base model is more efficient and at the same time effective as well.

I.6 Discussion about Orthogonality with Baselines

To provide additional evidence of the unique benefits of our approach, we conduct additional experiments to analyze the orthogonality between GATEAU and various baselines. Specifically, we calculate the overlap of the top 10% of long SFT data selected by other baselines and our method in Limited Short Instruction Data Settings. As shown in the Table 14, we can find a significant difference between the samples selected by the baselines focused on short SFT data selection (i.e., Cherry LLM and CaR) and those selected by our proposed GATEAU. This is because GATEAU is designed to identify the influential long samples enriched with long-range dependency relations to achieve better long context alignment instead of focusing on selecting short SFT data. Thus, GATEAU grasps important patterns that differ from the existing baselines. Furthermore, how to utilize such orthogonality to improve the final performance remains a promising research topic. We attempt to explore how to utilize it to further improve the final performance in our future work.

I.7 Discussion about Training Models on Mixed-Quality Data

We further attempt to explore whether there exists an optimal balance between low-scoring and high-scoring long SFT samples that enables the long-context LLM to perform even better than using the samples with high scores. Specifically, we use long SFT samples from the top 5% and bottom 5% to form the training samples, namely Mixed-Quality Data. As shown in Table 15, we find that this strategy does not improve the final performance of the LLMs This indicates that our strategy of using top-ranked samples is effective.

J Case Study for Characteristics of Selected Samples

We conduct a case study to show the difference between samples with high or low scores generated by GATEAU. In Figure 18, we show the sample with the highest score and the sample with the lowest score in Real-world Settings. We highlight the low-quality sentences. We can find that the sample with the highest score shows better faithfulness and fluency, showing the effectiveness of our method in selecting high-quality influential samples for long context alignment.

LongBench-Chat Evaluation Prompt

[Instructions] You are asked to evaluate the quality of the AI assistant's answers to user questions as an impartial judge, and your evaluation should take into account factors including correctness (high priority), helpfulness, accuracy, and relevance. The scoring principles are as follows:
1. Read the AI assistant's answer and compare the assistant's answer with the reference answer.
2. Identify all errors in the AI Assistant's answers and consider how much they affect the answer to the question.
3. Evaluate how helpful the AI assistant's answers are in directly answering the user's questions and providing the information the user needs.
4. Examine any additional information in the AI assistant's answer to ensure that it is correct and closely related to the question. If this information is incorrect or not relevant to the question, points should be deducted from the overall score.
Please give an overall integer rating from 1 to 10 based on the above principles, strictly in the following format: "[[rating]]", e.g., "[[5]]".

[Question] {}
[Reference answer begins] {} [Reference answer ends]
Below are several assistants' answers and their ratings:
[Assistant's answer begins] {} [Assistant's answer ends] Rating: [[{}]]
[Assistant's answer begins] {} [Assistant's answer ends] Rating: [[{}]]
[Assistant's answer begins] {} [Assistant's answer ends] Rating: [[{}]]

Please rate the following assistant answers based on the scoring principles and examples above:
[Assistant's answer begins] {} [Assistant's answer ends]
Rating:

Figure 8: LongBench-Chat evaluation prompt.

LongBench Evaluation Prompt for QA tasks

You are asked to evaluate the quality of the AI assistant's answers to user questions as an impartial judge, and your evaluation should take into account factors including correctness (high priority), and comprehensiveness (whether the assistant's answer covers all points). Read the AI assistant's answer and compare it against the reference answer, and give an overall integer rating of 1, 2, or 3 (1 = wrong or irrelevant, 2 = partially correct, 3 = correct and comprehensive) based on the above principles, strictly in the following format: "[[rating]]", e.g., "[[2]]".

Question: {*Question*}
Reference answer: {*Groundtruth*}
Assistant's answer: {*Response*}
Rating:

Figure 9: LongBench evaluation prompt for QA tasks.

LongBench Evaluation Prompt for Summarization Tasks

You are asked to evaluate the quality of the AI assistant's generated summary as an impartial judge, and your evaluation should take into account factors including correctness (high priority), comprehensiveness (whether the assistant's summary covers all points), and coherence. Read the AI assistant's summary and compare it against the reference summary, and give an overall integer rating on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the evaluation criteria, strictly in the following format: "[[rating]]", e.g., "[[3]]".

Reference summary: {*Groundtruth*}
Assistant's summary: {*Response*}
Rating:

Figure 10: LongBench evaluation prompt for summarization tasks.

Evaluation Prompt for the Coherence of Long Input Contexts

You are asked to evaluate the Long Input Contexts as an impartial judge, and your evaluation should follow these scoring principles:
1. Read the given Long Input Contexts carefully.
2. Evaluate the fluency and coherence of Long Input Contexts.
3. Evaluate whether the Long Input Contexts are focused and relevant.

Please give an overall integer rating from 1 to 5 based on the above principles, strictly in the following format:"[[rating]]", e.g. "[[5]]".

Please rate the following Long Input Contexts based on the scoring principles:

[Long Input Contexts begins]
{*Long Input Contexts*}
[Long Input Contexts ends]

Rating:

Figure 11: Evaluation prompt for the coherence of long input contexts.

Evaluation Prompt for the Necessity of Long Input Contexts

You are asked to evaluate the Long Input Contexts as an impartial judge, and your evaluation should follow these scoring principles:
1. Read the given Instruction, Long Input Contexts and Assistant's answer carefully.
2. Evaluate how difficult it is for the Assistant to follow the given Instruction without the given Long Input Contexts.
3. Evaluate how necessary the given Long Input Contexts are to get the Assistant's answer. If the Long Input Contexts are meaningless or irrelevant, points should be deducted from the overall score.

Please give an overall integer rating from 1 to 5 based on the above principles, strictly in the following format: "[[rating]]", e.g., "[[5]]".

[Instruction begins]
{*Instruction*}
[Instruction ends]

[Long Input Contexts begins]
{*Long Input Contexts*}
[Long Input Contexts ends]

Please rate the following assistant answers based on the scoring principles:

[Assistant's answer begins]
{*Assistant's answer*}
[Assistant's answer ends]

Rating:

Figure 12: Evaluation prompt for necessity of long input contexts.

Evaluation Prompt for the Faithfulness of Response

You are asked to evaluate the AI assistant's answers to user questions as an impartial judge, and your evaluation should follow these scoring principles:
1. Read the given Instruction, Long Input Contexts, and Assistant's answer carefully.
2. Identify all errors in the AI Assistant's answers and consider how much they affect the answer to the question.
3. Evaluate how faithful the AI assistant's answers are to follow the Instruction, i.e., how correct and closely related to the Instruction.
4. Evaluate how faithful the AI assistant's answers are based on the Long Input Contexts, i.e., how correct and closely related to the Long Input Contexts.

Please give an overall integer rating from 1 to 5 based on the above principles, strictly in the following format: "[[rating]]", e.g., "[[5]]".

[Instruction begins]
{*Instruction*}
[Instruction ends]

[Long Input Contexts begins]
{*Long Input Contexts*}
[Long Input Contexts ends]

Please rate the following assistant answers based on the scoring principles:

[Assistant's answer begins]
{*Assistant's answer*}
[Assistant's answer ends]

Rating:

Figure 13: Evaluation prompt for faithfulness of response.

Evaluation Prompt for the Helpfulness of Response

You are asked to evaluate the AI assistant's answers to user questions as an impartial judge, and your evaluation should follow these scoring principles:
1. Read the given Instruction and Assistant's answer carefully.
2. Identify all errors in the AI Assistant's answers and consider how much they affect the answer to the question.
3. Evaluate how helpful the AI assistant's answers are in directly answering the user's questions and providing the information the user needs.

Please give an overall integer rating from 1 to 5 based on the above principles, strictly in the following format: "[[rating]]", e.g. "[[5]]".

[Instruction begins]
{*Instruction*}
[Instruction ends]

Please rate the following assistant answers based on the scoring principles:

[Assistant's answer begins]
{*Assistant's answer*}
[Assistant's answer ends]

Rating:

Figure 14: Evaluation prompt for helpfulness of response.

Evaluation Prompt for the Complexity of Instruction

You are asked to evaluate the Instruction as an impartial judge, and your evaluation should follow these scoring principles:
1. Read the given Instruction carefully.
2. Evaluate the scope of the Instruction, i.e., whether the Instruction encompasses information necessary for successful completion.
3. Evaluate the depth of the Instruction, i.e., whether the Instruction provides thorough details and nuances.
4. Evaluate whether the Instruction integrates multiple steps or concepts that require careful attention and understanding.
5. If the Instruction is too easy to follow, points should be deducted from the overall score.
Please give an overall integer rating from 1 to 5 based on the above principles, strictly in the following format: "[[rating]]", e.g., "[[5]]".

Please rate the following Instruction based on the scoring principles and examples above:

[Instruction begins]
{*Instruction*}
[Instruction ends]

Rating:

Figure 15: Evaluation prompt for complexity of instruction.

The Principles of Human Evaluation

You are asked to evaluate the responses generated by different models, and your evaluation should follow these principles:
1. **Correctness**: Focus primarily on whether the response covers the key points in the reference answer. For reference answers containing multiple key points, look for how many of these the response accurately addresses and score accordingly.
2. **Helpfulness**: Evaluate how helpful the AI assistant's answers are in directly answering the user's questions and providing the information the user needs.
3. **Faithfulness**: Examine any additional information in the AI assistant's answer to ensure that it is correct and closely related to the question. If this information is incorrect or not relevant to the question, points should be deducted from the overall score.
4. **Coherence**: Evaluate how fluent and coherent the AI assistant's answers are. Also, consider deducting points for overly verbose responses or those that are excessively generalized.

Finally, please make a decision among the 3 opinions, including Win, Tie, and Loss.

Figure 16: The principles of human evaluation.

OOD Case Study

User:
[Debate transcript begins] {}
[Debate transcript ends]
[Instruction begins]
As a professional summarizer, create a concise and comprehensive summary of the provided debate, while adhering to these guidelines:
1. Craft a summary that is detailed, in-depth, and complex, while maintaining clarity and conciseness.
2. Use proper grammar, punctuation, and spelling.
3. Write in a formal, objective, and impartial tone.
4. Provide a clear and accurate summary of the debate, without bias or personal opinions.
[Instruction ends]
Summary:

Ours: The CNN presidential debate between President Joe Biden and former President Donald Trump was held in Atlanta, Georgia. The debate was moderated by CNN anchors Jake Tapper and Dana Bash. The debate covered various topics including the economy, immigration, foreign policy, democracy, and more. On the economy, President Biden defended his administration's actions, stating that he inherited a disastrous situation from the previous administration. He highlighted his administration's job creation efforts and infrastructure spending. Former President Trump criticized Biden's policies, claiming that Biden's tax hikes and regulations were hurting the economy. On immigration, President Biden defended his administration's handling of the border crisis, stating that he had increased border patrol and asylum officers. Former President Trump criticized Biden's immigration policies, claiming that Biden was allowing "terrorists" to enter the country. On foreign policy, President Biden criticized former President Trump's handling of foreign affairs, particularly in regards to Russia and Ukraine. He stated that Trump had encouraged Putin to invade Ukraine and that Trump had not done enough to stop the war. Former President Trump defended his actions, stating that he had put pressure on Ukraine to investigate Biden's son and that he had not done enough to stop the war because he was focused on China. On democracy, President Biden criticized former President Trump's handling of the January 6th attack on the Capitol, stating that Trump had encouraged the protesters to go to the Capitol. Former President Trump defended his actions, stating that he had spoken peacefully and patriotically. Overall, the debate was contentious and both candidates strongly defended their records and policies.

Full - 100%: The debate between President Joe Biden and former President Donald Trump was held in Atlanta, Georgia, on October 22, 2020. The debate was moderated by CNN anchors Jake Tapper and Dana Bash. The debate focused on various issues such as the economy, immigration, foreign policy, democracy, and more. On the issue of the economy, President Biden argued that he inherited a disastrous economy from Trump, with high unemployment rates and no jobs. He claimed that his policies have helped create jobs and increase wages for working-class Americans. Trump, on the other hand, claimed that he had the best economy in history and that Biden's policies have caused inflation and driven up prices. On the issue of immigration, Trump claimed that Biden was allowing "millions of people to come in and destroy our country." He also claimed that Biden was responsible for the worst border crisis in history. Biden argued that Trump had left the border in chaos and that his policies had caused the crisis. On foreign policy, Trump claimed that he had brought the United States to the brink of war with Iran and that he had made a deal with North Korea that was a "disaster." Biden argued that Trump had put the United States in a dangerous position with his rhetoric and that he had made the world a more dangerous place. On democracy, Trump claimed that the election was rigged and that he had won the election. Biden argued that Trump had incited a violent insurrection at the Capitol and that he had undermined democracy. In conclusion, the debate was contentious and both candidates presented their respective arguments and policies. However, it remains to be seen how the voters will interpret the debate and how it will impact the election outcome.

Figure 17: OOD case study.

Case Study for Characteristics of Selected Samples

😄 Sample with the highest score:

[Long Input Contexts begins]
{Long Input Contexts}
[Long Input Contexts ends]

Instruction: Since what year has SAS been giving customers around the world THE POWER TO KNOW®?

Response: Based on the information in the doc, SAS has been giving customers around the world THE POWER TO KNOW® since 1976. Specifically, the last paragraph states: "SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®." Therefore, the answer is that SAS has been giving customers THE POWER TO KNOW® since 1976.

☹ Sample with the lowest score:

[Long Input Contexts begins]
{Long Input Contexts}
[Long Input Contexts ends]

Instruction: Bookertainment服务的主要目标市场是哪里?该服务是否会扩大到其他国家?

Response: 根据文档的内容, Bookertainment 服务的主要目标市场是芬兰。文档提到芬兰目前有约550万居民,并且娱乐行业的就业人数也在增长,这表明Bookertainment服务在芬兰市场存在机会。但是文档也提到, Bookertainment 的最初目标市场是芬兰,如果业务 successful,可以逐步扩展到斯堪的纳维亚其他国家。所以服务目前只针对芬兰市场,但未来有可能扩大到斯堪的纳维亚地区。文档没有提及会扩展到其它非斯堪的纳维亚国家。

Figure 18: Case study for characteristics of selected samples.