Beyond Binary Preferences: Semi-Online Label-Free GRACE-KTO with Group-Wise Adaptive Calibration for High-Quality Long-Text Generation

Jingyang Deng[1, *], **Ran Chen**[1, *], **Jo-Ku Cheng**[1], **Jinwen Ma**[1,2, **]
[1]School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China
[2]Institute of Systems Science, Beijing Wuzi University, Beijing 101149, China

Abstract

Generating high-quality long-text remains challenging for Large Language Models (LLMs), as conventional supervised fine-tuning fails to ensure overall quality due to its teacher-forcing nature. Kahneman-Tversky Optimization (KTO), as a model alignment method that can holistically optimize generation quality, overcomes the need for paired preference data required by previous methods. However, it still suffers from binary supervision that inadequately reflects varying quality degrees. To address this, we propose GRACE-KTO, a semi-online framework that transforms KTO's binary signals into dynamically calibrated intra-group rewards. Specifically, GRACE-KTO aggregates responses to identical queries into groups, computes rank-sum scores across multiple linguistic quality dimensions, and applies group-wise and global normalization to adaptively redistribute sample importance. We adopt a semi-online training strategy to reduce costly online sampling while outperforming offline variants. By leveraging query generation with seed data, we minimize labeled data dependency, using the model's own knowledge to enhance its long-text generation capabilities. Additionally, we extend the context window to 32k tokens using YaRN during inference, enabling the model to generate longer texts while maintaining perplexities. Experiments demonstrate GRACE-KTO's superiority over vanilla KTO on both automatic metrics and LLM-as-a-Judge evaluations, advancing long-text generation through group-wise adaptive calibration.

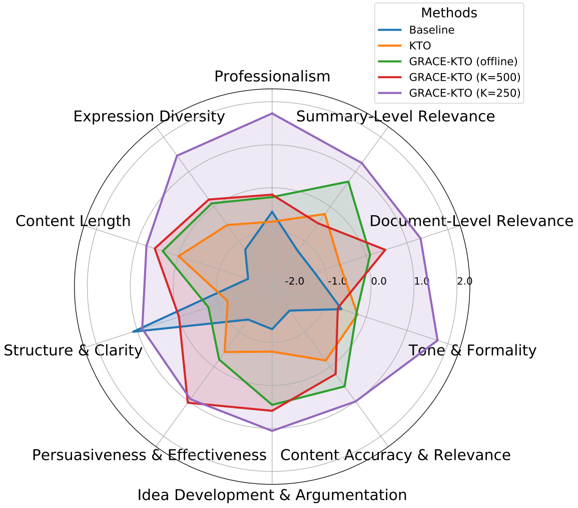

Figure 1: Z-score Normalized Metrics of Different Methods.

1 Introduction

Ensuring high-quality long-text generation remains a formidable challenge for Large Language Models (LLMs). While long-context LLMs have made remarkable progress in understanding lengthy texts with context lengths reaching 1M tokens or more (GLM et al., 2024; Yang et al., 2025), long-text generation presents a distinct and more complex task. This task demands that models produce content significantly longer than the input text. Even powerful open-source models like Qwen2.5-72B-Instruct (Yang et al., 2024), despite their impressive capabilities, face limitations and can only generate up to 8k tokens. This disparity highlights the ongoing need for advancements in training methodologies to overcome the inherent challenges of long-text generation.

Conventional supervised fine-tuning (SFT) for long-text generation relies on teacher-forcing (Williams and Zipser, 1989) to imitate training sequences stepwise, but has limited capacity to enhance holistic text quality. This arises from exposure bias (Li et al., 2024): models trained on ground-truth contexts are incapable of handling errors that accumulate during autoregressive generation. In long-text generation, early inaccuracies trigger cascading errors through subsequent to-

[*] These authors contributed equally to this work.
[**] Corresponding author, Email:jwma@math.pku.edu.cn.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17550–17562
November 4-9, 2025 ©2025 Association for Computational Linguistics

kens, worsening the training-inference gap between teacher-forced optimization and free-generation execution. This gap ultimately restricts SFT's ability to improve the overall quality of the responses.

To address this limitation, various alignment methods have been proposed to optimize the generation quality from a more comprehensive perspective. Among them, Kahneman-Tversky Optimization (Ethayarajh et al., 2024) (KTO) has shown promising results. Unlike previous methods like Proximal Policy Optimization (Ouyang et al., 2022) (PPO) and Direct Preference Optimization (Rafailov et al., 2023) (DPO) that require paired preference data, KTO can work with unpaired data, offering greater flexibility. KTO is based on Kahneman-Tversky's prospect theory (Tversky and Kahneman, 1992), which models human utility in a way that captures biases such as loss aversion. However, KTO still employs binary supervision signals, which are insufficient to capture the varying degrees of quality among different generated texts. This limitation hinders KTO's ability to fully utilize the available data and accurately guide the model towards generating higher-quality long-text.

In this work, we propose **G**roup-**R**eward **A**daptive **C**alibration for **E**nhanced KTO (GRACE-KTO), a semi-online framework aimed at enhancing long-text generation quality by addressing the limitations of KTO. GRACE-KTO refines KTO's binary signals into dynamic intra-group rewards by harnessing KTO's flexibility with unpaired data. The process begins by grouping responses to identical queries and calculating rank-sum scores across key linguistic dimensions, including content length, expression diversity, professionalism, and relevance to the query. These scores are then normalized within each group to provide a comprehensive reflection of each response's quality. Moreover, global normalization is applied across all samples to adaptively adjust the importance of each sample, thereby determining their respective weights in the KTO training process.

To improve both efficiency and effectiveness, we adopt a semi-online training strategy. Unlike fully online algorithms that require frequent real-time model updates for sampling small batches of data—a time-consuming process particularly for large models like 72B models—our semi-online approach efficiently harnesses the concurrent processing power of frameworks like vLLM (Kwon et al., 2023) to significantly enhance sampling efficiency. Furthermore, this strategy proves to be more effective than offline training methods. To reduce the reliance on labeled data, we utilize the training set as seed data and prompt the LLM to generate new queries for training. This allows the model to improve the quality of its responses by leveraging its inherent knowledge rather than depending solely on labeled data. Additionally, by extending the context window to 32k tokens using YaRN (Peng et al., 2024) interpolation during inference, the model can generate longer texts with low perplexities.

Through experiments, we show that GRACE-KTO surpasses vanilla KTO in both automatic metrics and LLM-as-a-Judge (Zheng et al., 2023) evaluations. By dynamically calibrating rewards across groups, GRACE-KTO enables the model to better learn from varying quality degrees in generated texts, resulting in more coherent and contextually consistent long-text outputs. In summary, our work makes the following key contributions:

- We propose GRACE-KTO, a novel semi-online framework that enhances KTO by transforming its binary signals into dynamic intra-group rewards. This allows for more nuanced optimization of text generation quality by effectively capturing varying quality degrees in the generated texts.
- We develop a semi-online training strategy that improves upon the time inefficiency of online sampling. This approach also minimizes reliance on labeled data by treating the training set as seed data and prompting the LLM to generate new queries for the alignment dataset. Thus, the model leverages its inherent knowledge rather than external annotations, enhancing its long-text generation capabilities in a more efficient and self-sustaining manner.
- We extend the model's context window to 32k tokens using YaRN, enabling the generation of longer texts while maintaining perplexity.

2 Related Work

2.1 Challenges in Long-Text Generation

Long-text generation poses significant challenges for LLMs, as it requires coherent and contextually consistent outputs across extended sequences. Various benchmarks have been developed to assess this capability. LongLaMP (Kumar et al., 2024)

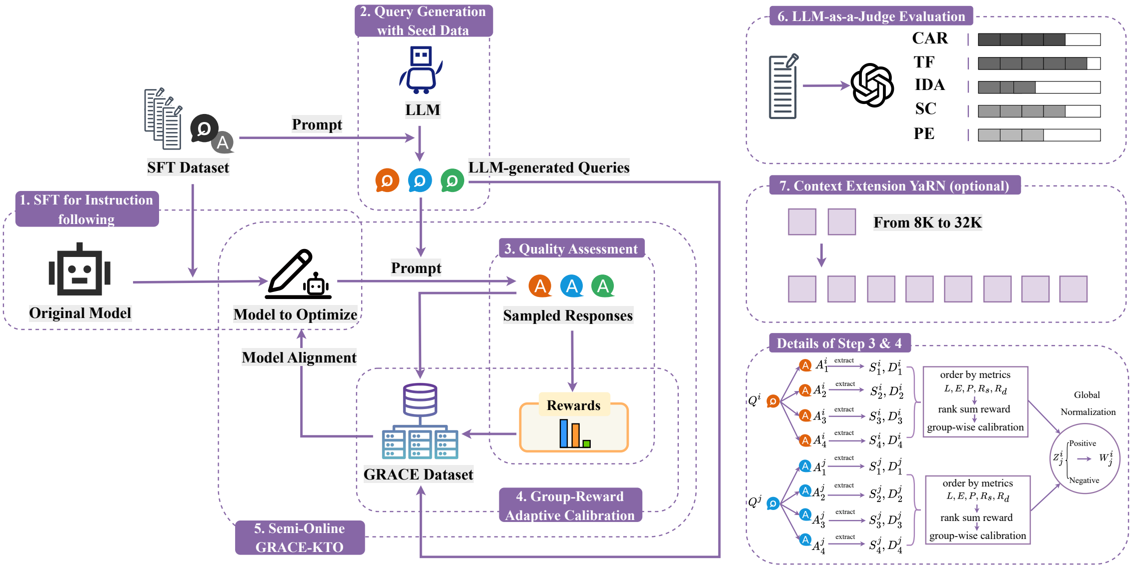
2. Query Generation with Seed Data
LLM
SFT Dataset
Prompt
LLM-generated Queries
1. SFT for Instruction following
Original Model
Model to Optimize
Prompt
3. Quality Assessment
Sampled Responses
Model Alignment
Rewards
GRACE Dataset
4. Group-Reward Adaptive Calibration
5. Semi-Online GRACE-KTO
6. LLM-as-a-Judge Evaluation
CAR
TF
IDA
SC
PE
7. Context Extension YaRN (optional)
From 8K to 32K
Details of Step 3 & 4
Q^i
A_1^i extract S_1^i, D_1^i
A_2^i extract S_2^i, D_2^i
A_3^i extract S_3^i, D_3^i
A_4^i extract S_4^i, D_4^i
order by metrics L, E, P, R_s, R_d
rank sum reward
group-wise calibration
Q^j
A_1^j extract S_1^j, D_1^j
A_2^j extract S_2^j, D_2^j
A_3^j extract S_3^j, D_3^j
A_4^j extract S_4^j, D_4^j
order by metrics L, E, P, R_s, R_d
rank sum reward
group-wise calibration
Global Normalization
Z_j^i Positive Negative W_j^i

Figure 2: Overview of GRACE-KTO.

provides a benchmark for personalized long-text generation but is limited to shorter output length. HelloBench (Que et al., 2024) evaluates long-text generation across multiple tasks and shows that most LLMs struggle to produce texts longer than 4000 words without quality loss. LongGenBench (Wu et al., 2025) further evaluates models' ability to generate long texts while following complex instructions, demonstrating the challenges faced by even state-of-the-art LLMs as text length increases. These benchmarks collectively highlight the gap between current LLM capabilities and the requirements of real-world applications that demand high-quality long-text generation.

2.2 Comparison with Existing Approaches

Approaches to enhancing long-text generation can be categorized into two groups: methods leveraging external agents or tools, and those focusing on model intrinsic training. AgentWrite (Bai et al., 2025) uses an agent-based pipeline to break down ultra-long generation tasks into manageable subtasks. Similarly, RAL-Writer (Zhang et al., 2025) employs retrieval-augmented generation to mitigate the "lost-in-the-middle" issue. These approaches depend on external mechanisms rather than enhancing the model's inherent capabilities. In contrast, our work focuses on improving the model's intrinsic ability to generate high-quality long texts.

Methods focusing on model training itself have shown promise. For instance, Self-Lengthen (Quan et al., 2024) presents an iterative training framework that expands responses through iterative SFT without auxiliary data, using two models to iteratively produce longer responses. However, this approach still relies on SFT, which may suffer from teacher-forcing issues in long-text generation. LongWriter (Bai et al., 2025) demonstrates that incorporating extended-output datasets into model alignment can unlock longer generation capabilities, and finds that DPO training outperforms SFT when using the AgentWrite-constructed dataset. Suri (Pham et al., 2024) explores multi-constraint instruction following for long-text generation, proposing the I-ORPO algorithm, which still requires paired preference data.

Our proposed GRACE-KTO stands out with a semi-online framework that transforms KTO's binary signals into dynamically calibrated intra-group rewards, removing the need for paired preference data that PPO (Ouyang et al., 2022), DPO (Rafailov et al., 2023), and ORPO (Hong et al., 2024) depend on. By aggregating responses to identical queries and adjusting sample importance via group-wise and global normalization, GRACE-KTO offers a more nuanced approach that better reflects varying quality degrees. This makes it more efficient and effective for long-text generation.

3 Methodology

In this section, we introduce GRACE-KTO, our proposed framework for enhancing the quality of long-text generation. As depicted in Figure 2, GRACE-KTO refines the conventional binary feedback of KTO through group-wise adaptive calibration. This process enriches the training information by capturing diverse quality degrees. Our semi-online training approach not only improves sampling efficiency but also outperforms purely offline training. By using the training set as seed data and employing query generation to construct an alignment training dataset, GRACE-KTO leverages the model's own knowledge to enhance its performance. Furthermore, by extending the context window, we enable the generation of longer texts while maintaining low perplexity.

3.1 SFT for Instruction Following

We propose a specialized system prompt to guide the model in generating well-structured long-text for formal document requests. The prompt instructs the model to first generate a summary and outline, followed by the complete document, with each section enclosed within designated tags. Additionally, we have reformatted the SFT dataset to incorporate these elements, which not only enhances the model's ability to adhere to instructions but also simplifies the subsequent evaluation of generated content through regular expression extraction.

System Prompt for Structured Long-Text Generation

> A conversation between User and Assistant. User presents a request for a formal document, and the Assistant generates a comprehensive and well-structured long-text based on the request. The assistant first conceives a summary and outline, and then produces the complete document. The summary, outline and document are enclosed within <summary></summary>, <outline></outline> and <answer></answer> tags, respectively, i.e., <summary>summary here</summary><outline>outline here</outline><answer>document here</answer>

3.2 Query Generation with Seed Data

To prevent overfitting from reusing SFT queries, we construct new queries for the alignment phase. We treat the training set queries as seed data and use a large language model to generate new alignment-phase queries. Specifically, we uniformly sample five queries from the training set as examples and use a prompt to direct a large language model to create new, theme-related long-text requests. The new queries, formatted similar to the examples, diversify our dataset, ensuring broad topic and intent coverage for robust training.

Prompt for Query Generation (Translated from Chinese)

> I want to generate question-answer pairs for the long-text generation task of my Large Language Model. Please generate new long-text generation requests that users may ask around the topics covered by the following examples. The output format should be similar to the examples.
> query #1: {queries[0]}
> query #2: {queries[1]}
> query #3: {queries[2]}
> query #4: {queries[3]}
> query #5: {queries[4]}
> Please organize the output in Python list format and use queries.extend(generated_query) to update the output.

3.3 Multi-Response Sampling and Quality Assessment

For each collected query Q^i, we generate m diverse responses $A_1^i, A_2^i, \ldots, A_m^i$ using top-p sampling. Using regular expressions, we extract the summary S_j^i, outline O_j^i, and document D_j^i from each response A_j^i, i.e., $S_j^i, O_j^i, D_j^i \triangleq \text{extract}(A_j^i)$. These extracted components facilitate subsequent quality evaluations across several dimensions, detailed as follows:

Content Length (L). The content length $L(A_j^i)$ measures the richness of the response A_j^i by counting the number of tokens in the extracted document D_j^i:

$$L(A_j^i) = \text{len}(\text{tokenizer}(D_j^i)). \tag{1}$$

Expression Diversity (E). The expression diversity $E(A_j^i)$, reflecting lexical variation in document D_j^i, is defined as the percentage of unique n-grams relative to the total n-grams in D_j^i:

$$E(A_j^i) = \frac{N_{\text{unique}}(D_j^i)}{N_{\text{total}}(D_j^i)} \times 100\%, \tag{2}$$

where $N_{\text{unique}}(D_j^i)$ is the count of unique n-grams in D_j^i, and $N_{\text{total}}(D_j^i)$ is the total number of n-grams in D_j^i.

Professionalism (P). The professionalism score $P(A_j^i)$ evaluates the domain specificity of document D_j^i by comparing the perplexity of two language models: a general foundation model M_{G} and its domain-adapted counterpart M_{D}, obtained via continual pretraining on domain-specific data. The metric is defined as:

$$P(A_j^i) = \mathcal{P}_{M_{\mathrm{G}}}(D_j^i) - \mathcal{P}_{M_{\mathrm{D}}}(D_j^i), \tag{3}$$

where $\mathcal{P}_M(D_j^i)$ represents the perplexity of D_j^i as assessed by model M. The domain-adapted model M_{D}, building on the general knowledge of M_{G}, shows lower perplexity for professional terminology due to domain-specific pretraining. A higher $P(A_j^i)$ indicates better alignment with the target domain's linguistic norms.

Relevance (R). The relevance score combines summary-level (R_s) and document-level (R_d) components computed between query Q^i and response components S_j^i/D_j^i. Following BGE-M3 (Chen et al., 2024), R_s and R_d can be computed as

$$\begin{aligned} R_s(A_j^i) &= 1.0r_{\text{dense}}^S + 0.3r_{\text{lex}}^S + 1.0r_{\text{mul}}^S \\ R_d(A_j^i) &= 0.15r_{\text{dense}}^D + 0.5r_{\text{lex}}^D + 0.35r_{\text{mul}}^D, \end{aligned} \tag{4}$$

where $r_{\text{dense}}, r_{\text{lex}}, r_{\text{mul}}$ represent the similarity scores from Dense Retrieval, Lexical Retrieval, and Multi-Vector Retrieval methods introduced in (Chen et al., 2024), respectively. Superscripts S and D denote similarities computed between: 1) Q^i-S_j^i for summary-level ($r_{\text{dense}}^S, r_{\text{lex}}^S, r_{\text{mul}}^S$) and 2) Q^i-D_j^i for document-level ($r_{\text{dense}}^D, r_{\text{lex}}^D, r_{\text{mul}}^D$). For further implementation details of these automatic metrics, please see Appendix A.

To establish comparative quality assessment within query groups, we propose a non-parametric rank sum method. For each query Q^i with m candidate responses $\{A_1^i, ..., A_m^i\}$, the composite score is computed as:

$$\begin{aligned} \rho_{\text{total}}^{(i,j)} =& \rho_L^{(i,j)} + \rho_E^{(i,j)} + \rho_P^{(i,j)} \\ &+ 0.8\rho_{R_d}^{(i,j)} + 0.2\rho_{R_s}^{(i,j)} + \rho_{-\mathcal{P}_{\mathrm{D}}}^{(i,j)}, \end{aligned} \tag{5}$$

where $\rho_X^{(i,j)} \in \{1, ..., m\}$ denotes the rank position of response A_j^i based on metric X within its query group Q^i, with lower ranks indicating better performance. The weighted combination coefficients (0.8/0.2) reflect the relative importance of document-level versus summary-level relevance, reflecting the greater emphasis on document-level relevance while acknowledging that summary-level relevance can aid in generating a more relevant document. Notably, we additionally include the perplexity term $\rho_{-\mathcal{P}_{\mathrm{D}}}^{(i,j)}$ in $\rho_{\text{total}}^{(i,j)}$ to ensure that the model's output distribution remains close to the domain-specific pre-training corpus.

The rank-sum method offers inherent robustness to non-uniform metric distributions and extreme values, avoiding the sensitivity of linear normalization methods like z-score or min-max normalization to skewed scales. By converting absolute scores into ordinal ranks, it achieves scale invariance while preserving interpretable relative comparisons among responses within the same query group. In this non-parametric framework, lower aggregate ranks indicate superior holistic quality.

3.4 Group-Reward Adaptive Calibration

Building upon the KTO algorithm, we introduce Group-Reward Adaptive Calibration to refine training signals for long-text generation. The method transforms binary rewards into dynamic quality-aware rewards through two key operations:

Group-wise Reward Calibration. For each query group Q^i with m responses $\{A_1^i, \dots, A_m^i\}$, we convert the rank-sum metric $\rho_{\text{total}}^{(i,j)}$ into normalized rewards Z_j^i via two operations: polarity inversion and intra-group standardization. The calibrated reward is computed as:

$$Z_j^i = \frac{-\rho_{\text{total}}^{(i,j)} - \mu_{(-\rho)}^{(i)}}{\sigma_{(-\rho)}^{(i)}}, \tag{6}$$

where the group statistics are derived from the negated rank-sums:

$$\begin{aligned} \mu_{(-\rho)}^{(i)} &= \frac{1}{m}\sum_{k=1}^{m}(-\rho_{\text{total}}^{(i,k)}), \\ \sigma_{(-\rho)}^{(i)} &= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(-\rho_{\text{total}}^{(i,k)} - \mu_{(-\rho)}^{(i)}\right)^2}. \end{aligned} \tag{7}$$

This transformation ensures higher original quality (lower ρ_{total}) translates to higher rewards. Meantime, it normalizes the response rewards of different queries onto the same scale, which shares similar spirit with GRPO (Shao et al., 2024), ensuring the comparability of response quality across different queries.

Global Normalization. Within this phase, the z-normalized rewards Z_j^i are subjected to a polarity-aware mass redistribution process encompassing the entire dataset. For instances where samples are positive, characterized by $Z_j^i \geq 0$, the corresponding weights are redistributed in accordance with the following formulation:

Method	Metrics					Avg. Rank
	L	E	P	R_s	R_d	
Baseline	3878.7 (5)	86.58 (5)	0.217 (4)	1.983 (5)	0.982 (5)	4.8
KTO	4903.7 (4)	87.48 (4)	0.213 (5)	1.991 (3)	0.987 (4)	4.0
GRACE (offline)	5137.7 (3)	88.27 (3)	0.223 (3)	1.998 (2)	0.993 (3)	2.8
GRACE (K=500)	5255.1 (2)	88.41 (2)	0.224 (2)	1.989 (4)	0.996 (2)	2.4
GRACE (K=250)	**5377.1 (1)**	**90.01 (1)**	**0.257 (1)**	**2.002 (1)**	**1.003 (1)**	**1.0**
	+38.6%	+4.0%	+18.2%	+1.0%	+2.2%	

Table 1: Performance comparison of different methods across various automatic metrics: L (Content Length), E (Expression Diversity), P (Professionalism), R_s (Summary-Level Relevance), and R_d (Document-Level Relevance). The numbers in parentheses indicate the ranking of each method for the corresponding metric. The last row shows the percentage improvement of GRACE (K=250) over the baseline.

$$\begin{aligned} W_j^i &= \frac{Z_j^i}{\sum_{(k,l)\in\Omega^+} Z_l^k} \cdot |\Omega^+|, \\ \Omega^+ &= \{(k,l)|Z_l^k \geq 0\}. \end{aligned} \tag{8}$$

Conversely, for negative samples where $Z_j^i < 0$, their weights are adjusted through scaling by the magnitude of their deviations, as expressed below:

$$\begin{aligned} W_j^i &= -\frac{Z_j^i}{\sum_{(k,l)\in\Omega^-} (Z_l^k)} \cdot |\Omega^-|, \\ \Omega^- &= \{(k,l)|Z_l^k < 0\}. \end{aligned} \tag{9}$$

This globally applied scaling mechanism ensures that the summation of weights for positive samples remains unchanged at $|\Omega^+|$, and likewise, the summation of weights for negative samples is preserved at $-|\Omega^-|$.

3.5 Semi-Online Training Strategy

Our semi-online framework employs periodic dataset renewal. In each iteration, the model generates K new queries, replacing the previous set to construct the GRACE-KTO dataset $\mathcal{D}_{\text{GRACE}} = \{Q^i, A_j^i, W_j^i\}_{i=1}^{K}{}_{j=1}^{m}$. Each query Q^i is paired with m responses $\{A_1^i, ..., A_j^i\}$ and associated weights $W_j^i \in \mathbb{R}$. Positive samples and negative samples are determined by:

$$y(W_j^i) = \text{sign}(W_j^i) = \begin{cases} +1 & W_j^i \geq 0 \\ -1 & \text{otherwise.} \end{cases} \tag{10}$$

The GRACE-KTO objective integrates our quality-aware weighting with the original KTO loss (Ethayarajh et al., 2024):

$$\begin{aligned} \mathcal{L}_{\text{GRACE-KTO}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{GRACE}}} \Big[&|W_j^i| \cdot \lambda(W_j^i) \\ &\cdot \big(1 - \sigma(\beta y(W_j^i) \Delta r(Q^i, A_j^i))\big) \Big], \end{aligned} \tag{11}$$

where σ denotes the sigmoid function, $\lambda(W_j^i)$ is the weighting function with the value of λ_+ and λ_- being the positive and negative sample coefficients ($\lambda_+|\Omega^+| \approx \lambda_-|\Omega^-|$), and β is a hyperparameter. $\Delta r(\cdot)$ denotes the policy log ratio relative to the reference point z_{ref}, defined as

$$\begin{aligned} \Delta r(Q^i, A_j^i) &= \log \frac{\pi_\theta(A_j^i|Q^i)}{\pi_{\text{ref}}(A_j^i|Q^i)} - z_{\text{ref}}, \\ z_{\text{ref}} &= \mathbb{E}_{\mathcal{D}_{\text{GRACE}}} \left[\text{KL}(\pi_\theta(A_j^i|Q^i) \| \pi_{\text{ref}}(A_j^i|Q^i)) \right]. \end{aligned} \tag{12}$$

Here, π_θ represents the current model, π_{ref} represents the reference model and the KL term represents the KL divergence between the model and reference model policies. By incorporating this reflection of response quality, W_j^i provides richer information during training.

3.6 Context Extension via YaRN

During inference, we utilize YaRN (Peng et al., 2024) interpolation integrated in vLLM (Kwon et al., 2023) to expand the model's context window to 32k tokens, enabling the generation of longer texts while maintaining perplexity.

4 Experiments

4.1 Experimental Settings

Model and Dataset. In our experiments, we employed the SOAEsV2-72B-Chat model provided by

Method	Metrics					Avg.	Avg. Rank
	CAR	TF	IDA	SC	PE		
Baseline	80.27 (5)	82.90 (4)	72.15 (5)	**70.75 (1)**	67.31 (5)	74.68	4.0
KTO	81.89 (4)	83.03 (2)	72.91 (4)	69.08 (5)	68.22 (4)	75.03	3.8
GRACE (offline)	82.73 (2)	83.02 (3)	74.71 (3)	69.42 (4)	68.43 (3)	75.66	3.0
GRACE (K=500)	82.33 (3)	82.87 (5)	74.91 (2)	69.94 (3)	**69.64 (1)**	75.94	2.8
GRACE (K=250)	**83.22 (1)**	**83.68 (1)**	**75.59 (1)**	70.59 (2)	69.54 (2)	**76.52**	**1.4**
Labeled Answer	82.02	80.67	74.04	63.65	67.02	73.48	-

Table 2: GPT-4o evaluations of different methods across multiple dimensions: CAR (Content Accuracy & Relevance), TF (Tone & Formality), IDA (Idea Development & Argumentation), SC (Structure & Clarity), and PE (Persuasiveness & Effectiveness)

(Deng et al., 2025). The model has undergone specialized pretraining and SFT on data related to Chinese state-owned assets and enterprises (SOAEs). Our dataset is derived from the report generation sub-dataset provided in their work. The specific task involves generating theme-related, content-rich, and professional reports based on a given research topic and outline. Specific query examples can be found in Appendix D.

Memory-Driven Method Selection. All experiments were performed on a single 8 × A800 (80 GB) node. To train a 72 B-parameter model within this ceiling, we combined 4-bit QLoRA (rank 64) (Dettmers et al., 2023) with DeepSpeed ZeRO-2 and FlashAttention-V2 (Dao, 2023).

Even with these optimizations, pairwise alignment methods (e.g., DPO, ORPO) exhaust memory because the “rejected” sample also involves in the forward–backward pass, instantly doubling the activation footprint. Online algorithms that alternate rollout and optimization (e.g., GRPO) encounter the same bottleneck. Consequently, only single-sample losses such as KTO remain executable; GRACE-KTO inherits this advantage without enlarging memory use. Hence, apart from the SFT baseline, KTO and proposed GRACE-KTO constitute our main experimental comparisons. For completeness, results under DeepSpeed ZeRO-3 + LoRA configuration are also provided in Appendix C.

Training Details. We first perform full-parameter SFT for one epoch (lr $= 1 \times 10^{-6}$) after re-formatting prompts to elicit instruction-following behavior. Alignment experiments are restricted to 1 000 unique queries; for each query we sample $m = 8$ responses, group them via regular-expression heuristics, and treat unmatched outputs as negatives ($W_j^i = -1$). All alignment runs (KTO, GRACE-KTO variants) keep the 4-bit QLoRA setting above, use a cosine LR schedule ($5 \times 10^{-6} \to 1 \times 10^{-6}$, 10 % warm-up), and are trained for one epoch with batch size 8. The KTO loss weight is $\beta = 0.05$; an auxiliary SFT loss on positive samples is added with weight $\mu = 0.1$.

Evaluation details. We computed automatic metrics on the test set as detailed in Section 3.3. To further evaluate our model’s responses, we employed a LLM-as-a-Judge method using GPT-4o (Achiam et al., 2023). The evaluation focused on five key criteria: Content Accuracy & Relevance (CAR), Tone & Formality (TF), Idea Development & Argumentation (IDA), Structure & Clarity (SC), and Persuasiveness & Effectiveness (PE). These dimensions are often subjective and complex, making them difficult to quantify with traditional metrics. Using GPT-4o for evaluation offers a more nuanced assessment than conventional automatic metrics alone. The detailed prompt is provided in Appendix E. To derive the final score, we converted the average score from these dimensions into a 0-100 scale by multiplying by 10, enabling a comprehensive quantitative assessment of our model’s response quality.

4.2 Main Results

Comprehensive Performance Evaluation. Figure 1 shows the z-score normalized performance across 10 metrics (5 automatic + 5 LLM-as-a-Judge), based on Tables 1 and 2. GRACE-KTO (K=250) achieves the highest z-scores in most dimensions, excelling in professionalism and Tone & Formality. Automatic metrics in Table 1 further shows our GRACE-KTO (K=250) produces a text length of 5377.1 tokens, a expression di-

Extension Method	8k (1×)		16k (2×)		32k (4×)		64k (8×)
	PPL	#Tokens	PPL	#Tokens	PPL	#Tokens	
Extrapolation	3.413	5377.1	3.321	6593.5	3.282	7238.8	-
Dynamic NTK	3.413	5377.1	3.549	6406.7	3.788	5842.8	-
YaRN	3.413	5377.1	**3.242**	**6614.9**	**3.188**	**8048.1**	OOM

Table 3: Performance comparison of different context window extension methods across various window sizes. PPL denotes perplexity, and #Tokens represents the number of tokens that can be processed.

versity score of 90.01, professionalism of 0.257, summary relevance of 2.002, and document relevance of 1.003, with the lowest average rank of 1.0. Compared to the SFT-only baseline, it shows improvements of 38.6% in length, 4.0% in diversity, 18.2% in professionalism, 1.0% in summary-level relevance, and 2.2% in document-level relevance. These results demonstrate GRACE-KTO's superiority over vanilla KTO.

Analysis of LLM-as-a-Judge Evaluation Results. Table 2 presents GPT-4o evaluations across five quality dimensions. GRACE-KTO (K=250) achieves the highest scores in CAR, TF, and IDA, with competitive performance in SC and PE. It attains the highest overall average score of 76.52 and the best average rank of 1.4. This demonstrates the effectiveness of our method in enhancing long-text generation quality.

Regarding evaluation results of labeled answers, it is important to note that in long-text generation task, it does not always represent the optimal response. This could be due to potential issues in the data labeling process, such as incomplete cleaning, or inherent ambiguities in long-text generation tasks where a single definitive answer may not exist. Consequently, the labeled data might not capture all aspects of high-quality responses, which further underscores the value of using advanced methods like GRACE-KTO for long-text generation tasks.

Comparison of Context Window Extension Methods. Table 3 compares three context window extension strategies—Extrapolation, Dynamic NTK (emozilla), and YaRN (Peng et al., 2024)—across various context lengths. With context lengths of 16k and 32k, YaRN achieves the lowest perplexity and the highest token throughput, demonstrating superior efficiency and generation quality. While Extrapolation shows moderate performance, Dynamic NTK suffers from higher perplexity and shorter generated contents, indicating less effective context extension. Despite encountering an out-of-memory issue at 64k, YaRN's strong performance at smaller context lengths establishes it as the most effective method. Its ability to extend context window size makes YaRN our preferred choice.

4.3 Ablation Studies

Effectiveness of Group-Reward Adaptive Calibration. As shown in Tables 1 and 2, offline GRACE-KTO consistently outperforms KTO across various evaluation metrics. For example, it achieves a notable increase in text length (5137.7 versus 4903.7), a higher diversity score (88.27 versus 87.48), and a significant improvement in Content Accuracy & Relevance (82.73 compared to 81.89). These enhancements can be attributed to the Group-Reward Adaptive Calibration mechanism. Unlike KTO, which relies on binary preference labels, offline GRACE-KTO incorporates nuanced, rank-based signals that better capture the degree of response quality. This richer feedback allows the model to more effectively learn from and optimize long-text generation.

Effectiveness of Semi-Online Strategy. As indicated in Tables 1 and 2, the transition from GRACE-KTO (offline) to GRACE-KTO (K=500), and subsequently to GRACE-KTO (K=250), underscores the advantages of our semi-online methodology. As previously established, GRACE-KTO (offline) already surpasses the baseline and KTO. Yet, the integration of the semi-online mechanism at K=500 further elevates performance across most metrics. The most pronounced enhancements are observed at K=250, demonstrating that a more aggressive semi-online strategy intensifies optimization. This configuration highlights the efficacy of our semi-online approach.

Training Dynamics of GRACE-KTO (K=250). Table 4 and Figure 3 illustrate the training trajectory of GRACE-KTO (K=250). As training progresses from 0% to 100%, the model demonstrates a gradual improvement across all five automatic metrics. These findings suggest that if the model can continuously generate diverse queries, its performance has the potential to be further enhanced.

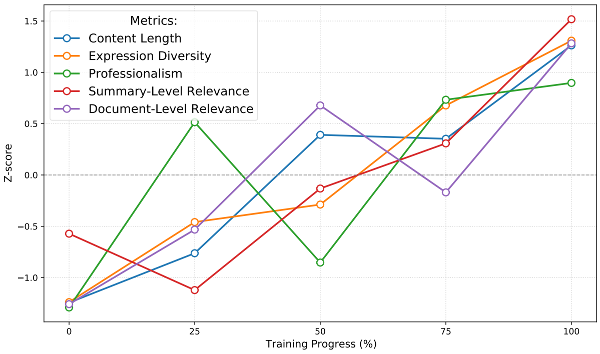

Figure 3: Z-score Normalized Metrics Dynamics During GRACE-KTO Training.

5 Conclusion

In this paper, we introduce GRACE-KTO, a semi-online framework designed to enhance the quality of long-text generation. GRACE-KTO aggregates responses into groups and calculates rank-sum scores across multiple linguistic quality dimensions. This allows it to adjust intra-group rewards and adaptively redistribute sample importance through group-wise and global normalization, fully utilizing KTO's flexibility with unpaired data. We implement a semi-online training strategy to minimize expensive online sampling costs and reduce reliance on labeled data by generating queries from seed data. Additionally, we extend the context window to 32k tokens using YaRN during inference. Experiments show GRACE-KTO outperforms vanilla KTO. Overall, our work provides a new and effective approach to long-context generation. Future research will focus on enhancing the reward strategy of GRACE-KTO by integrating more structured evaluation signals, and applying it to multi-domain text generation.

6 Limitations

In this work, we acknowledge several limitations that offer avenues for future improvement. Firstly, the quality and diversity of LLM-generated queries are critical yet constrained by the limited seed questions, potentially limiting the generation of diverse and high-quality training data. Secondly, constrained by application-driven requirements, our present experiments are exclusively conducted on Chinese data.

To solve these limitations, future directions should focus on:

- Scaling training data generation through hybrid human-AI collaboration frameworks to enhance both diversity and volume.
- Investigating cross-lingual capabilities beyond the current Chinese-language focus.

These enhancements could further strengthen GRACE-KTO's robustness for industrial-scale long-text generation scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grant 62471175 and the high-performance computing platform of Peking University.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Jingyang Deng, Ran Chen, Jo-Ku Cheng, and Jinwen Ma. 2025. Soaesv2-7b/72b: Full-pipeline optimization for state-owned enterprise llms via continual pre-training, domain-progressive sft and distillation-enhanced speculative decoding. *arXiv preprint arXiv:2505.04723*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

emozilla. Dynamically scaled rope further increases. https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/. Accessed: 2025-09-16.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, and 1 others. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Bolun Li, Zhihong Sun, Tao Huang, Hongyu Zhang, Yao Wan, Ge Li, Zhi Jin, and Chen Lyu. 2024. Ircoco: Immediate rewards-guided deep reinforcement learning for code completion. *Proceedings of the ACM on Software Engineering*, 1(FSE):182–203.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

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*.

Chau Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1722–1753.

Shanghaoran Quan, Tianyi Tang, Bowen Yu, An Yang, Dayiheng Liu, Bofei Gao, Jianhong Tu, Yichang Zhang, Jingren Zhou, and Junyang Lin. 2024. Language models can self-lengthen to generate long texts. *arXiv preprint arXiv:2410.23933*.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, and 1 others. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025. Longgenbench: Benchmarking long-form generation in long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Junhao Zhang, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye, and Yaowei Zheng. 2025. Lost-in-the-middle in long-text generation: Synthetic dataset, evaluation framework, and mitigation. *arXiv preprint arXiv:2503.06868*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Training Progress	Metrics					Avg. Rank
	L	E	P	R_s	R_d	
Baseline (0%)	3878.7 (5)	86.58 (5)	0.217 (5)	1.983 (4)	0.982 (5)	4.8
GRACE (25%)	4166.5 (4)	87.63 (4)	0.250 (3)	1.978 (5)	0.988 (4)	4.0
GRACE (50%)	4856.0 (2)	87.86 (3)	0.225 (4)	1.987 (3)	0.998 (2)	2.8
GRACE (75%)	4833.1 (3)	89.16 (2)	0.254 (2)	1.991 (2)	0.991 (3)	2.4
GRACE (100%)	**5377.1 (1)**	**90.01 (1)**	**0.257 (1)**	**2.002 (1)**	**1.003 (1)**	**1.0**

Table 4: Performance of GRACE-KTO (K=250) at different training stages. Automatic metrics include: L (Content Length), E (Expression Diversity), P (Professionalism), R_s (Summary-Level Relevance), and R_d (Document-Level Relevance).

A Implementation Details of Automatic Metrics

A.1 Professionalism Score Implementation

As described in Section 3.3, the professionalism score computation employs the following model configuration:

- **General-Purpose Model** M_G: Qwen2.5-7B (Yang et al., 2024) (base model without domain adaptation or task-specific tuning)

- **Domain Expert Model** M_D: SOAEsV2-7B (Deng et al., 2025) (domain-adapted model continual pre-trained from Qwen2.5-7B on domain-specific corpora)

Both models perform zero-shot scoring of responses, with the final professionalism score computed as a weighted combination of their individual scores.

A.2 Relevance Score Components

The relevance computation consists of three complementary retrieval approaches introduced in (Chen et al., 2024), using their BGE-m3 model:

- **Dense Retrieval**: Given input query q and passage p, their representations are derived from encoder hidden states:

$$\begin{aligned} e_q &= \text{norm}(\mathbf{H}_q[0]), \\ e_p &= \text{norm}(\mathbf{H}_p[0]), \\ r_{\text{dense}} &= \langle e_p, e_q \rangle = e_p^\top e_q. \end{aligned} \tag{13}$$

  where $\mathbf{H}_q[0]$ and $\mathbf{H}_p[0]$ denote the [CLS] token embeddings, and $\text{norm}(\cdot)$ is L2-normalization.

- **Lexical Retrieval**: Term weights are learned through neural projections:

$$\begin{aligned} w_{q_t} &= \max_{i \in \text{pos}(t)} \text{ReLU}(\mathbf{W}_{\text{lex}}^\top \mathbf{H}_q[i]), \\ w_{p_t} &= \max_{j \in \text{pos}(t)} \text{ReLU}(\mathbf{W}_{\text{lex}}^\top \mathbf{H}_p[j]), \\ r_{\text{lex}} &= \sum_{t \in q \cap p} w_{q_t} \cdot w_{p_t}. \end{aligned} \tag{14}$$

  where $\text{pos}(t)$ indicates positions of term t, $\mathbf{W}_{\text{lex}} \in \mathbb{R}^d$ is the learnable weight vector.

- **Multi-Vector Retrieval**: Utilizes full token embeddings with projection:

$$\begin{aligned} E_q &= \text{norm}(\mathbf{W}_{\text{mul}}^\top \mathbf{H}_q), \\ E_p &= \text{norm}(\mathbf{W}_{\text{mul}}^\top \mathbf{H}_p), \\ r_{\text{mul}} &= \frac{1}{N} \sum_{i=1}^{N} \max_{1 \le j \le M} E_q[i]^\top E_p[j]. \end{aligned} \tag{15}$$

  where $\mathbf{W}_{\text{mul}} \in \mathbb{R}^{d \times d}$ is the projection matrix, N and M are query/passage lengths.

In our framework, the query corresponds to Q^i while passages are either summaries S_j^i or documents D_j^i.

B Detailed Training Dynamics of GRACE-KTO

The results in Table 4 show a progressive improvement in performance across all metrics as training progresses. The 100% training stage achieves the best performance in all metrics, with the highest values for content length, expression diversity, professionalism, summary-level relevance, and document-level relevance. The average rank also improves, reaching the highest value of 1.0 at 100% training progress.

Method	L	E	P	R_s	R_d
Baseline	3878.7	86.58	0.217	1.983	0.982
ORPO (offline)	3707.3	86.04	0.190	1.984	**0.994**
DPO (offline)	3938.8	87.72	0.217	1.992	0.983
KTO (offline)	4555.0	88.32	0.224	**2.006**	0.990
GRACE (offline)	**5080.5**	**88.76**	**0.237**	1.999	0.991
GRPO (online)	OOM	OOM	OOM	OOM	OOM

Table 5: ZeRO-3 + LoRA results (no quantization). L = length, E = diversity, P = professionalism, R_s = summary relevance, R_d = document relevance.

Method	CAR	TF	IDA	SC	PE	Avg.
Baseline	80.27	82.90	72.15	70.75	67.31	74.68
ORPO (offline)	79.57	83.05	71.55	70.86	67.06	74.42
DPO (offline)	80.32	81.32	73.21	70.11	68.21	74.63
KTO (offline)	80.84	83.41	73.85	**71.01**	68.10	75.44
GRACE (offline)	**82.93**	**83.81**	**74.97**	70.50	**68.84**	**76.21**
GRPO (online)	OOM	OOM	OOM	OOM	OOM	OOM

Table 6: GPT-4o evaluation under ZeRO-3 + LoRA.

C DeepSpeed ZeRO-3 + LoRA Validation

Switching to ZeRO-3 + LoRA (rank 64, no quantization) lowers per-GPU memory pressure sufficiently to enable DPO and ORPO, while GRPO again exhausts memory during training; we therefore include these pairwise alignment methods in the comparison. All models were trained on the same dataset.

Automatic metrics (Table 5) show ORPO underperforming the SFT baseline on length, expression diversity and professionalism, while DPO also achieves gains compared to the SFT baseline. However, KTO surpasses both pairwise alignment methods, and GRACE-KTO further extends the lead across all axes.

LLM-as-a-Judge scores (Table 6) place ORPO and DPO slightly below the baseline on average, corroborating that pairwise preference optimization yields limited benefit for long-text generation. GRACE-KTO retains the highest average rating without additional memory cost, confirming that single-sample KTO plus group-wise calibration remains preferable in this setting.

D Example of Query for Report Writing Task

An Example of Query for Report Writing Task (Traslated from Chinese)

Please write a report titled "Management Practice of Grid-Source Integration Digitalization Construction in Northeast Combined Heat and Power (CHP) Enterprises", covering the following sections:

1. **Background**
* Company Profile and Reform Background
* Dilemmas before Separation and Transfer
* Challenges after Separation and Transfer

2. **Connotation of Grid-Source Integration Digitalization Construction Management Practice**
* Digitalization and Intelligent Management
* System Framework
* Multiple Mechanisms

3. **Practices of Grid-Source Integration Digitalization Construction Management**
* Platform Development
    * Digitalization and Intelligent Management Platform Construction
    * Big Data Analysis Platform
    * Precision Control
    * Secondary Network Hydraulic Balance
    * Pilot Operation of Model Demonstration Zone
* Management System and Mechanism Innovation
    * Management System Innovation
    * Performance Incentive Mechanism Innovation
    * Supervision and Management Innovation
    * Environmental Protection Red Line Management System

4. **Implementation Effects of Grid-Source Integration Digitalization Construction Management Practice**
* Preservation and Appreciation of State-owned Assets and Livelihood Protection
* Safety and Environmental Standards Compliance
* Innovation in Management System and Mechanism
* Green and Low-carbon Economy
* Green Emission Reduction
* Intelligent Regulation Technology
* Enhanced Social Influence

E Prompt for LLM-as-a-Judge Evaluation using GPT-4o

Prompt for LLM-as-a-Judge Evaluation

You are an expert evaluator of written responses on the topic of Chinese state-owned assets and enterprises (SOEs). Your role is to critically assess a candidate's response to a writing prompt, focusing on how effectively it addresses the original task. Your evaluation should be rigorous—not lenient—and should highlight meaningful distinctions in quality. Assess the response according to the following criteria, assigning each a score from 1 to 10 (1 = extremely poor, 5 = average, 10 = outstanding):
Evaluation Criteria:

1. **Content Accuracy and Relevance**: Does the response demonstrate a sound and accurate understanding of the issues raised, particularly Chinese government policies, official statements, and reform priorities concerning SOEs and state-owned assets?

2. **Tone and Formality**: Is the tone appropriate for an official or institutional context? Does it maintain a consistent level of formality throughout?

3. **Idea Development and Argumentation**: Are viewpoints clearly articulated and well-supported with logical reasoning, evidence, or policy references? Are the ideas developed thoroughly and insightfully?

4. **Structure and Clarity**: Is the writing logically organized, easy to follow, and coherent? Is the expression varied yet precise?

5. **Persuasiveness and Effectiveness**: Does the response communicate its points compellingly and persuasively, while maintaining clarity and professionalism?

<User Request >instruction </User Request ><Response >response </Response >Instructions: First, provide a concise overall analysis of the response, noting major strengths and weaknesses. Then, deliver a detailed evaluation in strict JSON format as follows: "Analysis": "Your analysis here.", "Content Accuracy and Relevance": score, "Tone and Formality": score, "Idea Development and Argumentation": score, "Structure and Clarity": score, "Persuasiveness and Effectiveness": score