Improving Conversational Abilities of Quantized Large Language Models via Direct Preference Alignment

Janghwan Lee[1*], **Seongmin Park**[1*], **Sukjin Hong**[1,2], **Minsoo Kim**[1],
Du-Seong Chang[2] and **Jungwook Choi**[1†]
[1]Hanyang University, [2]KT
Seoul, Republic of Korea
[1]{hwanii0288, skstjdals, sjhong7898, minsoo2333}@hanyang.ac.kr
[2]{sukjin.hong, dschang}@kt.com, [1†]choij@hanyang.ac.kr

Abstract

The rapid advancement of large language models (LLMs) has facilitated their transformation into conversational chatbots that can grasp contextual nuances and generate pertinent sentences, closely mirroring human values through advanced techniques such as instruction tuning and reinforcement learning from human feedback (RLHF). However, the computational efficiency required for LLMs, achieved through techniques like post-training quantization (PTQ), presents challenges such as token-flipping that can impair chatbot performance. In response, we propose a novel preference alignment approach, quantization-aware direct preference optimization (QDPO), that aligns quantized LLMs with their full-precision counterparts, improving conversational abilities. Evaluated on two instruction-tuned LLMs in various languages, QDPO demonstrated superior performance in improving conversational abilities compared to established PTQ and knowledge-distillation fine-tuning techniques, marking a significant step forward in the development of efficient and effective conversational LLMs.

Question from User
Explain what is meant by the circle of fifths.

Response by Assistant
(a) 16-bit Baseline (CSQA: 75.2%, MMLU: 46.6%)
The circle of fifths is a **musical** diagram that represents the relationship between the intervals of the octave, the fifth, and the fourth.
Token probability: **musical (46.1%)** vs. visual (14.9%)

(b) W4A16 AWQ (CSQA: 74.8%, MMLU: 45.1%)
The circle of fifths is a **visual** representation of the relationship between the intervals of the circle of fifths, which is a circle divided into 360 degrees.
Token probability: musical (17.4%) vs. **visual (32.4%)**

(c) W4A16 QDPO (CSQA: 74.3%, MMLU: 45.0%)
The circle of fifths is a **musical** diagram that represents the relationship between the notes of the octave in a musical scale.
Token probability: **musical (29.5%)** vs. visual (23.0%)

Figure 1: Example responses generated by Mi:dm-7B on 16-bit and 4-bit quantized inference.

1 Introduction

As large language models (LLMs) advance in understanding the context of language and generating relevant sentences, LLMs are evolving into conversational chatbots that can naturally respond to a wide array of user requests (OpenAI, 2023; Chiang et al., 2023; Team et al., 2023; Touvron et al., 2023b). Particularly noteworthy is the remarkable ability of LLMs to follow user instructions and align with human values, such as providing helpful and engaging responses through techniques like instruction tuning and reinforcement learning from human feedback (RLHF) (Taori et al., 2023; Longpre et al., 2023; Chung et al., 2022; Mukherjee et al., 2023; Ouyang et al., 2022). These advancements have greatly enhanced the capability to fine-tune pre-trained LLMs for various tasks and user preferences.

For the effective implementation of LLM-based chatbots, addressing LLMs' computational complexity is essential. Weight load overhead, a critical bottleneck in LLM deployment, has led to the development of weight quantization techniques like post-training quantization (PTQ). PTQ reduces storage requirements by applying quantization to the weights of trained LLMs, thereby decreasing the necessary bit count for weight data storage (Frantar et al., 2023; Lin et al., 2023). Techniques such as AWQ (Lin et al., 2023) address quantization-induced accuracy loss through methods like scaling data distribution and weight updates aimed at preserving accuracy. The effectiveness of these quantization strategies has been measured by task-dependent benchmarks to evaluate model accuracy instead of multifaceted conversational qualities.

Evaluating the conversational abilities of LLM-based chat assistants, especially for open-ended

*Equal contribution †Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11346–11364
August 11–16, 2024.

tasks requiring alignment with human preferences, challenges traditional score-based benchmarks due to the assistants' varied capabilities. To address this, new methods have been introduced for a more objective assessment of LLM chatbot performance (Chiang et al., 2023; Zheng et al., 2023). The "LLM as a Judge" approach (Zheng et al., 2023) employs advanced LLMs like GPT-4 (OpenAI, 2023) to evaluate responsiveness in multi-turn conversations across eight conversational categories, focusing on conversational continuity and adherence to instructions. Furthermore, FLASK (Ye et al., 2023) offers fine-grained evaluation criteria that dissect conversational skills linguistically. Yet, these methods mainly target full-precision chatbots, leaving the performance of cost-efficient quantized LLM chatbots less explored.

To assess quantization's effect on LLM-based chatbots' conversational abilities, we qualitatively compared the responses of quantized LLMs with a 16-bit baseline. Fig. 1 reveals that quantized models often fail to maintain engaging dialogues with repetitive phrases. We identify "token-flipping" — a phenomenon where quantization errors skew token distribution, causing incorrect token selection — as a crucial factor for this quality degradation. Traditional task-dependent evaluation metrics, such as Common Sense Question Answering (CSQA) (Talmor et al., 2019) and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), may not fully detect these nuances. For example, as shown in Fig. 1(a) and (b), 16-bit and W4A16 inference exhibit similar task accuracy, but W4A16 inference produces responses that are not helpful to the user. This observation underscores the need for a new quantization approach that preserves user-perceived effectiveness beyond the task-dependent benchmarks.

To address the issue of token-flipping in quantized LLMs, we propose a novel preference alignment method that aligns quantized LLMs with full-precision counterparts. Drawing inspiration from direct preference optimization (DPO) strategies (Rafailov et al., 2023; Liu et al., 2023a), our approach generates preference datasets directly from the quantized LLM and its full-precision counterpart to implement quantization-aware optimization for preference-reflective weight adjustments. Our quantization-aware direct preference optimization (QDPO) method improves the disparity between the top-1 and top-2 logits of token distribution, reducing token-flipping, and fostering more relevant and consistent text output. We rigorously tested QDPO on two instruction-tuned LLMs, Vicuna (Zheng et al., 2023) and Mi:dm (KT-AI, 2023), assessing their conversational performance in both English and Korean. The results, as illustrated in Fig. 1(c), demonstrate that QDPO markedly enhances conversational abilities beyond those achieved with established quantization techniques.

2 Background

2.1 Conversational Ability of LLM

In the pre-training phase, LLMs learn from a vast corpus of text data collected from various sources, including the internet, books, articles, and conversations (Raffel et al., 2019; Zhu et al., 2015; Gao et al., 2020; Penedo et al., 2023). Through this process, they acquire extensive knowledge on a wide range of topics, which forms the foundation that enables LLMs to flexibly respond to diverse conversational subjects (Zhang et al., 2022; Touvron et al., 2023a,b; Brown et al., 2020). Subsequently, LLMs develop the capability to follow instructions through instruction fine-tuning and learn to align with human preferences via RLHF (Taori et al., 2023; Longpre et al., 2023; Chung et al., 2022; Mukherjee et al., 2023; Ouyang et al., 2022). Through such processes, LLM-based chatbots like GPT-4 (OpenAI, 2023) and Vicuna (Chiang et al., 2023) have acquired the conversational ability to engage with humans on various topics over multiple turns, distinguishing them from conventional language models.

To evaluate LLM-based chatbots, it is essential to assess their conversational ability, which is their key capability. However, existing task-dependent benchmarks such as MMLU (Hendrycks et al., 2020) and HELM (Liang et al., 2023) do not adequately capture human preferences, rendering them insufficient for evaluating LLM-based chatbots. In response, proposals for new benchmarks such as MT-Bench (Zheng et al., 2023) and FLASK (Ye et al., 2023) are emerging, focusing on multi-turn questions or alignment with human preferences to effectively evaluate conversational abilities.

2.2 LLM Quantization

LLMs demand high serving costs due to their extensive number of parameters (Brown et al., 2020).

Weight quantization techniques (Lin et al., 2023; Frantar et al., 2023; Lee et al., 2023a; Kim et al., 2023; Lee et al., 2023b) address this issue by representing the model's weights in lower bit-precision, thereby reducing memory size, lowering memory load time, and speeding up inference. Post-training quantization (PTQ) changes the model's weights directly to lower precision without additional training, offering cost benefits. However, due to concerns about accuracy loss, PTQ utilizes a portion of the training samples to calibrate and minimize the layer-wise quantization error through methods such as AWQ (Lin et al., 2023). Quantization-aware training (QAT), on the other hand, maintains the performance of a quantized model by applying quantization during the forward pass and training the model accordingly. When applying QAT to LLMs, due to the insufficient information from the ground truth, techniques often use Knowledge Distillation (KD) by reducing the distance between the logits of the quantized model and the full-precision model (Kim et al., 2023; Liu et al., 2023b).

However, previous quantization studies have evaluated their methods on task-dependent benchmarks, which show a limited scope for comprehensive evaluation of conversational abilities. For example, AWQ (Lin et al., 2023) emphasizes that the quantized model achieves accuracy comparable to the baseline on CSQA. However, they do not analyze why only 35% of the sentences generated by the quantized model are considered as good as those from the baseline, according to GPT-4 (OpenAI, 2023)'s evaluation in assessing conversational abilities. In this research, we analyze how the model's quantization error impacts the conversational abilities of large language models and propose methods to enhance these capabilities.

2.3 Alignment with Human Preferences

The RLHF is an advanced method to improve the performance of LLMs by aligning with human preferences. It comprises three stages:

Supervised Fine-Tuning (SFT). SFT utilizes a dataset of human instructions to refine pre-trained LLMs.

Reward Modeling. This stage develops a reward model based on human preferences for LLM response pairs, using the Bradley-Terry (BT) model (Bradley and Terry, 1952) to quantify these preferences. It represents the distribution of preferences distribution p^* between y_1 and y_2:

$$p^*(y_1 \succ y_2|x) = \frac{e^{r^*(x,y_1)}}{e^{r^*(x,y_1)} + e^{r^*(x,y_2)}}, \quad (1)$$

r^* is defined as the optimal reward function. y_1 and y_2, assumed to be sampled from the optimal preference distribution p^* with prompt x, the parameterized reward model estimates the parameter using maximum likelihood.

Policy Optimization. The LLM policy optimization is guided by the reward model to generate responses that better align with human preferences for the training prompts. The reinforcement learning (RL) objective function is defined as follows:

$$\max_{\pi} \mathop{\mathbb{E}}_{\substack{x\sim D \\ y\sim\pi}} \left[r(x,y)\right] - \beta D_{KL}\left[\pi(y|x)\|\pi_{\text{ref}}(y|x)\right], \quad (2)$$

π represents the LLM policy, β is a control parameter that regulates variations with respect to π_{ref}. Recent approach (Ouyang et al., 2022) employs Proximal Policy Optimization (Schulman et al., 2017) for RL-based optimization, wherein the necessary reward is derived from a previously trained reward model.

DPO (Rafailov et al., 2023) aligns LLM policies with human preferences via supervised learning, leveraging Eq. (2) to relate the optimal reward to the optimal policy directly.

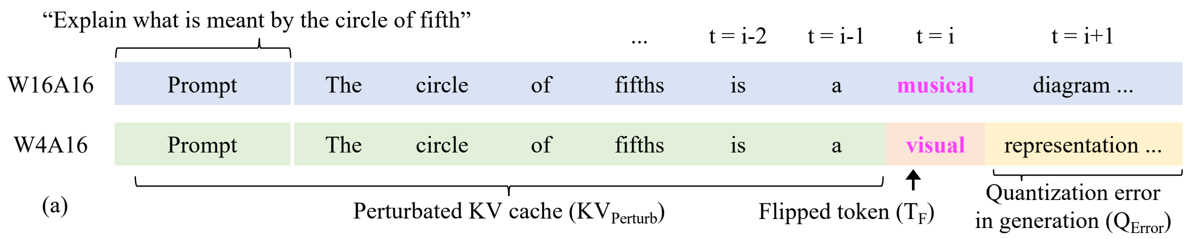
$$r^*(x,y) = \beta \log\left(\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \beta \log Z(x), \quad (3)$$

where $Z(x)$ is the partition function. The optimal reward function is fitted to the objective function of BT model, defining DPO loss as follows:

$$\mathop{\mathbb{E}}_{x,y_w,y_l\sim D}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \quad (4)$$

where σ is logistic function.

SRO (Liu et al., 2023a) criticizes the preference sampling method of DPO. Sampling data y_w and y_l from the π^* is the optimal way for estimating π_θ. However, all experiments in DPO use preference pairs not from the π^* but from π_{ref}, and there is a lack of research into the implications of this approach. SRO proposes a solution by constructing an additional reward-ranking model to directly form preference pairs from an approximated optimal policy and statistical rejection sampling.



Case	T_F	$KV_{Perturb}$	Q_{Error}	ROUGE-L
1	✓	✓	✓	0.3455
2	✓	✓	–	0.3736
3	✓	–	✓	0.3702
4	✓	–	–	0.4155
5	–	✓	✓	0.4921
6	–	✓	–	0.5858
7	–	–	✓	0.5624
8	–	–	–	1.0000

(b)

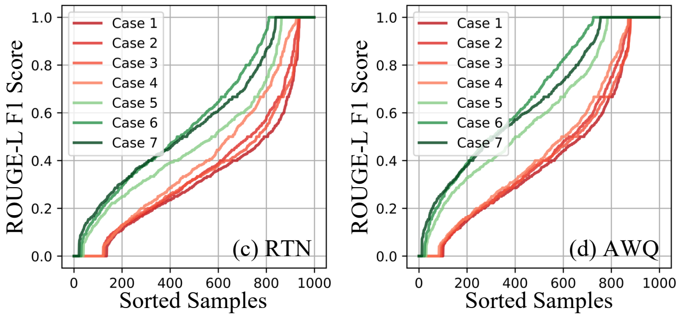


Figure 2: (a) Breakdown of factors influencing sentence generation in quantized models. (b) Case study on the impact of each factor. The ROUGE-L score is used to measure changes in sentences. More results for ROUGE-1/2 are in Fig. 7. (c-d) Case-wise ROUGE scores in models where W4A16 PTQ is applied with (c) RTN and (d) AWQ.

3 Conversational Abilities of Quantized LLMs

3.1 Observations

Recent advancements in PTQ have demonstrated that 4-bit quantized LLMs are effective for a variety of tasks, as evidenced by references such as AWQ and OPTQ (Lin et al., 2023; Frantar et al., 2023). However, our observations reveal that these quantized LLMs struggle to sustain engaging conversations, particularly in multi-turn chatbot interactions. For instance, Fig. 1 illustrates the contrast between the 16-bit baseline and 4-bit quantized LLMs in sentence generation. The baseline model begins its responses with “The circle of fifths is a musical diagram,” providing relevant answers. On the other hand, the 4-bit quantized model starts to deviate at the seventh token, switching its focus from “musical” to “visual,” and often generates limited and repetitive phrases. Although both models display similar task performance metrics, such as accuracy in multiple-choice benchmarks, there’s a noticeable difference in the logit probability for the seventh token in the 4-bit model, causing a change in the token from “musical” to “visual.” This issue of altered text generation, observed across multiple examples (see A.9 for additional examples), prompts an investigation into its underlying causes.

3.2 Breakdown Analysis

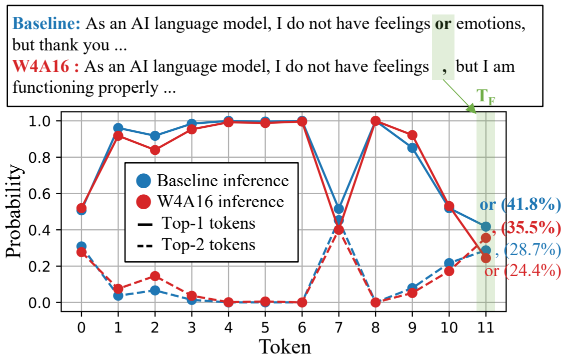
To understand the cause of altered text generation in quantized LLMs, we examine how quantization impacts text production. We pinpoint the initial deviation to a flipped token and identify three contributing factors as shown in Fig. 2(a):

- **Flipped token** (T_F)**:** Occurs when a quantized model selects a different token at timestep $t = i$ compared to the baseline, altering the input for subsequent token generation and leading to deviations.

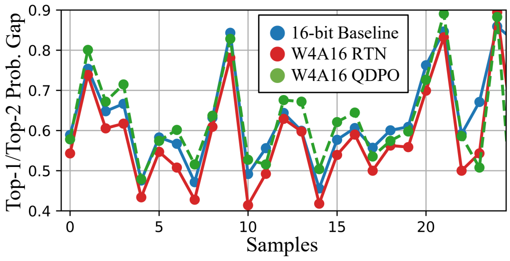
- **Perturbated KV cache** ($KV_{Perturb}$)**:** Despite identical token sequences up to timestep $t = i - 1$, quantization errors already affect the Transformer’s key-value caches, contributing to further deviations.

- **Quantization error in generation** (Q_{Error})**:** Starting from timestep $t = i + 1$, ongoing quantization errors continue to influence token generation, causing further divergence from the baseline.

Setup. To evaluate the impact of each identified factor, we analyze eight possible scenarios shown in Fig. 2(b). Case 1, where all three factors are present, mirrors the standard text generation of a quantized LLM, whereas Case 8, devoid of these factors, corresponds to the baseline model’s inference. For this analysis, we generate text using both 4-bit and 16-bit models with 1,000 instruction samples randomly chosen from the Alpaca dataset (Taori et al., 2023). We record the first token where discrepancies in text generation between



(a) Token flipping in quantized inference



(b) Average gap between Top-1/Top-2 probability per sample

Figure 3: (a) Auto-regressive inference probabilities for baseline and quantized models, token by token. (b) Difference in average probability between top-1 and top-2 tokens per sample (Mi:dm, from MT-Bench). See Fig. 9 for more on the AWQ case.

the two models occurred, along with the key-value cache status up to that point for each scenario. To quantify the deviation from the baseline text, we utilize the ROUGE-L (Lin, 2004) as a metric (the higher the better). More details of the implementation for the breakdown analysis are provided in A.2.

Results. To assess the contribution of each factor to deviations in sentence generation from the baseline, we contrast each scenario with Case 8. As depicted in Fig. 2(b), sentences become more divergent with the inclusion of additional factors. Specifically, from Case 4, it is evident that the Flipped Token ($\mathrm{T_F}$) significantly affects sentence variation, as indicated by the largest decrease in ROUGE scores. Conversely, the effects of perturbed KV cache ($\mathrm{KV_{Perturb}}$) and quantization error in generation ($\mathrm{Q_{Error}}$) are comparatively minor. This pattern is further highlighted in Fig. 2(c), where ROUGE scores, sorted by samples, show that Cases 1-4 cluster on the right, signifying greater deviation. This suggests that even a single token difference, resulting from quantization-affected probability shifts, can substantially alter the overall sentence structure in quantized inference.

Ablation: Advanced Quantization. Advanced quantization techniques designed to minimize errors may not fully address the issue of deviated text generation caused by flipped tokens. Recent PTQ methods that employ calibration using a small sample by scaling weight channels or adjusting quantization step sizes (Frantar et al., 2023; Lin et al., 2023; Lee et al., 2023b) aim to lessen layer-specific quantization errors. However, our observations indicate that while these calibrated PTQ models reduce quantization error effects, they do not mitigate the issues stemming from flipped tokens. The case study for a 4-bit quantized model calibrated with AWQ (Lin et al., 2023), shown in Fig. 2(d), reveals that although calibration decreases the impacts of $\mathrm{KV_{Perturb}}$ and $\mathrm{Q_{Error}}$, sentence variations are still predominantly influenced by $\mathrm{T_F}$. A similar trend can be observed by KD-based QAT (Fig. 8), highlighting the need for strategies that specifically address flipped tokens.

3.3 Why Token-Flipping Happens?

We hypothesize that token-flipping occurs due to inherently ambiguous token distributions in sentence generation, which become prone to flipping when quantization errors introduce alterations. To empirically validate this, Fig. 3(a) demonstrates token-flipping during text generation by a quantized model. It shows the probabilities for the top-1 and top-2 tokens throughout the auto-regressive generation. Notably, the 16-bit baseline and 4-bit quantized models produce nearly identical probabilities for most tokens. However, at certain points (e.g., $t = 0, 7, 11$), the probability margin between the top-1 and top-2 tokens is minimal. Token-flipping occurs when quantization-induced deviations in the probability distribution surpass this narrow margin, altering subsequent sentence generation and leading to unnatural phrasing.

Fig. 3(b) shows the average probability margin between the top-1 and top-2 tokens across each text sample. By feeding identical inputs to each model, we note that the 4-bit quantized model has a narrower average probability margin between the top-1 and top-2 tokens than the 16-bit baseline. This indicates a higher likelihood of the 4-bit model experiencing token-flipping due to quantization error-induced deviations exceeding this margin. Additionally, our examination of beam search (Graves, 2012) in Section 5.5 reveals its limited effectiveness in mitigating this issue. This underscores the need for strategies that ensure the quantized model retains clear decision-making capabilities.

4 QDPO: Quantization-aware Direct Preference Optimization

As described in Section 3, quantization significantly degrades the conversational ability of LLMs. To address this issue, we introduce an algorithm named Quantization-aware Direct Preference Optimization (QDPO), which aims to align the conversational abilities of quantized LLMs with those of LLMs prior to quantization. QDPO has two main contributions: 1) Providing an efficient method for generating the dataset $\mathcal{D}_{\text{QDPO}}$ without costly human annotations. 2) Offering a theoretical foundation that ensures the automatic distinction of preferences during dataset generation.

4.1 Method

Drawing inspiration from the success of DPO in aligning LLMs with human preferences, we have developed a novel approach that extends its application to overcome the challenges introduced by quantization.

The challenge in preference dataset generation arises from human labeling. To mitiagte this, we introduce for efficiently creating dataset $\mathcal{D}_{\text{QDPO}}$, which is composed of triplets $\{y_w, y_l, x\}$. Here, y_w denotes the response from the full-precision model π_{fp}, which is also referred to as the optimal policy. y_l represents the corresponding response from the quantized model π_{q}. x serves as the prompt. Specifically, y_w is obtained as $\arg\max_y \pi_{\text{fp}}(y|x)$ and y_l as $\arg\max_y \pi_{\text{q}}(y|x)$. The preference of y_w over y_l is ensured by Theorem 1. The proof is in A.8. Unlike conventional DPO methods, QDPO automatically distinguishes preferences without relying on expensive human-annotated datasets.

Theorem 1. *For any response y in the set of all possible responses Y, if $y_1 = \arg\max_{y \in Y} \pi_{fp}(y|x)$ and $y_2 = \arg\max_{y \in Y} \pi_q(y|x)$, then it is guaranteed that $p^*(y_1 \succ y_2) \geq p^*(y_2 \succ y_1)$.*

By precisely distinguishing preferences, we can clearly eliminate errors in data labeling. This leads to improved performance by accurately estimating the policy model's density. $\mathcal{L}_{\text{QDPO}}$ is define with high-quality preference data $\mathcal{D}_{\text{QDPO}}$ as follows:

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_{\text{QDPO}} \\ y_w \sim \pi_{\text{fp}}, y_l \sim \pi_{\text{q}}}} \left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{q}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{q}}(y_l|x)} \right) \right]. \quad (5)$$

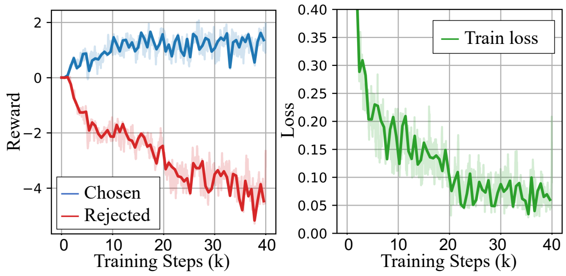


Figure 4: Training dynamics of QDPO showing chosen and rejected rewards (left), and loss (right) across steps.

Algorithm 1 Quantization-aware DPO

Input: prompt $\{x_1, x_2, \ldots, x_N\}$, full precision policy π_{fp}, quantized policy π_{q}, KL penalty β
Output: Updated policy π_θ
Initialize π_θ from π_{q}
Preference pairs dataset $\mathcal{D}_{\text{QDPO}} = \emptyset$
for $i = 1$ to N **do**
  $y_w^i = \arg\max_y \pi_{\text{fp}}(y|x^i)$
  $y_l^i = \arg\max_y \pi_{\text{q}}(y|x^i)$
  Add pair $\{y_w^i, y_l^i, x^i\}$ to $\mathcal{D}_{\text{QDPO}}$
end for
for each pair $\{y_w, y_l, x\}$ in $\mathcal{D}_{\text{QDPO}}$ **do**
  Calculate $\mathcal{L}_{\text{QDPO}}$ from eq. (5)
  Calculate the gradient with respect to θ
  $\frac{\partial \mathcal{L}_{\text{QDPO}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{QDPO}}}{\partial \theta_{\text{q}}} \cdot \frac{\partial \theta_{\text{q}}}{\partial \theta} \underset{\text{STE}}{\approx} \frac{\partial \mathcal{L}_{\text{QDPO}}}{\partial \theta_{\text{q}}}$
  Update π_θ by minimizing $\mathcal{L}_{\text{QDPO}}$
end for
return Updated policy π_θ

4.2 Implementation

Given π_θ as the quantized model's policy, integrating $\mathcal{L}_{\text{QDPO}}$ with QAT adjusts for quantization effects. The quantization technique we employ uniformly quantizes each channel across its entire min-max range, ensuring comprehensive accommodation of the full spectrum of values within each channel. To overcome the challenge posed by the non-differentiable rounding within the quantization process, we employ the Straight-Through Estimator (STE) for gradient approximation, facilitating effective gradient approximation and ensuring smooth training despite quantization. As shown in Fig. 4, QDPO demonstrates an increase in the chosen reward and a decrease in the rejected reward throughout the training process, indicating effective loss convergence. The complete procedure of QDPO is described in Algorithm 1. Details of the training settings and hyperparameters for QDPO can be found in A.1.

5 Experiments

5.1 Experimental Settings

Models. We evaluate QDPO on two representative conversational LLMs. Vicuna-v1.5 (Zheng et al., 2023), instruction-finetuned from LLaMA2 for improved conversational ability, and a bilingual (English-Korean) LLM, Mi:dm (KT-AI, 2023), to confirm QDPO's effectiveness to support multiple languages. All these models have 7B parameters.

Benchmarks. For a comprehensive evaluation of conversational abilities, we employ three distinct benchmarks: MT-Bench (Zheng et al., 2023), Vicuna-Eval (Chiang et al., 2023), and FLASK (Ye et al., 2023). MT-Bench utilizes GPT-4 to evaluate the quality of two responses obtained from an initial question and an additional follow-up question, offering an evaluation of multi-turn responses. For assessing Korean capability, we also translate the MT-Bench dataset into Korean using GPT-4. Vicuna-Eval consists of 80 questions for evaluation by GPT-4 to determine which model generates better sentences. FLASK includes 1.7K samples designed to assess LLM's fine-grained language skills, such as robustness and harmlessness.

Quantization Methods. To understand the impact of quantization on conversational abilities, we consider variations of quantization methods:

- Baseline: 16-bit floating-point weight
- RTN (Jacob et al., 2018): 4-bit round-to-nearest weight quantization
- AWQ (Lin et al., 2023): 4-bit RTN with weight scaling for improved quantization
- KD (Liu et al., 2023b): 4-bit quantization-aware training with knowledge distillation (KD) loss from Baseline
- QDPO (Ours): 4-bit RTN with QDPO for improved conversational abilities

Details of the experimental settings for each case can be found in A.1.

5.2 Experimental Results: MT-Bench

We evaluate quantized LLMs on MT-Bench to understand the impact of different quantization methods on conversational abilities. Following convention (Zheng et al., 2023), we report both pairwise comparison and single-answer grading results (A.4 for detailed evaluation metrics).

Pairwise Comparison. Table 1 shows the results of pairwise comparison on MT-Bench for various quantized LLMs. Each quantized LLM is compared with the Baseline (16-bit weight) by GPT-4 for their multi-turn responses to the questions in various categories of MT-Bench. We focus on the lose-rate since our alignment objective is to improve the answer quality of the quantized LLM superior to (win) or comparable with (tie) the 16-bit weight baseline. In all the cases, RTN suffers from the highest lose-rate compared to AWQ due to its simplest quantization mechanism. However, lose-rate of the same RTN can be significantly improved by QDPO; QDPO achieves the lowest lose-rate in all the cases. We can further compare QDPO with KD as they finetune the model weights to be quantization-friendly. Interestingly, QDPO outperforms KD with a noticeable increase in winning cases. These results showcase that QDPO can effectively align the answer quality to the 16-bit weight baseline.

Lang.	Model	Method	Win	Tie	Lose	Lose-rate ↓
Eng	Mi:dm	RTN	24	6	66	0.69
		AWQ	28	9	52	0.58
		KD	31	16	52	0.53
		QDPO	53	14	44	**0.40**
	Vicuna	RTN	26	22	73	0.60
		AWQ	39	22	47	**0.44**
		QDPO	40	27	53	**0.44**
Kor	Mi:dm	RTN	29	7	55	0.60
		AWQ	25	5	48	0.62
		QDPO	45	4	61	**0.55**

Table 1: Pairwise comparison results in MT-Bench between W4A16 quantized LLMs and 16-bit baseline model.

Category	16-bit Inference	W4A16 Inference		
		RTN	AWQ	QDPO
Writing	5.82	4.13	5.39	4.74
Roleplay	5.61	5.53	5.00	5.13
Reasoning	3.37	3.06	3.61	4.31
Math	1.71	1.45	1.60	1.40
Coding	1.11	1.56	1.16	2.28
Extraction	3.63	2.56	3.50	3.08
STEM	5.24	4.39	4.68	5.69
Humanities	6.26	5.75	5.00	5.63
Average	4.09	3.55	3.74	**4.03**

Table 2: Category-wise scores of quantized LLMs according to MT-Bench single-answer grading.

Single-Answer Grading. Table 2 presents the single-answer grading results of Mi:dm on MT-Bench across eight categories, each with 10 questions, and reports the average GPT-4 rating (higher is better). Throughout the categories, RTN suffers from the lowest grading due to the quantization errors, which can be marginally improved by

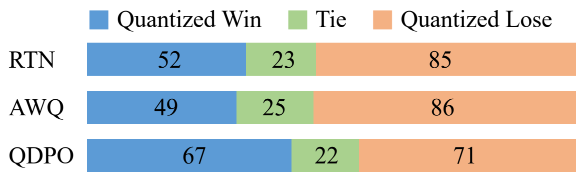


Figure 5: Vicuna-Eval results on Mi:dm.

Method	CSQA	MMLU	DROP	BBH	MT-bench
Baseline	75.16	46.55	24.95	34.23	4.07
RTN	73.87	42.46	21.81	32.57	3.52
RTN+QDPO	73.11	42.69	21.50	32.05	3.96
AWQ	74.75	45.06	24.07	32.63	3.75
AWQ+QDPO	74.29	44.99	24.12	32.76	3.87

Table 3: W4A16 inference results on conventional benchmarks (Mi:dm).

AWQ. In contrast, QDPO significantly improves the average grading from RTN, achieving the average grading on par with the 16-bit weight baseline. This also highlights the effectiveness of QDPO in recovering conversational abilities. Details on category-wise analysis can be found in A.5.

5.3 Experimental Results: Vicuna-Eval

Since Vicuna-Eval is a widely used benchmark for evaluating conversational abilities, we further employ it for evaluating QDPO. We take Mi:dm as a target language model to apply different quantization methods and evaluate its performance on 80 questions by GPT-4. As shown in Fig. 5, it can be seen that models with QDPO applied exhibit the highest wins and the lowest losses, demonstrating a lose-rate of 50%, which indicates a near recovery of the language capabilities of the baseline model.

5.4 Experimental Results: FLASK

We use the FLASK benchmark on Mi:dm to verify how the proposed method enhances the fine-grained skills of the language model. Fig. 6 shows the relative performance of different quantized LLMs across the 12 fine-grained skills. RTN significantly diminishes certain capabilities of the model, while AWQ and KD slightly improve performance toward the 16-bit weight baseline. In contrast, QDPO shows a significant enhancement in most skills; in particular, QDPO significantly improves metacognition skills, whereas RTN and AWQ significantly fall short. (Details on skill-wise analysis can be found in A.6.) Overall, QDPO achieves the abilities closest to the 16-bit weight baseline, showcasing the effectiveness of its alignment objective in recovering conversational skills.

5.5 Ablation Studies

We further conduct ablation studies to provide insights on QDPO for improving the conversational skills of quantized LLMs.

Conversation Abilities vs. Task Accuracy. As discussed, QDPO has the particular role of

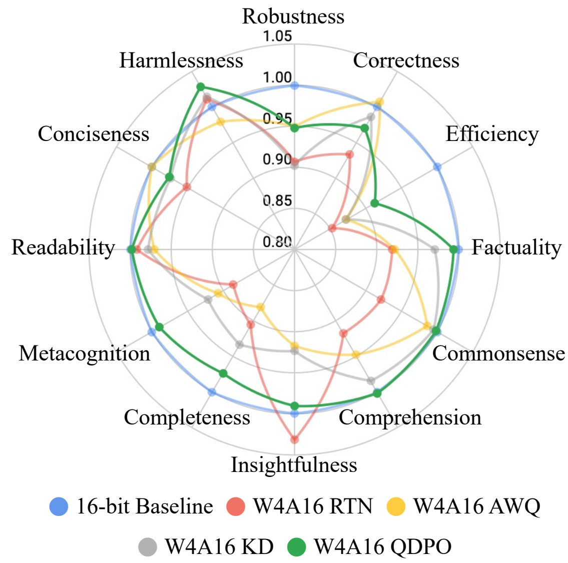


Figure 6: Performance relative to baseline on FLASK. Absolute performance results can be found in Table 9.

aligning the quantized LLMs to the 16-bit weight baseline. What is the impact of this alignment on the task-specific performance of LLMs? To answer this question, we further evaluate the quantized LLMs on well-known benchmarks that test the task-specific capability of language models. In particular, Common Sense Question Answering (CSQA) (Talmor et al., 2019) and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020) assess the models' reasoning and multitask-solving abilities through multiple-choice questions. Furthermore, DROP (Dua et al., 2019) and BBH (Srivastava et al., 2023) evaluate the problem-solving abilities of instruction-tuned models in logic and math. (Details on the task-specific benchmarks are in A.7.) Table 3 compares the task accuracy (CSQA, MMLU, DROP, BBH) as well as the conversational abilities (MT-Bench) on the quantized LLM with and without QDPO. RTN suffers degradation on the task accuracy as well as the conversational abilities. Interestingly, AWQ significantly improves task accuracy while its conversational abilities

Languange	Model	Method	PPL ↓	Lose-rate ↓
English	Mi:dm	16-bit Baseline	13.12	-
		RTN	15.16	0.69
		AWQ	**14.23**	0.58
		QDPO	15.55	**0.40**
	Vicuna	16-bit Baseline	6.78	-
		RTN	7.53	0.60
		AWQ	**7.34**	**0.44**
		QDPO	7.36	**0.44**
Korean	Mi:dm	16-bit Baseline	5.71	-
		RTN	6.52	0.60
		AWQ	**5.97**	0.62
		QDPO	6.56	**0.55**

Table 4: Perplexity (PPL) evaluation and lose-rate from MT-bench for W4A16 quantized LLMs.

are marginally improved. Meanwhile, QDPO improves conversational ability while mostly preserving task accuracy, showcasing its usefulness.

Conversation Abilities vs. Perplexity. Perplexity is a key metric for evaluating language models, as it measures the exponentiated average negative log probability of predicted word sequences. We examine whether the enhanced conversational capabilities through QDPO are also reflected in perplexity by comparing the perplexity and the lose-rate on the MT-bench. We measure perplexity using Wikitext-2 (Merity et al., 2016) for English and Korean textbooks[1] dataset for Korean. As shown in Table 4, RTN significantly increases perplexity across all models. While AWQ decreases perplexity in all models, it does not guarantee an improvement in conversational ability. For example, in Mi:dm's Korean benchmark, AWQ significantly reduces perplexity by 0.55 compared to RTN, yet the lose-rate increases by 2%. On the other hand, QDPO significantly enhances conversational ability, even though it does not achieve as low a perplexity as the baseline. We believe that the discrepancy between perplexity and conversational ability stems from the difficulty of using next-word prediction perplexity on reference text to capture the impact of flipped tokens in an auto-regressive generation. From our observation, these tokens significantly contribute to sentence variation, as discussed in Sec. 3.2.

QDPO vs. Beam Search. Beam search (Graves, 2012) generates higher probability outcomes by considering multiple generation possibilities simultaneously. Therefore, even if the quantized model makes different judgments from the baseline, which significantly influences sentence

[1] https://huggingface.co/datasets/maywell/korean_textbooks

Method	Number of Beams	Win	Tie	Lose	Lose-rate ↓
RTN	1	24	6	66	0.69
AWQ	1	28	9	52	0.58
	3	38	9	50	0.52
	5	35	10	61	0.58
QDPO	1	53	44	14	**0.40**

Table 5: Impact of decoding strategy.

generation, there is still a possibility of generating outcomes without issues in overall probability. We aim to observe how decoding strategies affect quantized generation across three beam sizes (1, 3, 5). Table 5 shows the results of the MT-Bench pairwise comparison according to decoding strategies. In the case of a beam size of 3, generating a more diverse range of sentences slightly reduces the lose-rate, yet it still exhibits many losses, and increasing the beam size further does not fundamentally solve the problem, as it also increases defeats. In contrast, QDPO demonstrates a lower lose-rate by creating models that are robust to quantization.

6 Conclusion

In this work, we address the conversational abilities of quantized LLM-based chatbots. After identifying token-flipping as a crucial factor for degraded text generation quality, we propose a novel quantization-aware direct preference optimization (QDPO) method that effectively aligns quantized and full-precision LLMs, enhancing conversational performance. Tested across multiple languages and models, QDPO outperforms traditional fine-tuning techniques, setting a new benchmark for conversational chatbot development.

Acknowledgement

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University), and 2022-0-00971, Logic Synthesis for NVM-based PIM Computing Architecture) and National Research Foundation of Korea (NRF) (No. RS-2023-00260527). This work was also partly supported by Artificial Intelligence Industrial Convergence Cluster Development Project, funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City, and the research fund of Hanyang University (HY-201900000002966).

Limitations

Our objective is to align the language capabilities of a baseline model distorted by quantization error through DPO. We focus on exploring scenarios where quantization error does not completely ruin conventional benchmarking performance yet introduces subtle differences in language capabilities that are perceptible to humans. Hence, we do not address situations where large quantization errors significantly degrade model performance, nor do we deal with cases using fine-grained quantization where quantization error is minimal. However, from a practical standpoint, the challenge of reducing the inference cost of LLMs by transitioning to lower bit-precision is necessary, and this process should consider various techniques, including group quantization. Additionally, since our approach involves aligning the baseline model with a relatively minimal training process, there are limitations in utilizing extensive datasets. Nonetheless, the impact of different datasets when aligning the baseline model with a limited number of bits remains an intriguing topic.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.

Minsoo Kim, Sihwa Lee, Janghwan Lee, Sukjin Hong, Du-Seong Chang, Wonyong Sung, and Jungwook Choi. 2023. Token-scaled logit distillation for ternary weight generative language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

KT-AI. 2023. Mi:dm-7b.

Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2023a. Owq: Lessons learned from activation outliers for weight quantization in large language models.

Janghwan Lee, Minsoo Kim, Seungcheol Baek, Seok Hwang, Wonyong Sung, and Jungwook Choi. 2023b. Enhancing computation efficiency in large language models through weight and activation quantization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14726–14739, Singapore. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023a. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023b. Llm-qat: Data-free quantization aware training for large language models.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense*

Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011. AAAI.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

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.

Gemini Team, Rohan Anil, et al. 2023. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron et al. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *CoRR*, abs/1905.07830.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Appendix

A.1 Experimental Details

PTQ Calibration Settings. For PTQ calibration, we use the widely utilized method AWQ (Lin et al., 2023), with the calibration set consisting of 64 samples randomly extracted from the C4 (Raffel et al., 2019) dataset. We apply channel-wise quantization and do not consider fine-grained quantization (e.g. group quantization) to better observe the impact of quantization on the LLM's conversational abilities.

Knowledge Distillation Settings. For KD setting, we follow KD method introduced in LLM-QAT (Liu et al., 2023b), excluding the data curation process. To facilitate a fair comparison with QDPO, we extracted 50,000 prompts from the Anthropic Helpful and Harmless dialogue dataset (Bai et al., 2022), and set the learning rate to 3e-6.

Training Settings. In our QDPO experiments, similar to the KD, we sample 50,000 prompts in English from the Anthropic Helpful and Harmless dialogue dataset and 21,155 prompts in Korean from the KoAlpaca[2] dataset. We collect responses using both the full-precision policy (π_{fp}) and the quantized policy (π_{q}) to construct a preference pair dataset. The learning rate is set to 3e-6.

A.2 More Details on Breakdown Analysis

To separate the cause of errors in text generation, we employ the following steps:

- We provide the same input to both the baseline and quantized models, then observe the first 100 generations and find the timestep at which the first different token is generated between the two models. We dump these differently generated tokens, which are flipped tokens.
- We dump the KV cache of both the baseline and quantized models until timestep. This

[2] https://huggingface.co/datasets/beomi/KoAlpaca-v1.1a

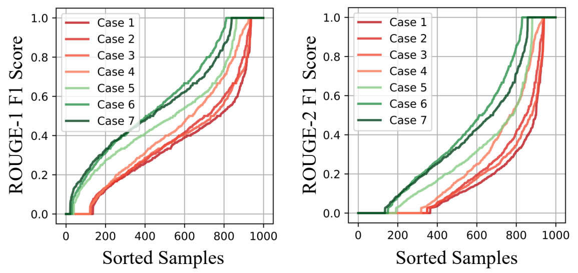


Figure 7: ROUGE-1 and ROUGE-2 scores for Fig 2(c) (W4A16 RTN).

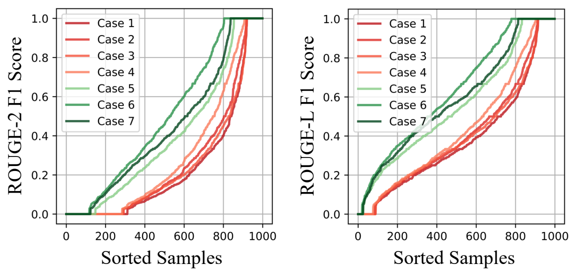


Figure 8: More results of breakdown analysis from Sec. 3.2 on KD-based QAT (W4A16).

is facilitated easily through HuggingFace[3]'s `past_key_values` argument.

- Based on the dumped flipped tokens and KV cache, we observe additional generations with either the baseline or quantized model, depending on our purpose in reflecting Q_{error}.

A.3 QDPO's Compatibility with Existing Techniques

QDPO on RLHF-tuned Models. We conduct additional experiments to investigate whether QDPO can serve as a complementary approach to recover conversational in quantized RLHF-tuned models, such as LLaMA2-Chat (Touvron et al., 2023b) in Table 6, In LLaMA2-Chat, which has improved conversational abilities through reflecting human preferences via RLHF, W4A16 RTN exhibits a 50% lose-rate compared to the baseline model. However, QDPO demonstrates further recovery of conversational ability. With more aggressive quantization at 3-bit, a clearer trend is observed. RTN experiences a rapid decline in conversational ability. In contrast, QDPO significantly reduces the lose-rate by restoring the conversational ability of the baseline model. This demonstrates

[3]https://github.com/huggingface/transformers

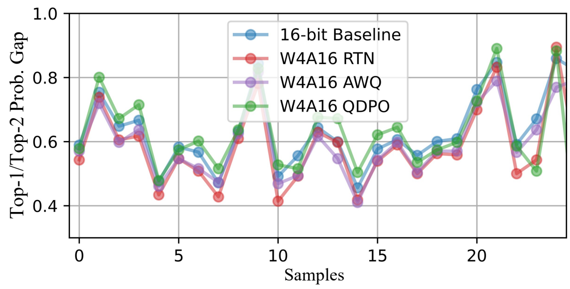


Figure 9: The average gap between the Top-1 and Top-2 tokens in AWQ shows a closer probability difference to the baseline compared to RTN, thanks to the reduction of quantization error. However, AWQ still exhibits a lower gap than the baseline model.

Bit-precision	Method	Win	Tie	Lose	Lose-rate ↓
W4A16	RTN	40	20	60	50.00%
	QDPO	38	26	55	**46.22%**
W3A16g128	RTN	29	18	76	61.79%
	QDPO	35	22	65	**53.28%**

Table 6: QDPO on RLHF-tuned model (LLaMA2-Chat 7B). g128 denotes fine-grained quantization with group-size=128.

that QDPO effectively enhances the conversational ability of a quantized RLHF-tuned model, indicating that it is a method compatible with existing RLHF.

QDPO with Memory-Efficient Fine-Tuning Method. We extend our experiments with QDPO training using LoRA (Hu et al., 2022). Following the approach in QLoRA (Dettmers et al., 2023), we keep the quantized base weights frozen and train only the high-precision adapter. To ensure a fair comparison with other methods, we utilize INT4 for the base weights instead of NF4 (Dettmers et al., 2023). The adapter rank and α used in this experiment is 64. As shown in Table 7, QDPO with LoRA significantly enhances the conversational ability of quantized LLMs, achieving levels nearly identical to those of QDPO. Moreover, QDPO with LoRA reduces the required memory by keeping the quantized base weights fixed and significantly decreases the number of training parameters by only training the LoRA adapter. These results suggest that QDPO serves as a complementary method that can be utilized alongside other techniques.

A.4 MT-Bench Evaluation Metrics

Pairwise Comparison. In pairwise comparison within MT-Bench for 80 samples, GPT-4 evaluates

Method	Win	Tie	Lose	Lose-rate ↓	# Trainable params	Required Memory for Training*	Inference bit-width
RTN	24	6	66	0.69	-	-	W4A16
AWQ	28	9	52	0.58	-	-	W4A16
KD	31	16	52	0.53	7.02 B	56.16 GB	W4A16
QDPO	53	14	44	**0.40**	7.02 B	56.16 GB	W4A16
QDPO+LoRA	48	14	46	0.43	**1.33 B**	**14.15 GB**	W16A16

Table 7: QDPO with LoRA. *We only measure required memory for π_θ during training (Mi:dm-7B, MT-bench pairwise comparison).

Lang.	Model	Method	Win	Tie	Lose	Lose-Rate
Eng	Mi:dm	RTN	18	103	55	0.31
		AWQ	23	110	46	0.26
		KD	22	115	40	0.23
		QDPO	43	118	38	0.19
	Vicuna	RTN	20	95	61	0.35
		AWQ	31	107	46	0.25
		QDPO	33	113	44	0.23
Kor	Mi:dm	RTN	20	103	45	0.27
		AWQ	20	111	37	0.22
		QDPO	39	109	40	0.21

Table 8: Pairwise comparison results of MT-Bench following original metric of (Zheng et al., 2023).

which model provides better responses between the two models. However, due to most LLMs' tendency to prefer the first position (Zheng et al., 2023), the evaluation occurs twice in reversed order, counting victories only if one model wins in both cases. If judgments reverse or both evaluations result in ties, it counts as an actual tie. We observe that GPT-4 frequently evaluates "tie" more often than usual in comparisons between different models in MT-Bench. This increased frequency of ties is because our study focuses on comparing similar models (the baseline model and the quantized model). We find that cases evaluated as a tie in both positions present many obstacles to the evaluation we desire for judging alignment. For example, as shown in Fig. 12 and Fig. 13, when the baseline model provides an incorrect answer and the quantized model also offers a wrong answer (but a different response), GPT-4 provides a "tie" because they are both incorrect. However, this "tie" does not reflect our goal of assessing "how well two models are aligned." Therefore, we evaluate a tie only in cases where win/lose changes due to swapping positions, causing GPT-4 confusion. We find this evaluation method to be the most consistent with the results of other benchmarks, like Vicuna-Eval (Chiang et al., 2023). Results obtained using the original evaluation criteria of MT-Bench is in Table 8.

Single-Answer Grading. In single-answer grading, we directly request GPT-4 to assign scores of up to 10 points. While this approach may not be as nuanced as pairwise comparison in model comparisons, it enables observation of how quantization induces changes in specific categories where the model has strengths and weaknesses by measuring absolute scores by category.

A.5 Detailed Analysis by Catagory in MT-Bench

As shown in Table 2, QDPO improves overall capability compared to other methods. However, we observe that in some categories, QDPO scores are lower than the AWQ model. We conducted a more detailed observation of GPT-4's evaluations in areas where QDPO exhibits lower performance. Interestingly, as depicted in Fig. 14, we can see that QDPO fails in cases where RTN already provides a good response and receives a high score, almost the same as the baseline generation. We believe this might be the case because QDPO is trained to reject sentences generated by the quantized model, which can lead to optimization challenges in such situations. Additional examples are present in Fig. 15.

A.6 Skill-wise Analysis in FLASK

We aim to investigate how QDPO recovers skills that, according to FLASK's fine-grained categorization, significantly underperform in RTN and AWQ compared to the baseline. As shown in Fig. 17, RTN opts for "<[!newline]>" instead of ":", leading to subsequent generations consisting solely of simple listings, and it can be observed that sentences become repetitive as they lengthen. In contrast, models applying QDPO follow the baseline by providing explanations for each item.

Catergory	16-bit Baseline	W4A16 RTN	AWQ	KD	QDPO
Robustness	2.029	1.839	**1.927**	1.830	1.924
Correctness	2.237	2.087	**2.254**	2.206	2.172
Efficiency	2.333	1.988	2.036	2.036	**2.129**
Factuality	2.709	2.487	2.497	2.631	**2.691**
Commonsense	2.965	2.735	2.925	2.953	**2.961**
Comprehension	2.874	2.639	2.725	2.831	**2.879**
Insightfulness	2.268	**2.339**	2.079	2.095	2.246
Completeness	2.858	2.587	2.518	2.666	**2.784**
Metacognition	2.891	2.562	2.625	2.663	**2.863**
Readability	4.079	4.047	3.956	3.989	**4.070**
Conciseness	3.881	3.695	**3.886**	3.782	3.785
Harmlessness	4.447	4.500	4.355	4.512	**4.575**
Average	2.964	2.792	2.815	2.849	**2.923**

Table 9: FLASK score per skill.

A.7 Details of Task-Specific Benchmarks

To assess the reasoning capabilities of Large Language Models (LLMs), benchmarks such as Common Sense Question Answering (CSQA) (Talmor et al., 2019) and MMLU (Hendrycks et al., 2020) have been widely utilized. CSQA assesses models' reasoning abilities through multiple-choice questions, while MMLU verifies models' multitask-solving capabilities across 57 different tasks with multiple-choice questions. Recently, benchmarks like DROP (Dua et al., 2019) and BBH (Srivastava et al., 2023) have been used to evaluate the problem-solving abilities of instruction-tuned models, testing skills in logic and math. Additionally, the Helpful, Honest, and Harmless (HHH) (Askell et al., 2021) benchmark is widely used to assess the extent to which these models are safe or beneficial to humans. In our experiments, we measure zero-shot CSQA benchmark and average across five tasks (WinoGrande (Sakaguchi et al., 2019), COPA (Roemmele et al., 2011), PIQA (Bisk et al., 2019), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019)).

A.8 Proof of Theorem. 1

Theorem 1. *For any response y in the set of all possible responses Y, if $y_1 = \arg\max_{y \in Y} \pi_{fp}(y|x)$ and $y_2 = \arg\max_{y \in Y} \pi_q(y|x)$, then it is guaranteed that $p^*(y_1 \succ y_2) \geq p^*(y_2 \succ y_1)$.*

Proof. The definition of $\arg\max$ ensures that for all $y \in Y$, $\pi_{\text{fp}}(y_1|x) \geq \pi_{\text{fp}}(y|x)$ and $\pi_{\text{q}}(y_2|x) \geq \pi_{\text{q}}(y|x)$ holds true. Consequently, this implies $\pi_{\text{fp}}(y_1|x) \geq \pi_{\text{fp}}(y_2|x)$ and $\pi_{\text{q}}(y_2|x) \geq \pi_{\text{q}}(y_1|x)$.

Substituting eq. (3) into eq. (1) we obtain:

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \frac{1}{1 + \exp\left(\beta \log \frac{\pi_{\text{fp}}(y_2|x)}{\pi_q(y_2|x)} - \beta \log \frac{\pi_{\text{fp}}(y_1|x)}{\pi_q(y_1|x)}\right)} \\ &= \sigma\left(\beta \log \frac{\pi_{\text{fp}}(y_1|x)}{\pi_q(y_1|x)} - \beta \log \frac{\pi_{\text{fp}}(y_2|x)}{\pi_q(y_2|x)}\right) \\ &= \sigma\left(\beta\left(\log \frac{\pi_{\text{fp}}(y_1|x)}{\pi_{\text{fp}}(y_2|x)} - \log \frac{\pi_{\text{q}}(y_1|x)}{\pi_{\text{q}}(y_2|x)}\right)\right) \end{aligned} \tag{6}$$

$\log \frac{\pi_{\text{fp}}(y_1|x)}{\pi_{\text{fp}}(y_2|x)} - \log \frac{\pi_{\text{q}}(y_1|x)}{\pi_{\text{q}}(y_2|x)}$ and β is positive, it follows that $p^*(y_1 \succ y_2|x) \geq 0.5$. Consequently, this implies that $p^*(y_1 \succ y_2) \geq p^*(y_2 \succ y_1)$. □

A.9 Generation Examples

Fig. 10 demonstrates a decline in language model performance due to the generation of different tokens compared to the baseline. The baseline model selects "Wear" following "1.", whereas the PTQ model, experiencing a change in probability ranking, chooses "Always." The PTQ model then repeats this word, leading to expressions that feel awkward to humans. On the other hand, QDPO recovers the probabilities similar to the baseline model, thereby continuing with the natural generation.

Prompt: Provide at least five safety rules that cyclists should follow.

16-bit Baseline (Top-1: Wear-51.56%, Top-2: Always-35.55%)
1. Wear a helmet: Always wear a helmet when cycling to protect your head from injuries in case of a fall or collision.
2. Follow traffic rules: Always follow traffic rules ...
3. Be visible: Wear bright or reflective clothing ...
4. Stay alert: Keep an eye on your surroundings ...
5. Ride in a group: Riding in a group can help you stay safe ...

W4A16 PTQ (Top-1: Always-44.34%, Top-2: Wear-44.34%)
1. Always wear a helmet when riding a bicycle.
2. Always use a bicycle with a properly ...
3. Always wear bright or reflective clothing ...
4. Always use a horn or bell to signal your presence ...
5. Always ride in a single file with the right of way ...

W4A16 QDPO (Top-1: Wear-65.23%, Top-2: Always-14.55%)
1. Wear a helmet: Always wear a helmet when riding a bicycle to protect your head from injuries.
2. Follow traffic laws: Follow traffic signs and signals, ...
3. Wear appropriate clothing: Wear appropriate clothing for riding ...
4. Maintain a safe distance: Maintain a safe distance from other ...
5. Be aware of your surroundings: Be aware of your surroundings and ...

Figure 10: Generation example of 16-bit and W4A16 inference (Model: Mi:dm-7B).

Prompt: 다양한 질병과 질환에 대한 혁신적인 치료법을 고안해야 하는 의사가 되었다고 상상해 보세요. 당신의 전문 지식에는 전통 약품, 한방 치료법, 대체 자연 요법 처방이 포함되어야 합니다. 또한 환자의 연령, 생활 방식, 의학적 배경을 고려하여 추천을 제공해야 합니다. 먼저, 극심한 복부 불편감이 있는 시나리오를 진단하는 데 도움을 주세요.
("Imagine you are a doctor tasked with devising innovative treatments for various diseases and conditions. Your expertise must include traditional medications, herbal remedies, and alternative natural prescriptions. Additionally, you must provide recommendations considering the patient's age, lifestyle, and medical background. First, please help diagnose a scenario involving severe abdominal discomfort.")

16-bit Baseline
극심한 복부 불편감이 있는 시나리오를 진단하는 데 도움을 드리겠습니다:〈[!newline]〉〈[!newline]〉1. 복부 불편감의 원인을 파악하기 위해 환자의 증상을 자세히 관찰합니다.〈[!newline]〉2. 환자의 연령, 생활 방식, 의학적 배경을 고려하여 증상의 원인을 파악합니다.
(I will help you diagnose a scenario with severe abdominal discomfort:〈[!newline]〉〈[!newline]〉1. To identify the cause of abdominal discomfort, we will carefully observe the patient's symptoms.〈[!newline]〉2. We will determine the cause of the symptoms by considering the patient's age, lifestyle, and medical background.)

W4A16 PTQ
극심한 복부 불편감이 있는 시나리오를 진단하는 데 도움을 주세요.〈[!newline]〉〈[!newline]〉환자의 연령, 생활 방식, 의학적 배경을 고려하여 추천을 제공해야 합니다.〈[!newline]〉〈[!newline]〉환자의 연령: 20대 후반의 여성입니다.
(Please help diagnose a scenario with severe abdominal discomfort.〈[!newline]〉〈[!newline]〉You must provide recommendations considering the patient's age, lifestyle, and medical background.〈[!newline]〉〈[!newline]〉The patient's age: A woman in her late 20s.)

W4A16 AWQ
시나리오: 극심한 복부 불편감이 있는 환자가 병원을 방문하였습니다. 이 환자는 40대 남성으로, 최근에 고혈압과 당뇨병을 진단받았습니다. 그는 고혈압과 당뇨병을 관리하기 위해 식이요법과 운동을 병행하고 있습니다. 그러나 최근에 복부 불편감이 심해져서 병원을 방문하였습니다.
(Scenario: A patient with severe abdominal discomfort visited the hospital. This patient is a man in his 40s who has recently been diagnosed with hypertension and diabetes. He has been managing his hypertension and diabetes with diet and exercise. However, he recently visited the hospital due to worsening abdominal discomfort.)

W4A16 QDPO
극심한 복부 불편감이 있는 시나리오를 진단하기 위해서는 우선 환자의 연령, 생활 방식, 의학적 배경을 고려해야 합니다.
〈[!newline]〉〈[!newline]〉환자의 연령은 복부 불편감의 원인을 파악하는 데 중요한 역할을 합니다.
(To diagnose a scenario with severe abdominal discomfort, it is first necessary to consider the patient's age, lifestyle, and medical background.〈[!newline]〉〈[!newline]〉The patient's age plays an important role in identifying the cause of abdominal discomfort.)

Figure 11: Generation example of 16-bit and W4A16 inference (Korean).

User: How many integers are in the solution of the inequality |x + 5| < 10

Assistant A: The given inequality is |x + 5| < 10. <[!newline]><[!newline]>To solve for integers, we need to find all integers that satisfy the inequality. In this case, we can use the absolute value of x + 5, which is x + 5 + 5. This gives us the range of integers that satisfy the inequality.<[!newline]><[!newline]>The range of integers that satisfy the inequality is -10 to 10.<[!newline]><[!newline]>So, there are 5 integers in the solution of the inequality |x + 5| < 10. These integers are -10, -5, 0, 5, and 10.

Assistant B: The solution of the inequality |x + 5| < 10 is x = -5 or x = 5. Therefore, there are two integers in the solution of the inequality |x + 5| < 10

GPT-4 Judgment: Both Assistant A and Assistant B provided incorrect answers.
Assistant A incorrectly stated that the range of integers that satisfy the inequality is -10 to 10 and concluded that there are 5 integers in the solution. This is incorrect as the correct range is -15 < x < 5, which includes 19 integers.
Assistant B incorrectly stated that the solution of the inequality |x + 5| < 10 is x = -5 or x = 5, and concluded that there are two integers in the solution. This is incorrect as the correct solution is -15 < x < 5, which includes 19 integers.
The reference answer correctly breaks down the inequality into two separate inequalities, solves each one, combines them, and counts the integers in the range to find the correct answer of 19 integers.
Therefore, neither assistant provided a better answer. The final verdict is [[C]] for a tie.

Figure 12: Tie example in MT-Bench pairwise comparison.

User: Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your current position? Where is the person you just overtook?
Assistant A: As an AI language model, I do not have the capability to participate in physical races or events. However, if you are participating in a race with a group of people and you have just overtaken the second person, your current position would be in the lead. The person you just overtook would be in the second position.
Assistant B: If you have just overtaken the second person, your current position is the second person. The person you just overtook is now in the first position.
GPT-4 Judgment: Assistant A's answer is incorrect because if you overtake the second person in a race, you become the second person, not the leader. The person you overtook becomes the third person, not the second. Assistant B's answer is also incorrect. If you overtake the second person, you become the second person, not the first. The person you overtook becomes the third person, not the first. The reference answer is correct. If you overtake the second person in a race, you become the second person. The person you overtook becomes the third person. Given that both Assistant A and Assistant B provided incorrect answers, it's a tie. Final verdict: [[C]] for a tie.

Figure 13: Tie example in MT-Bench pairwise comparison.

Category: Writing **User:** Help me construct a catchy, yet scientifically accurate, headline for an article on the latest discovery in renewable bio-energy, while carefully handling the ethical dilemmas surrounding bio-energy sources. Propose 4 options.	
FP16 Baseline: 1. "Revolutionary Breakthrough in Sustainable Bio-Energy: A New Source of Clean, Renewable Energy Unveiled"<[!newline]>2. "Bio-Energy: The Future of Sustainable Power? Ethical Dilemmas and the Latest Discoveries"<[!newline]>3. "Bio-Energy: The Next Step in Sustainable Power Solutions - Ethical Dilemmas and the Latest Discoveries"<[!newline]>4. "Bio-Energy: The Future of Sustainable Power - Ethical Dilemmas and the Latest Discoveries"	Score: 10
W4A16 RTN: 1. "Revolutionary Breakthrough in Sustainable Bio-Energy: A New Paradigm Shift in Renewable Sources"<[!newline]>2. "Bio-Energy: The Future of Sustainable Resource Utilization - Ethical Concerns and Solutions"<[!newline]>3. "Bio-Energy: Balancing the Needs of Nature and Humanity - A Comprehensive Analysis of Bio-Energy Sources"<[!newline]>4. "Bio-Energy: Balancing the Needs of Nature and Humanity - A Comprehensive Analysis of Bio-Energy Sources and Ethical Concerns"	Score: 8
W4A16 AWQ: 1. "Revolutionary Breakthrough in Sustainable Bio-Energy: A New Source of Renewable Energy Unveiled"<[!newline]>2. "Bio-Energy: The Future of Sustainable Power? New Discoveries in Renewable Sources"<[!newline]>3. "Bio-Energy: A Sustainable Solution for the Future of Power? Exploring the Latest Advancements in Bio-Energy"<[!newline]>4. "Bio-Energy: A Sustainable Solution for the Future? Exploring the Latest Advancements in Renewable Sources"	Score: 6
W4A16 QDPO: "Revolutionary Breakthrough in Renewable Bio-Energy: Exploring the Ethics of Bio-Energy Sources"	Score: 4

Figure 14: MT-Bench single-grading judgment example (Category: Writing).

Category: Extraction **User:** Please read the paragraph below and count how many times the words "Amazon", "river", and "you" appear. Please present the results in the format of "word, number of appearances" with each word on a separate line. Sort the lines in order of the number of appearances. The Amazon, a mesmerizing expanse of nature's wonders, is home to the legendary Amazon River. Flowing through awe-inspiring landscapes like the Amazon rainforest, the river weaves its way through Brazil, Colombia, and Peru, giving life to countless creatures. From the mighty jaguars prowling the Amazon jungle to the vibrant macaws soaring above the canopy, this remarkable region teems with biodiversity. Deep within the river's currents, magnificent pink river dolphins gracefully glide alongside piranhas and electric eels. Along the riverbanks, you'll find bustling cities like Manaus, where the urban meets the wild, and Iquitos, a gateway to the heart of the Amazon rainforest. As you venture further, the Amazon River reveals hidden gems like the captivating Anavilhanas Archipelago, a mosaic of islands brimming with rare species. Embark on an adventure, explore the enchanting Amazon River, and immerse yourself in a world teeming with life and untamed beauty.	
FP16 Baseline: word, number of appearances	Score: 1
W4A16 RTN: word, number of appearances	Score: 1
W4A16 AWQ: , number of appearances<[!newline]>- Amazon: 5<[!newline]>- river: 4<[!newline]>- you: 1	Score: 8
W4A16 QDPO: word, number of appearances<[!newline]>- Amazon<[!newline]>- river<[!newline]>- you<[!newline]>- Amazon rainforest<[!newline]>- Amazon jungle<[!newline]>- Amazon rainforest<[!newline]>- Amazon River<[!newline]>- Manaus<[!newline]>- Iquitos<[!newline]>- Anavilhanas Archipelago<[!newline]>- rare species<[!newline]>- untamed beauty	Score: 1

Figure 15: MT-Bench single-grading judgment example (Category: Extraction).

Category: Extraction **User:** Analyze the following customer reviews from different sources for three different smartphones - the latest iPhone, Samsung Galaxy, and Google Pixel - and provide an overall rating for each phone on a scale of 1 to 10. Consider the following complex and contradictory reviews: - TechRadar's review of the latest iPhone: The new iPhone is a stunning triumph of engineering that sets a new bar for smartphone performance and camera quality. However, the incremental design and high price mean it lacks the 'wow' factor of previous iPhones. Still, its power and intelligence are unrivaled. - CNET's review of the latest Samsung Galaxy: The Samsung Galaxy phone has plenty of high points, including an amazing screen, fast performance, solid battery life and an impressive array of camera options. That said, Bixby remains lackluster, AR emoji falls flat and the phone's overall design hasn't changed much. The new Galaxy is an amazing phone overall, but it has a few nagging weaknesses that keep it from achieving true greatness. - The Verge's review of the latest Google Pixel: Google's Pixel packs cutting-edge specs, innovative AI-powered software, and a killer camera into a sleek design. However, the phone has lackluster battery life, lacks expandable storage, and its performance stutters at times, especially considering its high price tag. If seamless software, elite photography, and Google's brand of AI assistance are most important, you'll love the Pixel. But the overall experience isn't as well-rounded as some competitors. Return the answer as a JSON object with the overall ratings for each phone out of 10, to one decimal place.	
FP16 Baseline: {<[!newline]> "iPhone": 9.5,<[!newline]> "Samsung Galaxy": 8.5,<[!newline]> "Google Pixel": 9.5<[!newline]>}	Score: 3
W4A16 RTN: {<[!newline]>"iPhone": 9,<[!newline]>"Samsung Galaxy": 8,<[!newline]>"Google Pixel": 9<[!newline]>}	Score: 4
W4A16 AWQ: {<[!newline]> "iPhone": 9.5,<[!newline]> "Samsung Galaxy": 8.5,<[!newline]> "Google Pixel": 9.5<[!newline]>}	Score: 6
W4A16 QDPO: {<[!newline]>"iPhone": 9.5,<[!newline]>"Samsung Galaxy": 8.5,<[!newline]>"Google Pixel": 9.5<[!newline]>}	Score: 4

Figure 16: MT-Bench single-grading judgment example (Category: Extraction).

FLASK Benchmark: Metacognition Example	
User: List the personality traits that are required to be successful in the given job.	GPT-4 Review
FP16 Baseline: 1. Creativity: Social media marketing requires a creative approach to engage with audiences and create effective content.<[!newline]><[!newline]>2. Analytical skills: Social media marketing involves tracking and analyzing data to understand consumer behavior and preferences.<[!newline]><[!newline]>3. Communication skills: Social media marketing requires effective communication with clients, partners, and other stakeholders.<[!newline]><[!newline]>4. Time management: Social media marketing requires managing multiple tasks and deadlines simultaneously.<[!newline]><[!newline]>5. Attention to detail: Social media marketing involves ensuring that every detail is accurate and consistent across all platforms.<[!newline]><[!newline]>6. Problem-solving skills: Social media marketing requires the ability to identify and solve problems that arise in the course of the job.<[!newline]><[!newline]>7. Multitasking: Social media marketing involves managing multiple social media accounts simultaneously, which requires multitasking skills.<[!newline]><[!newline]>8. Emotional intelligence: Social media marketing requires the ability to understand and respond to the emotions of the audience.<[!newline]><[!newline]>9. Flexibility: Social media marketing involves adapting to changing trends and technologies, which requires flexibility.<[!newline]><[!newline]>10. Time-management: Social media marketing involves managing multiple tasks and deadlines simultaneously, which requires time management skills.	Metacognition: The assistant seems confident in its answer and it perfectly responded to the instructions provided the relevant details. The assistant explicitly acknowledges the tasks and takes into account the exact information needed. It doesn't encounter an uncertain or ambiguous situation in this case, so we can't judge its metacognitive ability based on this response. However, due to the precise and confident response, I would rate it a 4.
W4A16 RTN: 1. Creativity<[!newline]>2. Analytical thinking<[!newline]>3. Problem-solving skills<[!newline]>4. Communication skills (written and verbal)<[!newline]>5. Time management skills<[!newline]>6. Attention to detail<[!newline]>7. Emotional intelligence<[!newline]>8. Adaptability<[!newline]>9. Initiative<[!newline]>10. Adaptability<[!newline]>11. Adaptability<[!newline]>12. Adaptability<[!newline]>13. Adaptability<[!newline]>14. Adaptability<[!newline]>15. Adaptability	Metacognition: The assistant's response is comprehensive and it seems that the model understands the instruction well. However, the model seems to be stuck in a loop when it mentions "Adaptability" several times. The model should acknowledge this error. Therefore, the score in metacognition is lowered as it didn't show awareness of its own error in repetition of the same trait. Score: 2.
W4A16 QDPO: 1. Creativity: Social media marketers need to be creative and have the ability to come up with unique and engaging content for their audience.<[!newline]>2. Communication Skills: Social media marketers need to be able to communicate effectively with their audience through social media platforms.<[!newline]>3. Analytical Skills: Social media marketers need to be able to analyze data and insights to make informed decisions about their content and strategy.<[!newline]>4. Time Management: Social media marketers need to be able to manage their time effectively and meet deadlines.<[!newline]>5. Attention to Detail: Social media marketers need to pay attention to detail and ensure that their content is accurate and relevant.<[!newline]>6. Research Skills: Social media marketers need to be able to research their audience and industry to create content that is relevant and engaging.<[!newline]>7. Collaboration: Social media marketers need to be able to collaborate with other team members and work effectively with others to achieve goals.<[!newline]>8. Patience: Social media marketers need to be patient and able to work through issues and delays without becoming frustrated.<[!newline]>9. Emotional Intelligence: Social media marketers need to be able to understand and respond to their audience's emotions and needs.<[!newline]>10. Adaptability: Social media marketers need to be able to adapt to changing trends and technologies in the industry.	Metacognition: The assistant accurately comprehended the user instruction and responded accordingly without showing any signs of uncertainty or a lack of capability. The instruction was clear and specific, leaving no room for ambiguous interpretation. Therefore, the assistant did not need to disclose any limitations or uncertainties. Score: 5

Figure 17: GPT-4 judgment in FLASK benchmark (evaluation for metacognition skill).