Conformal Feedback Alignment: Quantifying Answer-Level Reliability for Robust LLM Alignment

Anonymous ACL submission

Abstract

Preference-based alignment like Reinforcement Learning from Human Feedback (RLHF) learns from pairwise preferences, yet the labels are often noisy and inconsistent. Existing uncertainty-aware approaches weight preferences, but ignore a more fundamental factor: the reliability of the *answers* being compared. To address the problem, we propose Conformal Feedback Alignment (CFA), a framework that grounds preference weighting in the statistical guarantees of Conformal Prediction (CP). CFA quantifies answer-level reliability by constructing conformal prediction sets with controllable coverage and aggregates these reliabilities into principled weights for both DPO- and PPO-style training. Experiments across different datasets show that CFA improves alignment robustness and data efficiency, highlighting that modeling *answer-side* uncertainty complements preference-level weighting and yields more robust, data-efficient alignment.

1 Introduction

Large Language Models (LLMs) have achieved remarkable capabilities (Anil et al., 2023; Achiam et al., 2023; Touvron et al., 2023), largely driven by alignment techniques that fine-tune them on human preferences, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023). To overcome the cost and scalability limitations of human annotation, the field is increasingly adopting Reinforcement Learning from AI Feedback (RLAIF), where preference data is generated by capable AI evaluators (Yu et al., 2024; Lee et al., 2023; Li et al., 2024). While this method is powerful, it generates new challenges: preference labels can be noisy and inconsistent, which creates a fundamental noisy label problem that can degrade the quality and robustness of the final aligned model (Banerjee and Gopalan, 2024; Wang et al., 2024).

Recent studies have explored uncertainty-aware preference alignment, aiming to handle feedback by estimating uncertainty at the preference level. For example, Banerjee and Gopalan (2024) estimates reward-model uncertainty through ensemble variance, Lodkaew et al. applies weights to DPO losses, Xu et al. (2024b) introduces a Bayesian formulation that models preference uncertainty within a risk-sensitive policy framework, and WPO (Zhou et al., 2024) reweights pairs by likelihood to mitigate off-policy drift. All these approaches reduce the effect of noisy preference pairs by modeling uncertainty on preferences, i.e., how reliable the preference relationship is.

However, these methods overlook a more fundamental dimension of uncertainty: the intrinsic reliability of the individual answers that constitute the preference pair. A preference judgment can only be as trustworthy as the responses it compares. When both answers are of low confidence and quality, the preference carries little meaningful information, regardless of how we model uncertainty for the preference itself. This dimension of answer-level reliability represents a critical blind spot in current research, as it addresses the quality of the data at its very source. Therefore, we argue that each model-generated answer carries its own reliability, which directly contributes to the uncertainty observed in preference comparisons.

To address this distinct dimension of answer-level reliability, we introduce Conformal Feedback Alignment (`CFA`), a novel framework that grounds preference weighting in the statistical guarantees of Conformal Prediction (CP) (Quach et al., 2023; Su et al., 2024). Instead of focusing on the downstream learning process, CFA directly assesses the quality of the answer itself. The core of our method is to first use CP to construct statistically valid prediction sets for responses with a pre-defined and controllable coverage. A higher-coverage prediction set is considered more reliable than a lower-

coverage set. We then introduce a set-wise uncertainty aggregation function that translates the set memberships of responses into a single, principled weight that quantifies the trustworthiness of the preference label. This weight is integrated into both PPO-style and DPO-style alignment, forcing the model to prioritize judgments built upon high-confidence answers. Overall, our primary contributions are:

- We identify and address a critical, under-explored dimension in uncertainty-aware alignment: the intrinsic reliability of individual answers, which is distinct from uncertainty measured at the preference level.
- To the best of our knowledge, `CFA` is the first to construct statistically valid prediction sets with Conformal Prediction for AI feedback, which is flexible for both black- and white-box settings as well as DPO- and PPO-style model training.
- Through comprehensive experiments, we demonstrate that by focusing on answer reliability, CFA improves alignment performance and data efficiency, outperforming standard baselines and offering comparable performance over methods that address preference-level uncertainty.

2 Related Works

2.1 LLM Alignment

LLMs have achieved strong performance on a wide range of tasks (Touvron et al., 2023; Grattafiori et al., 2024; Wang et al., 2023; Team et al., 2024; Achiam et al., 2023; Abdin et al., 2024; Guo et al., 2025). A key driver of this success is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), which aligns model outputs with human preferences. RLHF is traditionally implemented as a **reward-model-based method**, typically using Proximal Policy Optimization (PPO) (Schulman et al., 2017). This approach requires training both a reward model and a policy model. To simplify this process, several alternatives have been proposed. REINFORCE++ (Hu, 2025) eliminate the need for a critic model.

Beyond reward-based methods, **reward-model-free method** Direct Preference Optimization (DPO) (Rafailov et al., 2023) directly optimizes LLM parameters based on pairwise preferences, bypassing reward model training. Extensions such as IPO (Azar et al., 2024), α-DPO (Wu et al., 2024), CPO (Xu et al., 2024a), TPO (Saeidi et al., 2024), and KTO (Ethayarajh et al., 2024) improve DPO's stability, adaptiveness, or data efficiency.

To address the cost of human labels, recent work has explored Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022). AI-generated preference data can reduce labeling cost and even improve alignment performance (Lee et al., 2023; Li et al., 2024; Williams, 2024), while AI-generated preferences might be more noisy. To address this problem, several methods that provide weight to the preference pair are proposed (Ye et al., 2025; Lin et al., 2023; Xu et al., 2024b; Wang et al., 2024). However, all of them are focusing on the reliability of the preference without considering the reliability of responses.

2.2 Uncertainty Quantification

Uncertainty quantification (UQ) has been studied in LLMs (Lin et al., 2023; Chen et al., 2025; Liu et al., 2025), but most prior work focuses on token-level or output-level generation uncertainty. For example, Kadavath et al. (2022) and Band et al. (2024) show that LLMs often produce poorly calibrated self-assessments. Conformal Prediction (CP) (Shafer and Vovk, 2008) provides a non-parametric, distribution-free approach for constructing confidence sets with formal guarantees. Recent extensions adapt CP to LLMs in both white-box settings (Quach et al., 2023) and black-box API scenarios (Su et al., 2024). However, existing UQ, including CP methods, are rarely integrated into preference learning or alignment pipelines.

3 Preliminaries

3.1 Preference-Based Alignment

Reinforcement learning from human feedback (RLHF) and its variants optimize large language models (LLMs) by learning from preference comparisons. Given a prompt x, two outputs y^+ (preferred) and y^- (dispreferred), the dataset is denoted as: $\mathcal{D}_{\text{pref}} = (x, y^+, y^-)$. Two main approaches exist for preference-based alignment:

Reward-Model-Based Methods (PPO-style). RLHF typically trains a reward model $r_\phi(x, y)$ using pairwise comparisons. The standard loss is:

$$\mathcal{L}_{\text{RM-Orig}}(\phi) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{pref}}} \log \sigma\left(r_\phi(x, y^+) - r_\phi(x, y^-)\right), \quad (1)$$

where $\sigma(\cdot)$ is the sigmoid function.

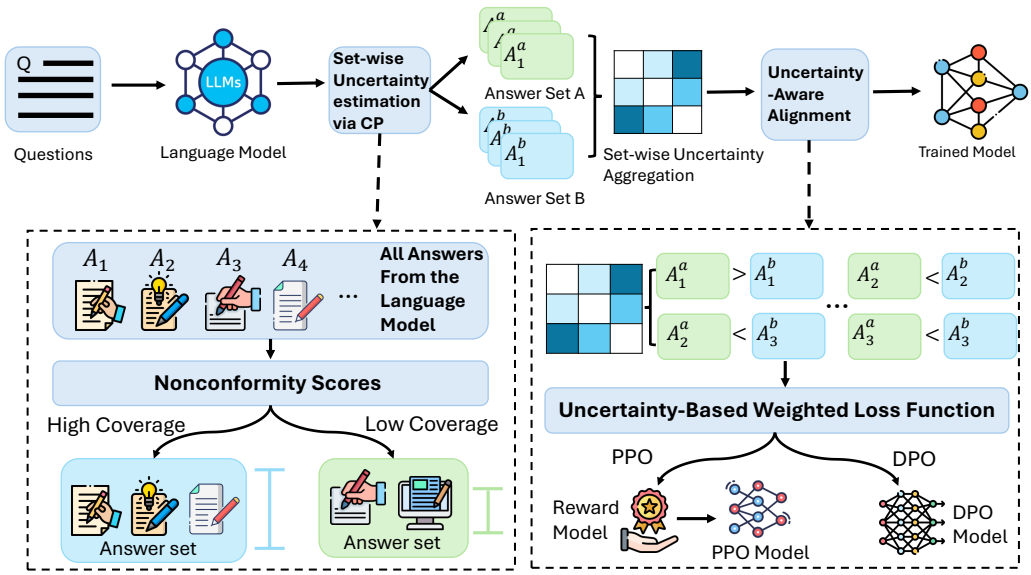

Figure 1: The pipeline of our method. In our paper, we use conformal prediction, which can be applied for both black-box and white-box settings, to estimate the uncertainty and then use a weighted loss to conduct alignment.

After training r_ϕ, a policy π_θ is optimized using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with $r_\phi(x, y)$ as the reward.

Reward-Free Methods (DPO-style). Direct Preference Optimization (DPO) (Rafailov et al., 2023) directly optimizes the policy π_θ without training a reward model. The loss is:

$$\mathcal{L}_{\text{DPO-Orig}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{pref}}} \log \sigma\left(\beta \cdot \Delta_{\theta,\text{ref}}\right), \tag{2}$$

where β is a temperature parameter, and:

$$\Delta_{\theta,\text{ref}} = \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}. \tag{3}$$

Here, π_{ref} is a frozen reference policy, typically the supervised fine-tuned model.

3.2 Conformal Prediction

Uncertainty aims to provide reliable confidence in prediction results (Kadavath et al., 2022; Band et al., 2024). Conformal Prediction (CP) is a distribution-free method that provides statistically valid uncertainty estimates. The core idea of CP is to compute a nonconformity score $s(x, y)$ for each output y given a prompt x. This score quantifies how "unusual" or "atypical" the output is. Lower scores indicate that the output is typical or confident, while higher scores suggest uncertainty.

Given a calibration set with known outputs, nonconformity scores are computed for each example, and a quantile threshold q_α is determined corresponding to the desired coverage level $1 - \alpha$. The conformal set is:

$$\mathcal{C}_{1-\alpha}(x) = \{y \mid s(x, y) \leq q_\alpha\}. \tag{4}$$

This set is guaranteed to include a correct output with probability at least $1 - \alpha$.

• **White-Box CP:** In the white-box setting, token-level log-probabilities are available. Therefore, the nonconformity score can be defined as the negative log-likelihood of the output (Quach et al., 2023):

$$s(x, y) = -\log p_\theta(y|x). \tag{5}$$

This choice of scoring function has a straightforward interpretation. If the model assigns high probability to y given x, then $s(x, y)$ will be small, indicating that the model is confident in this output. Conversely, if the model assigns low probability to y, the score will be large, reflecting higher uncertainty.

• **Black-Box CP:** In the black-box setting, where token-level probabilities are unavailable, the nonconformity score is estimated from the multiple model-generated samples and is given by (Su et al., 2024):

$$s(x, y) = -\text{Freq}(y) + \lambda_1 \cdot \text{NE}(x) - \lambda_2 \cdot \text{Sim}(y, y_{\text{top}}), \tag{6}$$

where $\text{Freq}(y)$ is the count of y among the sampled responses for prompt x, $\text{NE}(x)$ is the normalized entropy of the sampled response distribution, $\text{Sim}(y, y_{\text{top}})$ is the similarity to the most

frequent sample y_{top}, and λ_1, λ_2 are weighting coefficients. It suggests that responses that occur more frequently and are more similar to the most frequent sample yield lower nonconformity scores, indicating higher confidence and lower uncertainty.

4 Method

This section presents Conformal Feedback Alignment (CFA). Our goal is to improve the robustness of alignment using the answer reliability from CP.

4.1 Overview

Figure 1 presents an overview of our framework, which has two components: (1) estimating the uncertainty of AI-generated answers, which is considered as the answer reliability, using CP and (2) incorporating this uncertainty into both reward-model-based (e.g., PPO) and reward-free (e.g., DPO) alignment. For each preference pair, CFA produces an uncertainty score $u \in [0, 1]$ from the answer reliability, which is used to weight the learning signal during policy optimization, prioritizing reliable feedback upon high-reliability answers while reducing the effect of comparisons for less reliable answers. The next sections describe set-wise uncertainty estimation (Section 4.2) and its integration into alignment (Section 4.3).

4.2 Set-wise Uncertainty Estimation via CP

To account for the reliability of responses, a principled approach to uncertainty estimation is essential. However, LLMs are known to produce uncalibrated uncertainty scores (Band et al., 2024). We address this by employing CP (Section 3.2) to generate statistically valid prediction sets. After employing CP, we can obtain two sets A and B with different coverage α. The properties of these sets form the basis for our **Set-wise Uncertainty Aggregation** method, which translates answer-level reliability into a weight for the preference pair.

For a given preference pair (y^+, y^-), we define a set-wise uncertainty score $u(x, y^+, y^-)$ based on the conformal sets to which the two responses belong. If both y^+ and y^- belong to the set A, we use the corresponding quantile q_a as the confidence. If they both belong to set B, we use q_b. If the two outputs belong to different sets, we take the average of the two quantiles. Formally, we define:

$$u(x, y^+, y^-) = \begin{cases} q_a & \text{if both } y^+, y^- \in \text{set A}, \\ q_b & \text{if both } y^+, y^- \in \text{set B}, \\ \frac{q_a + q_b}{2} & \text{if } y^+, y^- \text{ are from different sets.} \end{cases} \tag{7}$$

This design reflects the overall reliability of the comparison based on the reliability of the answers. When the outputs come from uncertain sets or the sets are mismatched, the confidence is reduced because the answers are not reliable. This set-wise approach allows us to calibrate the preference signal based on the uncertainty in the model's predictions.

4.3 Uncertainty-Aware Alignment

Standard alignment objectives, as used in PPO and DPO (Section 3), assume all comparisons are equally reliable. To address this limitation, we use uncertainty scores to adjust the training process. We introduce an uncertainty-weighted loss function, where each preference pair's contribution is scaled by its estimated reliability. This ensures that more trustworthy comparisons have a greater influence on alignment.

Reward-Model-Based Optimization (PPO-Style) In the standard PPO-style approach, a reward model is trained by optimizing the preference loss shown in Eq. 1. We adapt this objective to be uncertainty-aware by introducing an uncertainty weight u for each comparison:

$$\begin{aligned} &\mathcal{L}_{\text{RM-Uncert}}(\phi) \\ &= -\mathbb{E}_{(x,y^+,y^-,u)} u \cdot \log \sigma \left(r_\phi(x, y^+) - r_\phi(x, y^-) \right) \end{aligned} \tag{8}$$

Reward-Free Optimization (DPO-Style) For reward-free learning, we build on the DPO objective defined in Eq. 2. DPO optimizes the policy directly by maximizing the log odds of the preferred output relative to the reference policy. To account for uncertainty, we introduce the uncertainty weight u from Set-wise Uncertainty Aggregation into this objective, multiplying the log-sigmoid term by u:

$$\mathcal{L}_{\text{DPO-Uncert}}(\theta) = -\mathbb{E}_{(x,y^+,y^-,u)} u \cdot \log \sigma \left(\beta \cdot \Delta_{\theta,\text{ref}} \right). \tag{9}$$

The strategy of our method is to introduce an uncertainty-weighted loss, which reframes preference comparisons from 'hard labels' to 'soft evidence' following Zhou et al. (2024). This allows the learning process to prioritize preference with reliable answers. We apply this principle to both PPO and DPO-style optimization. In the PPO-style approach, the weights are used during reward model training to produce a more reliable reward signal $r_\phi(x, y)$. In the DPO-style approach, the weights are applied directly to the preference loss to modulate gradient updates. While the mechanisms differ, both methods ensure the final policy is shaped

Algorithm 1 Conformal Feedback Alignment (CFA)

Require: Initial policy π_θ, reference policy π_{ref}

for each training iteration **do**
  Collect preference pairs (x, y^+, y^-) using AI or human feedback
  for each pair (x, y^+, y^-) **do**
    Estimate nonconformity scores $s(x, y^+), s(x, y^-)$ via CP
    Compute set-wise uncertainty score $u(x, y^+, y^-)$ according to Eq.(7)
  end for
  Form augmented dataset $\mathcal{D}_{\text{uncert}} = (x, y^+, y^-, u)$
  if PPO-style **then**
    Train reward model r_ϕ with Eq.(8)
    Fine-tune policy π_θ using PPO with r_ϕ
  else
    Update policy π_θ with Eq.(9)
  end if
end for

more significantly by high-confidence data for better alignment.

4.4 Algorithm Description

We provide a summary of the procedure in Algorithm 1. This algorithm describes the complete pipeline of CFA, covering both the uncertainty estimation and the policy optimization stages. Our algorithm begins by initializing the policy π_θ and reference π_{ref}, and calibrating nonconformity score thresholds. Each iteration involves two stages. First, we collect preference data (x, y^+, y^-) and estimate an uncertainty score u for each sample via CP (Section 4.2), yielding an uncertainty-augmented dataset $\mathcal{D}_{\text{uncert}}$. Second, we update the policy π_θ using this dataset. This can be done with uncertainty-aware alignment introduced in Section 4.3 with PPO-style or DPO-style training. This algorithm allows the policy to focus on preference comparisons that are built from reliable answers, while reducing the influence of noisy or ambiguous feedback. By integrating CP into the learning loop, the method calibrates the learning signal dynamically and adapts to the uncertainty present in the data.

5 Experiments

In this section, we mainly conduct experiments to answer the following research questions:

- **RQ1:** Does the proposed method outperform normal alignment or post-training methods overall?
- **RQ2:** How Does CFA perform under different scales of model parameters and data?
- **RQ3:** How does CFA perform against preference-level uncertainty-aware methods, and how is it affected by white-box versus black-box settings?

5.1 Experimental Setup

Models For a comprehensive evaluation of the performance of our proposed method, we use in total three different open-source models as the pre-trained model. In detail, we use Llama2-7b (Touvron et al., 2023) and Llama3.1-8b (Grattafiori et al., 2024) to show that our method works for different versions of the popular open-source models. We also use Qwen series, including Qwen2.5-7B (Yang et al., 2024) and Qwen3 series (Yang et al., 2025) as the pre-trained model to show that our method could be generalized to different model architectures. For all models, we use a standard framework that first does a supervised fine-tuning and then does alignment.

Dataset In this paper, we use three datasets:

- Webgpt Comparisons (Webgpt) (Nakano et al., 2021): A dataset of question-answering outputs from the WebGPT model, used to evaluate long-form answers grounded in web search results.
- Synthetic Instruct GPT-J Pairwise (Pairwise) (Alex et al., 2021): A general-purpose instruction following dataset for alignment via synthetic prompts and responses in a wide range of tasks.
- Summarize from Feedback (Summarize) (Stiennon et al., 2020): A summarization dataset targeting content compression, focusing on Reddit posts and learning from preference-based feedback.

For each dataset, our method will generate new answers using the question in the original dataset with conformal prediction for each model. Table 2 summarizes the detailed sizes of the augmented dataset for each model to be evaluated.

Evaluation Following previous work (Zhang et al., 2024), all outputs from trained models are evaluated by LLM-as-a-judge, which we use GPT-4o (Achiam et al., 2023). This methodology aligns with various studies (Gilardi et al., 2023; Alizadeh et al., 2023) highlighting the capabilities of LLMs to produce high-quality text evaluation that aligns with or surpasses human. Specifically, four different dimensions: (1) *Accuracy*, which assesses

Dataset	Summarize			Pairwise			WebGPT		
Models	SFT	Base	CFA	SFT	Base	CFA	SFT	Base	CFA
PPO									
Llama2-7B	63.91	65.88	**67.39 (+1.51)**	89.32	91.15	**91.89 (+0.74)**	69.54	72.25	**72.78 (+0.53)**
Llama3.1-8B	56.18	56.93	**57.59 (+0.66)**	79.22	81.38	**82.25 (+0.87)**	72.85	75.02	**75.56 (+0.54)**
Qwen2.5-7B	52.42	54.29	**55.21 (+0.92)**	88.25	90.17	**90.65 (+0.48)**	73.37	75.64	**75.95 (+0.31)**
DPO									
Llama2-7B	63.91	65.68	**67.30 (+1.62)**	89.32	90.88	**92.12 (+1.24)**	69.54	71.68	**72.25 (+0.57)**
Llama3.1-8B	56.18	56.7	**57.44 (+0.74)**	79.22	81.55	**82.90 (+1.35)**	72.85	74.70	**75.19 (+0.49)**
Qwen2.5-7B	52.42	53.85	**55.01 (+1.16)**	88.25	89.82	**90.97 (+1.15)**	73.37	75.91	**76.42 (+0.51)**

Table 1: Performance comparison of our methods with baselines across various datasets and models using LLM-as-a-Judge. The average scores are reported. The higher the score, the better the performance. SFT is supervised fine-tuning only without any preference alignment. Base is the normal PPO or DPO alignment without uncertainty. The **best** result is shown. The numbers in parentheses indicate improvements over the base alignment method. CFA outperforms base methods across different models and datasets consistently.

	Webgpt	Pairwise	Summarize
Llama2-7B	53199	68067	104614
Llama3.1-8B	55465	62101	101191
Qwen2.5-7B	52101	68029	103845

Table 2: Augmented dataset size for each model.

whether the content of the answer or summary correctly reflects the information and intent of the original prompt. (2) *Relevance*, which checks whether the answer or summary closely aligns with the subject of the prompt. (3) *Completeness*, which evaluates whether the response includes all essential points and details from the prompt. (4) *Expression*, which considers whether the language used in the answer or summary is clear and easy to understand. Each criterion is scored up to 100 in increments of 5. Due to page limits, we only report the average of these four scores in the main results. The evaluation prompts can be found in the Appendix.

Implementations Due to hardware constraints, for all experiments, our experiment applies a batch size of 1. We use the coverage $\alpha = 0.8$ and $\alpha = 0.5$ in the main results. We set the maximum generation length for both the CP process and test process after training to 1024. We use the AdamW optimizer (Zhuang et al., 2022) and use the learning rate $1e-6$. For black-box CP settings, we use a lower temperature = 0.15 since we need repeated answers. For white-box CP settings, we use a default temperature = 0.7 as in previous work (Quach et al., 2023). For all datasets, we regenerate the samples according to the instructions in the data and use GPT-4 (Achiam et al., 2023) and Alpacafarm (Dubois et al., 2023) to obtain AI feedback on preference. For the CP calibration set, we randomly sample 100 samples from each dataset as the calibration set and ensure there is no overlap between the calibration set and the training set. All of experiments on done using 4 Nvidia-A100 GPUs.

5.2 Comparison with Baselines (RQ1)

To evaluate the effectiveness of our method, CFA, we compare it against several standard alignment baselines: Supervised Fine-tuning (SFT), PPO, and DPO. The evaluation is conducted on three distinct language models across three benchmark datasets. Following the experimental setup detailed in Section 5.1, we use GPT-4o for automated evaluation. The main results are presented in Table 1, from which we draw the following observations:

- Our method consistently achieves superior performance across all evaluated scenarios. For instance, on the Llama2-7B model with the Summarize dataset, the improvement of CFA over the base model is +1.62, which is comparable to the +1.77 gain achieved by the base model over SFT.
- Considering the performance gain using DPO and PPO, we observe that our method combined with DPO-style training yields greater improvements. One potential reason is that our method is used to train the reward model for PPO-style training. However, in PPO, the reward model is only one component, and the overall training process is more complex and harder to optimize effectively.

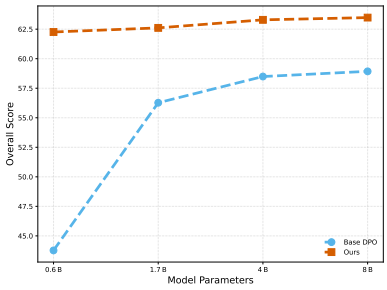

Figure 2: Performance comparison on Llama2-7B and the Summarize dataset with different sizes of model parameters.

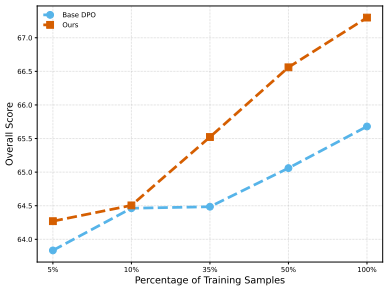

Figure 3: Performance comparison on Llama2-7B and the Summarize dataset with different sizes of training samples.

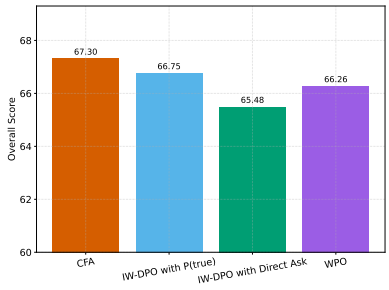

Figure 4: Performance comparison on Llama2-7B and the Summarize dataset using CFA and other methods.

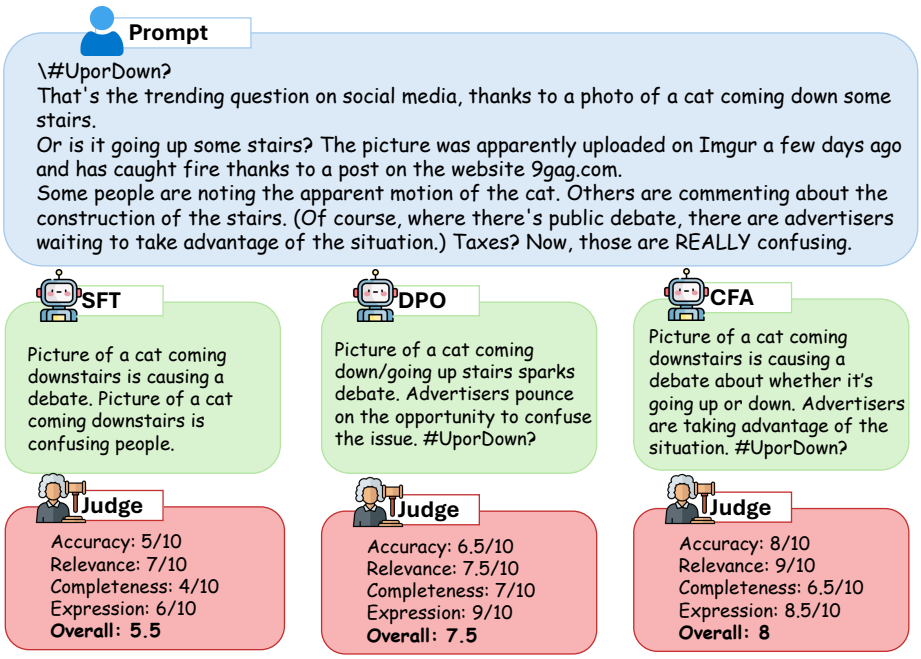

Figure 5: A case study on the Summarization dataset compared with the summarization output from the model with SFT, the model with Base DPO, and the model with our methods. The output results and scores show a clear advantage of our method.

5.3 Scalability and Efficiency Analysis (RQ2)

To evaluate the scalability and data efficiency of our method, we examine its performance across different model sizes and varying proportions of training data. This helps assess its practical applicability under diverse resource and data constraints. All experiments use black-box CP.

Influence of Different Model Parameters To understand how different sizes of model parameters influence performance, we train Qwen3 series (Yang et al., 2025), which contains models from 0.6B to 8B, allowing for comprehensive analysis. In detail, we conduct CFA with DPO on four different sizes of models from the Qwen3 series and the Summarize dataset (Stiennon et al., 2020). In Figure 2, we show the overall score for four different models. The results show that CFA works stably for different sizes of models, and the performance gain is even larger for the smaller model, which shows the effectiveness of CFA.

Influence of Different Sizes of Training Samples To understand how different sizes of training samples, we train Llama2-7B with 5%, 10%, 20%, 35%, 50% of the data in the Summarize dataset. The results can be found in Figure 3. We could see that for different proportions of the training data usage, CFA consistently outperforms the base DPO even when the data proportion is as low as 5%. And with the increase in training samples, CFA shows a

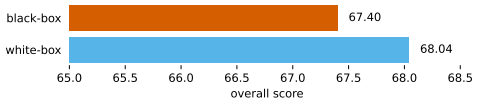

Figure 6: The comparison using white-box and black-box CP on Llama2-7B and Summarize Dataset. The result shows that using white-box CP could lead to an even better performance.

higher performance gain, which demonstrates the scalability of CFA.

5.4 Ablation Study and Sensitivity Analysis (RQ3)

In this section, we are going to answer How does CFA perform against preference-level uncertainty-aware methods and how different settings of conformal prediction influence the performance. More experiments about the settings of conformal prediction can be found in the Appendix.

Performance for preference-level uncertainty-aware methods To demonstrate the effectiveness of using answer-level reliability, we compare CFA with WPO (Zhou et al., 2024) as well as IW-DPO (Lodkaew et al.) with the following uncertainty signals as the weight:

• *Direct Ask*: When obtaining preferences of answers, directly ask LLMs to output the confidence of this preference.

• *p(true)* (Kadavath et al., 2022): When obtaining preferences of answers, using the output probability of the chosen preference as the confidence.

Entropy

We train the Llama2-7B on the Summarize dataset with DPO, and the detailed comparison can be found at Figure 4. The results show that CFA outperforms IW-DPO using different UQ methods and WPO. For IW-DPO using *Direct Ask* as the weight, it even shows a worse performance than the base DPO, showing the unreliability of uncertainty from LLMs themselves.

White-box v.s. Black-box Previously, we used black-box CP to ensure the generalization ability of CFA. Here, to see how white-box CP influences the performance, we show a performance comparison using white-box and black-box CP on Llama2-7B with the Summarize dataset. The results in Figure 6 show that using white-box CP, which allows for generating diverse answers, could lead to an even better performance, showing the potential of CFA.

5.5 Case Analysis

To better understand the effect of uncertainty-aware preference modeling, we examine a representative example from the Summarize dataset in Figure 5. The prompt involves summarizing an optical illusion debate featuring a photo of a cat. From the case, we can see that the SFT model's output is highly repetitive and uninformative, a fact reflected in its low overall score of 5.5 out of 10, which demonstrates why we need reinforcement learning.

The Base DPO model attempts to solve the problem from SFT model by incorporating more detail but misinterprets a critical part of the prompt. Specifically, it states that "advertisers pounce on the opportunity to confuse the issue," incorrectly implying that advertisers are actively creating confusion, which is wrong. This type of subtle misinterpretation is a key problem when a model learns from all feedback indiscriminately. It treats all feedback as equally valid, which prevents it from capturing the true, nuanced relationship described in the source text.

Our uncertainty-aware method excels because it corrects this specific failure. It achieved the highest scores for accuracy and relevance, 8 and 9, respectively, because it learns to disregard such noisy, misleading signals. By focusing on high-confidence feedback, it develops a more precise understanding, correctly identifying that advertisers are "taking advantage of the situation." This demonstrates its ability to filter out ambiguity and generate a summary that is not only fluent but also factually accurate and contextually relevant.

6 Conclusion

In this paper, we presented a new framework for Conformal Feedback Alignment (CFA), which improves the alignment of large language models by using reliability of responses. Our method provides both reward-model-based and reward-free training paradigms under white-box and black-box settings. Empirical results show that incorporating confidence from reliability from answers into preference learning not only enhances alignment quality but also scales well with limited data and smaller models. Overall, this work highlights the importance of reliability of answers in alignment pipelines and provides a practical supported solution. Future research may explore extending this framework to multimodal feedback, or human-in-the-loop calibration under real-world constraints.

Limitation

While Conformal Feedback Alignment (CFA) demonstrates promising improvements in robustness for LLM alignment, several limitations remain. First, CFA relies on the quality of conformal prediction calibration. When the calibration set is small, the resulting reliability estimates may be inaccurate. Second, the current experiments are conducted on text-based models and dataset, lacking the exploration for the multi-modal large language model. Finally, our approach does not combine the answer reliability and preference uncertainty. Future work should investigate how to jointly model them to further enhance alignment performance. We use LLM for grammar check only.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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*.

Beatrice Alex, Clare Llewellyn, Pawel Orzechowski, and Maria Boutchkova. 2021. The online pivot: Lessons learned from teaching a text and data mining course in lockdown, enhancing online teaching with pair programming and digital badges. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 138–148.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 101.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. *arXiv preprint arXiv:2404.00474*.

Debangshu Banerjee and Aditya Gopalan. 2024. Towards reliable alignment: Uncertainty-aware rlhf. *arXiv preprint arXiv:2410.23726*.

Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Vagelis Papalexakis, and Hua Wei. 2025. Uncertainty quantification of large language models through multi-dimensional responses. *arXiv preprint arXiv:2502.16820*.

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

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, and 1 others. 2024. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *arXiv preprint arXiv:2403.08309*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*.

Thanawat Lodkaew, Tongtong Fang, Takashi Ishida, and Masashi Sugiyama. Importance weighting for aligning language models under deployment distribution shift. *Transactions on Machine Learning Research*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

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

Amir Saeidi, Shivanshu Verma, Aswin RRV, and Chitta Baral. 2024. Triple preference optimization: Achieving better alignment with less data in a single step optimization. *arXiv preprint arXiv:2405.16681*.

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

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

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

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yikun Wang, Rui Zheng, Liang Ding, Qi Zhang, Dahua Lin, and Dacheng Tao. 2024. Uncertainty aware learning for language model alignment. *arXiv preprint arXiv:2406.04854*.

Marcus Williams. 2024. Multi-objective reinforcement learning from ai feedback. *arXiv preprint arXiv:2406.07295*.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. $alpha$-dpo: Adaptive reward margin is what direct preference optimization needs. *arXiv preprint arXiv:2410.10148*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Sheng Xu, Bo Yue, Hongyuan Zha, and Guiliang Liu. 2024b. Uncertainty-aware preference alignment in reinforcement learning from human feedback. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others.

2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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*.

Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, and Chengchun Shi. 2025. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Jinghan Zhang, Xiting Wang, Yiqiao Jin, Changyu Chen, Xinhao Zhang, and Kunpeng Liu. 2024. Prototypical reward network for data-efficient rlhf. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13871–13884.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*.

Zhenxun Zhuang, Mingrui Liu, Ashok Cutkosky, and Francesco Orabona. 2022. Understanding adamw through proximal methods and scale-freeness. *arXiv preprint arXiv:2202.00089*.

A Code

Our sample code and dataset can be found at *https://anonymous.4open.science/r/RLUF-5D7D*.

B Evaluation Prompt

As said in Section 5.1, we show our evaluation prompt for GPT-4o in Figure 7. The evaluation code with the same evaluation prompt can be found in our code.

Prompt for Evaluation

You are an expert evaluator. You are given an original input (question or source text) and an AI-generated response (answer or summary). Your task is to evaluate the response based on four criteria: Accuracy, Relevance, Completeness, and Expression. Each criterion should be scored from 0 to 10 in increments of 0.5. Provide a brief justification for each score. Then, calculate the average of the four scores and present it as the Overall Score.

Scoring Criteria:

1. Accuracy (Acc): Does the response accurately reflect the content and intent of the original prompt?
2. Relevance (Rel): Is the response closely aligned with the topic and requirements of the prompt?
3. Completeness (Comp): Does the response address all essential aspects or key points in the prompt?
4. Expression (Expr): Is the response clear, well-written, and easy to understand?

Please *only* return the four line-scores and the Overall Score, in this exact format:

```
**Accuracy (Acc):** [score]/10
**Relevance (Rel):** [score]/10
**Completeness (Comp):** [score]/10
**Expression (Expr):** [score]/10
**Overall Score:** [average]/10
```

Figure 7: Prompt for Evaluation in our Test Stage.

C More experiments

In this section, we present the results that better demonstrate the effectiveness of our method. In detail, all the experiments conducted in this section are using a black-boc CP, which is the same as the main paper.

Win Rate Comparisons To better understand the win rates between the responses of models trained with `CFA` and normal DPO. Same with the main text, we are using GPT-4o to judge which answer is

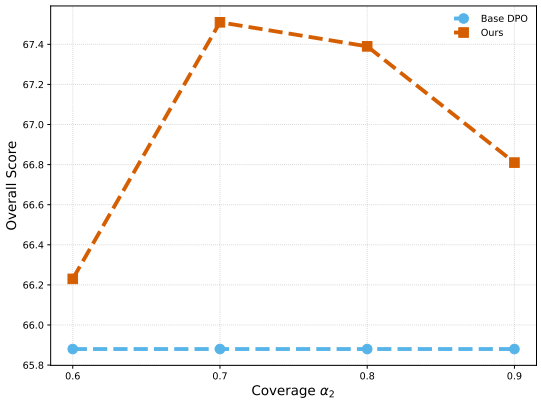

Figure 8: The comparison using different coverage α_2 in CFA on Llama2-7B and Summarize Dataset. The result shows there are some sweet points for the Coverage α_2, but CFA consistently outperforms the base DPO.

better. The result can be found in Table 3. Results show a clear advantage over CFA.

WinRate	Llama2	Llama3.1	Qwen2.5
CFA v.s Base_DPO	56.18%	64.33%	57.61%

Table 3: The Win rates compared with CFA with Base DPO on three models on the Summarize dataset. The results clearly show that CFA has an advantage.

Different Coverage in CP In the main experiments, we are using coverage $\alpha_1 = 0.5$ and $\alpha_2 = 0.8$. One advantage of using Conformal Prediction is that users can customize the coverage in the CP process. Therefore, in this section, we change the different α_2 to see how CFA performs. In detail, we change α_2 from 0.6 to 0.9 with Llama2-7B and the Summarize Dataset. The results are shown in Figure 8. The results show that using a coverage of 0.7-0.8 can have the best overall performance. When the α_2 is 0.6, we get the worst performance because α_2 is too close to $\alpha_1 = 0.5$, resulting in the reduction of the effect of weights in CFA. However, CFA consistently outperforms the base DPO, showing CFA is better for alignment of LLMs.