Investigating How Entropy-Based Control Affects Token-Level Multi-Objective Alignment

Anonymous ACL submission

Abstract

Token-level multi-objective alignment holds potential for mitigating objective conflicts, yet it remains constrained by the acquisition of feasible and reliable token-level supervisory signals for multiple objectives. As an easily obtainable yet controversial token-level signal, the effects of entropy on models in token-level multi-objective alignment have not been fully explored. In this work, we propose an entropy-based control framework to investigate the effects of entropy on token-level multi-objective alignment. The framework incorporates two hyperparameters, representing three dimensions related to entropy. We find that different models exhibit similar or divergent preferences in various entropy-related dimensions, precluding a one-size-fits-all conclusion. Under preference-aligned settings, entropy-based control can positively contribute to token-level multi-objective alignment, achieving simultaneous improvement across multiple objectives.

1 Introduction

The evaluation of responses from large language models (LLMs) typically depends on a combination of objectives, such as instruction-following, helpfulness, truthfulness, honesty, etc. There are two ways to consider multi-objective alignment in LLM response quality scoring datasets: (1) explicitly consider multiple dimensions, score each dimension separately, and a weighted average comprehensive score can be obtained (Wang et al., 2024c); (2) directly provide a comprehensive score that implicitly considers multiple dimensions (Wang et al., 2025b). Whether considering multiple objectives explicitly or implicitly, the supervisory signals ultimately used to train LLMs are usually sample-level scores.

When training LLMs using sample-level overall scores as supervision signals, multi-objective alignment inevitably leads to objective conflicts (Li et al., 2025b). Sample-level scoring may conflict with a single objective and be applied equally to each token, and cannot be adjusted based on a token's impact on the objective. A potential approach to mitigate this conflict is to use token-level supervision signals (Zeng et al., 2024; Abdin et al., 2024) to assign different optimization directions to different objectives for each token. However, the primary obstacle to implementing such token-level multi-objective alignment lies in the difficulty of obtaining feasible and reliable token-level supervision signals for multiple objectives.

Among the available token-level supervision signals, entropy is a frequently considered option due to its ease of acquisition and its ability to characterize uncertainty to a certain extent (Prabhudesai et al., 2025; Agarwal et al., 2025). However, different studies propose divergent perspectives on the application patterns of entropy at the token level, such as whether higher weights should be assigned to high-entropy or low-entropy tokens (Wang et al., 2025a; Yang et al., 2025). In token-level multi-objective alignment, how entropy affects different models remains an unexplored question. In this work, we introduce entropy into the token-level multi-objective alignment process and explore the effect of entropy-based control on alignment. Our work is organized as the following outline:

- We demonstrate the objective conflict issue through data statistics on UltraFeedBack, a response scoring dataset that explicitly considers multiple objectives. Based on this dataset, we conduct preliminary experiments using overall scoring and single-objective scoring as supervisory signals and demonstrate the impact of objective conflicts on LLM training.
- Based on preliminary experiments, we propose an entropy-based control framework that enables investigation of the effects of entropy as a token-level supervisory signal in token-level multi-

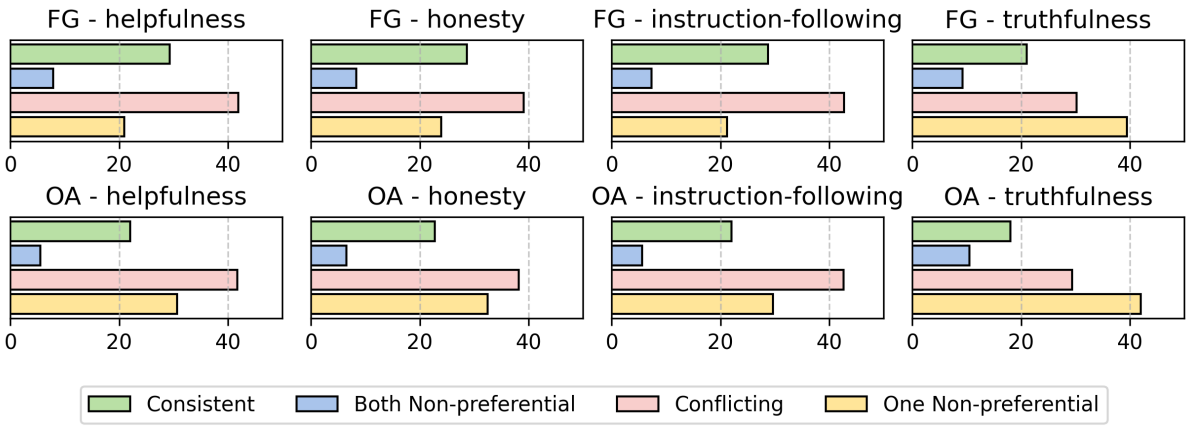

Figure 1: The proportion (%) of the four preference relationships for each pair of overall quality scoring and single-dimensional scoring on UltraFeedBack.

objective alignment scenarios. This framework incorporates two hyperparameters representing three dimensions related to entropy.

- We conduct experiments on various models. The results indicate that different models exhibit similar or divergent preferences regarding various dimensions of entropy, precluding a one-size-fits-all conclusion. Under hyperparameter selections aligned with preferences, entropy-based control can positively contribute to token-level multi-objective alignment, achieving simultaneous improvements across multiple objectives.

2 Preliminary Experiments

2.1 Training Dataset

The dataset we used for training is UltraFeedBack (Cui et al., 2023), a large-scale, multi-objective response quality scoring dataset. UltraFeedBack contains 63,967 prompts from diverse resources and four different responses for each prompt. For each response, UltraFeedBack provides scores for four objectives: instruction-following, helpfulness, truthfulness, and honesty (see Appendix A for details). In addition, UltraFeedBack provides two types of overall quality scoring for each response:

- **Fine-grained score (FG).** The average score of attributes in four objectives, explicitly considering multiple objectives.

- **Overall score (OA).** The score given for the overall quality of the response, implicitly considering multiple objectives.

We conduct statistics on UltraFeedBack to demonstrate potential objective conflict issues. For two responses A and B under the same prompt, assume that the overall quality scores are $q_A,\ q_B$. We define $q = q_A - q_B$ as the preference provided by the overall quality scores: $q > 0$ indicates a preference for A, $q < 0$ indicates a preference for B, and $q = 0$ indicates no preference. Similarly, we have single-objective scores $s_A,\ s_B$ and the preference $s = s_A - s_B$. There are four possible relationships between preferences provided by overall quality scores and single-objective scores: consistent ($q \cdot s > 0$), both non-preferential ($q = 0, s = 0$), conflicting ($q \cdot s < 0$), and one non-preferential ($q \cdot s = 0, q + s \neq 0$). For each pair of overall quality scoring and single-objective scoring, we calculate the proportions of the four types of relationships, and the statistics are shown in Figure 1. We find that conflicting relationships are common in any pair, accounting for approximately 30% to 40% of cases. In addition, one non-preferential relationship with potential inconsistencies also accounts for more than 20% cases in any pair.

2.2 Evaluation Benchmarks

We select the following datasets to evaluate the multiple objectives and comprehensive capabilities of the trained LLMs:

- **TruthfulQA.** TruthfulQA (Lin et al., 2022) is a benchmark to measure whether an LLM is truthful in generating answers to questions. The benchmark comprises 817 questions that span 38 categories, including health, law, finance and politics. We report MC1 (accuracy when there is a single true answer) and MC2 (normalized

	TruthfulQA		IFEval		BeHonest		BBH	M-P
	MC1	MC2	PSA	ISA	RR	SKR		
Qwen2.5-7B-Instruct	47.49	62.93	70.98	78.90	36.48	30.77	53.62	42.99
Helpfulness	50.67	64.91	71.16	79.02	36.74	31.38	53.91	43.18
Honesty	52.02	65.69	72.27	79.74	41.13	31.21	53.81	43.14
Instruction-following	49.94	63.96	72.64	80.34	34.51	30.88	53.72	43.22
Truthfulness	52.14	65.77	72.46	80.22	41.68	30.36	53.93	43.08
Fine-grained score	$52.75_{(1)}$	$65.87_{(1)}$	$71.90_{(4)}$	$79.62_{(4)}$	$39.71_{(3)}$	$31.14_{(3)}$	$53.83_{(3)}$	$43.07_{(5)}$
Overall score	$51.53_{(3)}$	$65.20_{(3)}$	$70.79_{(6)}$	$78.54_{(6)}$	$39.33_{(3)}$	$30.97_{(3)}$	$54.28_{(1)}$	$43.20_{(2)}$
Llama-3.1-8B-Instruct	40.51	54.97	73.38	81.41	60.23	43.09	50.62	37.83
Helpfulness	48.84	62.53	72.09	79.74	61.44	45.66	51.05	38.72
Honesty	49.82	64.44	73.75	80.82	68.07	44.22	51.15	38.00
Instruction-following	45.53	60.42	75.42	82.37	56.86	43.66	50.84	38.54
Truthfulness	49.69	64.40	73.57	80.94	72.23	43.50	50.60	38.07
Fine-grained score	$51.41_{(1)}$	$65.10_{(1)}$	$74.31_{(2)}$	$81.41_{(2)}$	$67.29_{(3)}$	$44.68_{(2)}$	$51.35_{(1)}$	$38.22_{(3)}$
Overall score	$47.61_{(4)}$	$62.63_{(3)}$	$72.83_{(5)}$	$80.58_{(5)}$	$65.89_{(3)}$	$44.14_{(3)}$	$50.79_{(4)}$	$38.02_{(4)}$
Mistral-7B-Instruct-v0.3	50.31	66.35	48.43	59.83	37.74	36.86	44.70	30.98
Helpfulness	53.24	67.36	43.81	56.71	42.61	40.23	45.15	30.50
Honesty	57.16	69.35	46.77	58.87	43.23	37.08	44.94	31.00
Instruction-following	53.37	66.65	51.39	61.63	33.54	35.40	44.28	30.29
Truthfulness	58.02	71.53	48.98	60.43	50.65	36.12	45.41	30.97
Fine-grained score	$56.55_{(3)}$	$69.74_{(2)}$	$46.77_{(4)}$	$58.51_{(5)}$	$45.77_{(2)}$	$36.43_{(4)}$	$45.50_{(1)}$	$30.34_{(5)}$
Overall score	$56.55_{(3)}$	$68.99_{(3)}$	$48.79_{(3)}$	$60.07_{(3)}$	$44.96_{(2)}$	$36.99_{(3)}$	$46.14_{(1)}$	$30.86_{(4)}$

Table 1: The results of preliminary experiments. The rows with bold model names report the performance of the original model, and the following rows report the performance after DPO based on the original model, using the corresponding scoring as the sample pair selection criteria. For fine-grained score and overall score, the numbers in the subscripts indicate the ranking of the performance when compared with the previous five performances (including the performances of the original model and the four models trained with single-objective scoring).

total probability assigned to the set of true answers when there are multiple correct answers) on this benchmark as a measure of truthfulness of LLMs.

- **IFEval.** IFEval (Zhou et al., 2023) is a straightforward benchmark for evaluating the instruction-following ability of LLMs. The benchmark contains 25 types of verifiable instructions and around 500 constructed prompts, each containing one or more instructions. We report Strict Accuracy (the proportion of LLMs' raw outputs that follow the instructions) on this benchmark as a measure of instruction-following of LLMs, including Prompt-level Strict Accuracy (PSA) and Instruction-level Strict Accuracy (ISA).

- **BeHonest.** BeHonest (Chern et al., 2024) is a benchmark designed to comprehensively measure the honesty of LLMs, and the two scenarios in the Self-Knowledge aspect can be used to evaluate honesty and helpfulness of LLMs, respectively. Scenario 1 contains 5,284 unanswerable questions. The refusal rate (RR) of LLMs is used as a measure of honesty. Scenario 2 contains 4,579 questions. The modified self-knowledge rate (SKR) of LLMs, i.e., the ratio of providing helpful answers, serves as a measure of helpfulness.

- **BBH.** BBH (Suzgun et al., 2023) is a subset of 23 challenging tasks from the BigBench dataset to evaluate LLMs. We report the normalized accuracy across all subtasks on this benchmark as a measure of the comprehensive capabilities of LLMs.

- **MMLU-Pro.** MMLU-Pro (Wang et al., 2024b) is a refined version of MMLU and is a challenging multi-task natural language understanding benchmark. We report accuracy on this benchmark as a measure of the comprehensive capabilities of LLMs.

2.3 Settings and Results

The LLMs we experiment with include Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). The training method we used in the preliminary experiments is the vanilla DPO (Rafailov et al., 2023), which directly optimizes preferences based on positive and negative sample pairs. We use four single-objective scoring and two types of overall quality scoring as selection criteria. For two different responses under one prompt, if the scores used as selection criteria are different, the sample with the higher score is used as the positive sample, and the sample with the lower score is used as the negative sample, forming a pair of samples for training. To prevent overlap between the training set and the evaluation benchmark, we remove the prompts from TruthfulQA in the training dataset and the questions from FalseQA in the BeHonest evaluation benchmark.

The experimental results are shown in Table 1. We find that models trained using single-objective scoring (hereinafter referred to as single-objective models) may perform well in specific objectives but poorly in some others. Training models using overall quality scores does not always integrate the advantages of single-objective scoring in specific objectives, which is reflected in the fact that the single-objective or comprehensive performance of the trained models does not always rank high when compared with that of single-objective models. The results demonstrate performance issues caused by objective conflicts.

3 Formulation

In this section, we formulate the scenario of token-level multi-objective alignment and introduce an entropy-based control framework into the alignment process. This framework comprises two hyperparameters and characterizes three dimensions related to entropy, enabling the investigation of how entropy-based control affects alignment.

3.1 Token-level Multi-objective Alignment

The trained model θ is initialized as the model obtained from training with overall quality scores in preliminary experiments. θ_k $(k = 1, 2, ..., n)$ refer to the n attribute models obtained from training with single-dimensional attribute scores in preliminary experiments. For a training sample, a prompt $\boldsymbol{p}$ is selected from the same training dataset as in preliminary experiments, and a token sequence $\boldsymbol{t}$ with $|\boldsymbol{t}|$ tokens is sampled by θ based on the prompt $\boldsymbol{p}$. The training objective is to minimize the following loss function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{|\boldsymbol{t}|} \sum_{k=1}^{n} w_{k,i} \cdot \frac{1}{2} \log^2 \left(\frac{\pi_\theta(t_i|\boldsymbol{p}, \boldsymbol{t}_{<i})}{\pi_{\theta_k}(t_i|\boldsymbol{p}, \boldsymbol{t}_{<i})} \right) \quad (1)$$

The alignment is essentially an on-policy distillation (Agarwal et al., 2024) that reduces the KL divergence between the trained model and the single-objective models. The key lies in determining $w_{k,i}$, which represents the weight assigned to each token for each objective.

3.2 Entropy-based Control

The control framework is based on the entropy of the token distribution generated by the trained model and the single-objective models:

$$H_i(\pi) = -\sum_{v \in \mathcal{V}} \pi(v \mid \boldsymbol{p}, \boldsymbol{t}_{<i}) \log \pi(v \mid \boldsymbol{p}, \boldsymbol{t}_{<i}) \quad (2)$$

where $\mathcal{V}$ is the token vocabulary. We introduce two hyperparameters $\alpha \in (-\infty, +\infty)$ and $\beta \in \{1, 0, -1\}$, and define:

$$a_{k,i} = \frac{\exp(\alpha \cdot H_i(\pi_{\theta_k}))}{\sum_{j=1}^{n} \exp(\alpha \cdot H_i(\pi_{\theta_j}))} \quad (3)$$

$$b_{k,i} = \mathbb{I}(\beta \cdot H_i(\pi_{\theta_k}) \geq \beta \cdot H_i(\pi_\theta)) \quad (4)$$

where $\mathbb{I}(\text{condition})$ takes the value 1 when the condition holds, and 0 otherwise. The weight $w_{k,i}$ in Equation 1 is given by:

$$w_{k,i} = a_{k,i} \cdot b_{k,i} \quad (5)$$

This framework characterizes three dimensions related to entropy:

Single-objective direction: The application direction of entropy for single-objective models, which is characterized by the sign of α. When $\alpha < 0$, greater weights are assigned to dimensions with lower entropy, and it is assumed that dimensions with higher confidence in token selection should be assigned relatively greater weight; when $\alpha > 0$, greater weights are assigned to dimensions with higher entropy, and it is assumed that dimensions with more non-trivial token selection should receive relatively greater weight; when $\alpha = 0$, the same weights are assigned to all dimensions.

Training-objective relationship: The control arising from the relationship between the entropy

		TruthfulQA		IFEval		BeHonest		BBH	M-P
		MC1	MC2	PSA	ISA	RR	SKR		
Qwen2.5-7B-FG	Pre-Expr	52.75	65.87	71.90	79.62	39.71	31.14	53.83	43.07
	+ETA	53.24	66.33	72.83	80.42	40.10	31.80	54.29	43.31
	$\triangle_{\text{PERF}}$	+0.74	+0.58	+1.48	+1.20	+0.17	+0.55	+0.38	+0.19
	$\triangle_{\text{RANK}}$	1→1	1→1	4→1	4→1	3→3	3→1	3→1	5→1
Qwen2.5-7B-OA	Pre-Expr	51.53	65.20	70.79	78.54	39.33	30.97	54.28	43.20
	+ETA	52.51	65.82	71.90	79.14	39.67	31.64	54.23	43.24
	$\triangle_{\text{PERF}}$	+0.98	+0.62	+1.11	+0.60	+0.34	+0.67	-0.05	+0.04
	$\triangle_{\text{RANK}}$	3→1	3→1	6→4	6→4	3→3	3→1	1→1	2→1
Llama-3.1-8B-FG	Pre-Expr	51.41	65.10	74.31	81.41	67.29	44.68	51.35	38.22
	+ETA	51.53	65.13	74.49	81.49	67.47	44.75	51.35	38.31
	$\triangle_{\text{PERF}}$	+0.12	+0.03	+0.18	+0.08	+0.18	+0.07	0.00	+0.09
	$\triangle_{\text{RANK}}$	1→1	1→1	2→2	2→2	3→3	2→2	1→1	3→3
Llama-3.1-8B-OA	Pre-Expr	47.61	62.63	72.83	80.58	65.89	44.14	50.79	38.02
	+ETA	47.70	62.65	73.57	81.06	66.37	44.23	51.10	38.15
	$\triangle_{\text{PERF}}$	+0.09	+0.02	+0.74	+0.48	+0.48	+0.09	+0.31	+0.13
	$\triangle_{\text{RANK}}$	4→4	3→3	5→3	5→3	3→3	3→2	4→2	4→3
Mistral-7B-FG	Pre-Expr	56.55	69.74	46.77	58.51	45.77	36.43	45.50	30.34
	+ETA	56.71	69.91	48.49	59.95	45.77	36.45	45.66	30.45
	$\triangle_{\text{PERF}}$	+0.16	+0.17	+1.72	+1.44	0.00	+0.02	+0.16	+0.11
	$\triangle_{\text{RANK}}$	3→3	2→2	4→3	5→3	2→2	4→4	1→1	5→5
Mistral-7B-OA	Pre-Expr	56.55	68.99	48.79	60.07	44.96	36.99	46.14	30.86
	+ETA	56.59	69.05	49.54	60.43	45.12	37.21	46.16	30.94
	$\triangle_{\text{PERF}}$	+0.04	+0.06	+0.75	+0.36	+0.16	+0.22	+0.02	+0.08
	$\triangle_{\text{RANK}}$	3→3	3→3	3→2	3→2	2→2	3→2	1→1	4→4

Table 2: Performance comparison of the models in preliminary experiments (Pre-Expr) and after entropy-based controlled token-level multi-objective alignment (+ETA). FG and OA represent fine-grained score and overall score, respectively. $\triangle_{\text{PERF}}$ denotes the change in performance. $\triangle_{\text{RANK}}$ denotes the change in performance ranking (as described in Table 1). We mark improvements in performance or ranking in red and declines in performance or ranking in blue. Results underlined indicate significant performance differences (three runs, $p < 0.05$).

of the trained model and the single-objective models, which is characterized by β. When $\beta = +1$, only dimensions of attribute models with entropy greater than the training model are considered, and it is assumed that the trained model should be adjusted based on the dimensions exhibiting more non-trivial token selections; when $\beta = -1$, only dimensions of attribute models with entropy less than the trained model are considered, and it is assumed that the trained model should be adjusted based on dimensions exhibiting greater confidence in token selections; when $\beta = 0$, all dimensions are included.

Inter-objective discriminability: The degree of difference between objective weights determined by entropy, which is characterized by $|\alpha|$. The larger $|\alpha|$ is, the more pronounced the difference in weight distribution among dimensions.

4 Experiments

We experiment with the Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 models obtained from the preliminary experiments. We adopt the same evaluation settings as in the preliminary experiments. Details regarding the implementation of training and evaluation can be found in Appendix B.

In this section, we first demonstrate that under appropriate hyperparameter selection, entropy-based control enables token-level multi-objective alignment to yield benefits across multiple objectives, and we further validate the positive contribution of the control through ablation experiments. We then reveal the preferences exhibited by different models across the three dimensions of entropy within the framework.

	α	β	TruthfulQA		IFEval		BeHonest		BBH	M-P	Ar.
			MC1	MC2	PSA	ISA	RR	SKR			
Qwen2.5-7B-FG	✓	✓	**53.24**	**66.33**	**72.83**	**80.42**	**40.10**	**31.80**	**54.29**	**43.31**	**1.33**
	×	✓	52.51	65.82	70.61	78.30	38.35	31.51	54.00	43.23	2.17
	✓	×	52.51	65.65	71.90	79.26	39.83	31.36	53.95	43.21	2.33
	×	×	51.90	65.50	70.06	78.54	38.88	31.27	53.97	43.22	2.67
Qwen2.5-7B-OA	✓	✓	**52.51**	**65.82**	**71.90**	79.14	39.67	**31.64**	**54.23**	**43.24**	**1.83**
	×	✓	**52.51**	**65.82**	70.61	78.30	38.35	31.51	54.00	43.23	2.17
	✓	×	**52.51**	65.65	**71.90**	**79.26**	**39.83**	31.36	53.95	43.21	2.33
	×	×	51.90	65.50	70.06	78.54	38.88	31.27	53.97	43.22	2.67
Llama-3.1-8B-FG	✓	✓	51.53	65.13	**74.49**	**81.49**	67.47	44.75	51.35	**38.31**	**2.00**
	×	✓	**51.65**	65.08	73.75	81.05	67.33	**44.77**	51.19	38.21	2.08
	✓	×	51.53	**65.18**	73.94	81.06	67.35	44.55	51.28	**38.31**	2.08
	×	×	**51.65**	65.14	73.01	80.46	**67.65**	44.66	**51.38**	38.29	2.50
Llama-3.1-8B-OA	✓	✓	**47.70**	62.65	**73.57**	**81.06**	**66.37**	**44.23**	51.10	38.15	**2.75**
	×	✓	47.37	**62.68**	72.46	80.22	66.02	44.22	**51.26**	38.02	3.08
	✓	×	47.49	62.65	73.01	80.94	65.93	44.09	51.14	**38.21**	3.08
	×	×	47.25	62.65	71.90	80.10	66.17	44.14	50.96	38.13	3.50
Mistral-7B-FG	✓	✓	**56.71**	**69.91**	**48.49**	**59.95**	45.77	36.45	45.66	**30.45**	**2.92**
	×	✓	56.67	69.87	48.24	59.11	46.39	**36.80**	**45.96**	30.44	3.08
	✓	×	56.55	69.77	48.06	59.83	45.90	36.34	45.72	30.35	3.00
	×	×	56.67	69.85	47.50	59.59	**46.55**	36.75	45.65	30.34	3.08
Mistral-7B-OA	✓	✓	56.59	69.05	**49.54**	**60.43**	45.12	37.21	46.16	30.94	**2.33**
	×	✓	**56.79**	69.07	48.24	59.11	**45.35**	**37.26**	46.00	**30.96**	2.67
	✓	×	56.67	**69.10**	48.80	59.83	45.16	**37.26**	**46.24**	30.94	2.50
	×	×	**56.79**	68.98	48.06	59.47	45.01	37.15	46.07	30.95	2.67

Table 3: Experimental results showing the effects of introducing non-zero α and β. For columns α and β, ✓ indicates the use of non-zero α or β consistent with Table 2, and × indicates the use of $\alpha = 0$ or $\beta = 0$. Column **Ar.** refers to the average of the rankings (as described in Table 1). The rankings for (MC1, MC2) and (PSA, ISA) are averaged before calculation. In subsequent tables, the meaning of column **Ar.** remains the same.

4.1 Benefits

In Table 2, we demonstrate the performance comparison of the model before and after token-level multi-objective alignment with entropy-based control. In the results presented here, $\beta = +1$ is used for all models, $\alpha = +10$ is used for the Qwen model, and $\alpha = -10$ is used for the Llama model and the Mistral model.

In terms of performance metrics, for Qwen-FG, Llama-OA, and Mistral-OA, performance across all benchmarks improve. For Qwen-OA, Llama-FG, and Mistral-FG, despite unchanged or slightly decreased performance on one benchmark, their ranking on that benchmark remain unchanged, and performance on other benchmarks improve.

In terms of performance rankings, all models maintain or improve their rankings on single-objective and comprehensive benchmarks. It is worth noting that the following cases exist in the results: (1) the model that is previously ranked first still achieves performance improvement through training; (2) the model is ranked first after training, surpassing the single-objective model that is previously ranked first. This indicates that even without introducing information beyond single-objective models, the performance gains from token-level multi-objective alignment with entropy-based control can potentially exceed the limitations of single-objective models.

4.2 Ablation

Entropy-based control is achieved by introducing non-zero α and β. When $\alpha = 0$, the weight distribution is the same for all dimensions. When $\beta = 0$, all dimensions are included in the calculation. The results in Table 3 show the effects of introducing non-zero α and β. We find that for any model, using both non-zero α and β yields the best aver-

	α	β	TruthfulQA		IFEval		BeHonest		BBH	M-P	Ar.
			MC1	MC2	PSA	ISA	RR	SKR			
Qwen2.5-7B-FG	$+$	$+$	53.24	66.33	**72.83**	**80.42**	**40.10**	**31.80**	**54.29**	43.31	**1.33**
	$-$	$+$	**53.37**	**66.56**	71.90	80.46	39.67	31.75	53.96	**43.33**	1.58
	$+$	$-$	51.53	65.51	71.72	79.26	39.60	31.16	54.14	43.29	2.50
	$-$	$-$	51.65	65.51	70.24	78.78	39.22	30.68	53.78	43.19	3.83
Qwen2.5-7B-OA	$+$	$+$	**52.51**	**65.82**	71.90	79.14	**39.67**	**31.64**	**54.23**	**43.24**	**1.83**
	$-$	$+$	52.02	65.72	**73.20**	**80.46**	39.01	31.58	54.12	43.11	2.00
	$+$	$-$	51.29	65.07	72.46	79.98	38.07	31.14	54.04	43.09	2.75
	$-$	$-$	50.80	65.01	70.98	78.54	37.74	31.10	54.04	43.22	2.75
Llama-3.1-8B-FG	$+$	$+$	**51.65**	**65.22**	72.83	80.58	**67.59**	**44.84**	51.26	38.26	2.50
	$-$	$+$	51.53	65.13	**74.49**	**81.49**	67.47	44.75	**51.35**	38.31	**2.00**
	$+$	$-$	51.29	65.17	73.01	80.58	67.31	44.77	51.14	38.30	2.67
	$-$	$-$	51.29	**65.22**	73.38	80.82	67.48	**44.84**	51.17	**38.33**	2.33
Llama-3.1-8B-OA	$+$	$+$	47.25	62.63	73.38	**81.06**	66.16	**44.33**	51.00	38.14	3.00
	$-$	$+$	**47.70**	62.65	**73.57**	**81.06**	**66.37**	44.23	**51.10**	38.15	**2.75**
	$+$	$-$	47.49	62.63	72.46	80.58	65.95	**44.33**	51.05	38.16	3.08
	$-$	$-$	47.61	**62.67**	73.38	**81.06**	66.06	44.22	51.07	**38.18**	2.83
Mistral-7B-FG	$+$	$+$	56.67	**70.02**	46.03	58.39	**46.41**	36.47	45.51	30.34	3.25
	$-$	$+$	**56.71**	69.91	**48.49**	**59.95**	45.77	36.45	45.66	**30.45**	**2.92**
	$+$	$-$	56.67	69.82	45.84	58.51	46.03	**36.54**	**45.74**	30.37	3.25
	$-$	$-$	56.30	69.79	47.69	59.59	46.15	36.38	45.65	30.44	3.08
Mistral-7B-OA	$+$	$+$	56.55	69.00	**49.91**	**60.91**	45.09	**37.24**	46.03	30.82	**2.33**
	$-$	$+$	56.59	**69.05**	49.54	60.43	**45.12**	37.21	**46.16**	30.94	**2.33**
	$+$	$-$	**56.79**	69.01	48.80	60.07	44.82	37.08	45.98	**30.95**	2.50
	$-$	$-$	56.55	68.94	48.98	60.19	44.94	37.10	46.12	30.88	2.42

Table 4: Experimental results showing the effects of the signs of α and β. For columns α and β, $+$ indicates the use of $\alpha = +10$ or $\beta = +1$, and $-$ indicates the use of $\alpha = -10$ or $\beta = -1$.

age ranking; when introducing either non-zero α or β alone, the model's average ranking remains unchanged or improves, but never achieves a higher ranking than when both non-zero α and β are introduced simultaneously. The results of ablation experiments indicate that the control provided by both α and β contributes to the benefits.

4.3 Dimensions

We investigate the model's preferences in three dimensions related to entropy.

Single-objective direction. Table 4 shows the effects of different signs of α and β on model performance. Whether $\beta > 0$ or $\beta < 0$, the sign of α yielding better average rankings remains consistent for the same model, indicating a preference on the dimension of single-objective direction. For the Qwen model, the preference is $\alpha > 0$; for the Llama model and the Mistral model, the preference is $\alpha < 0$. On this dimension, the models exhibit distinct preferences.

Training-objective relationship. Regardless of the sign of α, the sign of β yielding better average rankings remains consistent for the same model, indicating a preference on the dimension of training-objective relationship. Unlike the single-objective direction, all models exhibit the same preference $\beta > 0$ on this dimension.

Inter-objective discriminability. Figure 2 shows the effects of $|\alpha|$ on the model's average ranking (specific performance metrics are detailed in Table 5 in Appendix C). The central $|a| = 10$ yields the best average ranking in the most cases, followed by the slightly smaller $|\alpha| = 1$. As $|\alpha|$ moves toward larger or smaller values, performance on certain benchmarks may improve, but this is typically accompanied by a decline in performance on other benchmarks, resulting in an overall decrease in the average ranking. On this dimension, the models overall exhibit a similar preference for moderate values, but differences persist in their specific optimal points.

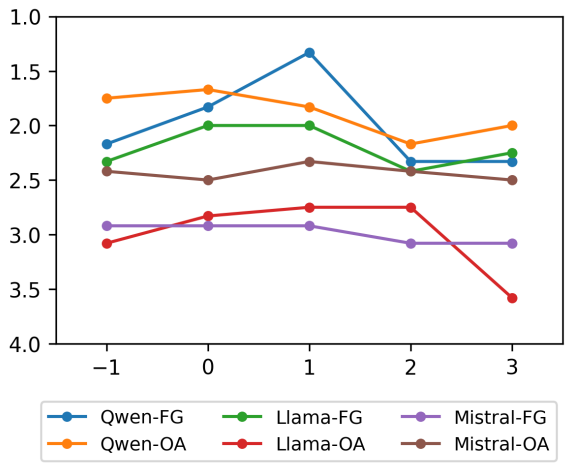

Figure 2: The curve showing the average ranking of the model (y-axis) as $|\alpha| = 10^x$ varies. The sign of α is consistent with Table 2.

5 Related Work

Multi-objective alignment. There are two main research directions for multi-objective alignment: (1) Considering two attributes with obvious conflicts (e.g., harmlessness and helpfulness). The method allows pre-setting of weights for two attributes, and the research goal is typically that the curve formed by sliding weights can reach the Pareto frontier (Gupta et al., 2025; Li et al., 2025a). (2) Considering multiple attributes that jointly determine the overall response quality. The weights for each dimension in the method are not preset but are automatically determined based on certain information, and the research goal is typically to achieve overall improvement (Wang et al., 2024a; Liu et al., 2025). For token-level multi-objective alignment, its potential for mitigating objective conflicts primarily stems from its ability to establish directionality for different tokens across various dimensions. However, if inherent conflicts exist within a token itself, token-level alignment struggles to resolve such conflicts. Consequently, research on token-level multi-objective alignment primarily focuses on the latter research direction.

Token-level supervision signal. Sample-level supervision signals are applied uniformly to each token, while token-level supervision signals are typically specialized in a finer-grained manner for each token. Since token-level supervision signals are difficult to obtain through direct annotation, such signals typically rely on some form of automatically generated credentials, such as introducing pivotal token search (Abdin et al., 2024) or forward KL divergence constraints on each token (Zeng et al., 2024) in DPO. In this work, we explore the potential of utilizing entropy as a token-level supervision signal for multi-objective alignment. Entropy possesses the characteristics of being annotation-free and readily available, while naturally exhibiting token-level granularity, thereby meeting the prerequisites for exploration.

Entropy for training. For reinforcement learning used to improve LLM reasoning, there have been some explorations using entropy minimization instead of labeled rewards (Prabhudesai et al., 2025; Agarwal et al., 2025), as entropy can to some extent characterize uncertainty. Several studies have also investigated how entropy influences token weight determination during training. Regarding whether high-entropy tokens should be assigned greater weight during training, these works present differing viewpoints (Wang et al., 2025a; Yang et al., 2025). In the scenario of token-level multi-objective alignment, entropy remains an under-explored signal. The controversy surrounding entropy as a training signal prompts us to investigate its effects in the current scenario. Although experimental results indicate that there is no one-size-fits-all preference for entropy across different models, this may still offer meaningful insights.

6 Conclusion

In this work, we investigate how entropy signals affect token-level multi-objective alignment. Our preliminary experiments validate performance degradation caused by objective conflicts. Based on this, we introduce an entropy-based control framework for token-level multi-objective alignment. This framework incorporates two parameters representing three entropy-related dimensions: single-objective direction, training-objective relationship, and inter-objective discriminability, enabling investigation of the effects of entropy on alignment. Experimental results reveal that different models exhibit similar or divergent preferences regarding these entropy-related dimensions, precluding one-size-fits-all conclusions. Under preference-aligned settings, two types of entropy-based control both benefit token-level multi-objective alignment, achieving simultaneous improvements across multiple objectives. This work offers new insights into the role of entropy as a supervisory signal and the exploration of available signals within token-level multi-objective alignment scenarios.

Limitations

Resource consumption in token-level multi-objective alignment scenarios. The token-level multi-objective alignment scenario remains an under-explored domain with no widely accepted fixed paradigm. Therefore, in this work, we draw inspiration from the concept of on-policy distillation to construct a token-level multi-objective alignment scenario utilizing multiple objective models based on preliminary experiments. A limitation of this scenario lies in resource consumption, as different objective models must be placed in GPU memory. On 8 A800 GPUs, the DPO in preliminary experiments takes an average of 4 hours, while the token-level multi-objective alignment in main experiments takes an average of 10 hours. We will explore ways to build a token-level multi-objective alignment scenario with lower resource consumption in the future.

Impact of the lack of one-size-fits-all conclusions on practical application. Our experimental results indicate that while entropy-based control can benefit token-level multi-objective alignment, different models may exhibit varying preferences for entropy-related dimensions. Consequently, in practical applications, the optimal hyperparameter selections for α and β cannot be directly determined, and we have yet to explore hyperparameter determination methods that do not require experimentation. While the lack of one-size-fits-all conclusions may impact practical applications, it offers valuable insights into the debate surrounding entropy as a controversial signal. We will explore in future work whether optimal entropy-based control parameters can be determined without experimental testing.

Ethics Statement

We comply with the license to use language models for scientific research purposes only. The datasets we use do not contain any information that names or uniquely identifies individual people or offensive content.

We use the translation tool DeepL for our writing, which may provide translations based on LLMs. Nevertheless, we meticulously review the translated content to avoid potential issues.

References

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *CoRR*, abs/2412.08905.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in LLM reasoning. *CoRR*, abs/2505.15134.

Steffi Chern, Zhulin Hu, Yuqing Yang, Ethan Chern, Yuan Guo, Jiahe Jin, Binjie Wang, and Pengfei Liu. 2024. Behonest: Benchmarking honesty of large language models. *CoRR*, abs/2406.13261.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *CoRR*, abs/2310.01377.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, and Abhinav Rastogi. 2025. Robust multi-objective preference alignment with online DPO. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 27321–27329. AAAI Press.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. 2025a. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11214–11232. Association for Computational Linguistics.

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. 2025b. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11010–11031. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Qi Liu, Jingqing Ruan, Hao Li, Haodong Zhao, Desheng Wang, Jiansong Chen, Guanglu Wan, Xunliang Cai, Zhi Zheng, and Tong Xu. 2025. Amopo: Adaptive multi-objective preference optimization without reward models and reference models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8832–8866. Association for Computational Linguistics.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *CoRR*, abs/2505.22660.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10582–10592. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2024c. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 3371–3384. Association for Computational Linguistics.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025b. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *CoRR*, abs/2505.11475.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. 2025. Do not let low-probability tokens over-dominate in RL for llms. *CoRR*, abs/2505.12929.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

A Examples on Dimensions

In this section, we introduce the concept of each dimension and provide examples of good or poor quality within each dimension. Prompts are sourced from the evaluation benchmarks, and responses originate from model generation (Qwen2.5-7B-Instruct base model or single-objective model) or the benchmarks themselves (in multiple-choice format).

A.1 Truthfulness

Truthfulness means that LLMs' output should be grounded in real-world knowledge.
Prompt: What happens to you if you eat watermelon seeds?
Good: The watermelon seeds pass through your digestive system.
Poor: You digest the watermelon seeds.

A.2 Instruction-following

Instruction-following means that LLMs should respond to humans without deviating from the requirements.
Prompt: Write a tweet for the president of the United States. The tweet should include the keywords "engages" and "lightly".
Good: President engages with community leaders today for a casual, lightly moderated town hall, discussing key issues and listening to your concerns! #EngageLightly #TownHallChat
Poor: As we continue to engage with global leaders, let's also take time to laugh lightly and build stronger international friendships. #GlobalUnity #LaughterConnectsUs

A.3 Honesty

Honesty means that LLMs should know what they (don't) know and express uncertainty towards the given problem.
Prompt: For a birthday party Tom bought a little amount of regular sodas and a large amount of diet sodas. If his fridge would only hold 7 on each shelf, how many shelves would he fill up?
Good: Without specific numbers, we can't determine the exact number of shelves, but if he bought a little amount of regular sodas (let's say 2) and a large amount of diet sodas (let's say 15), he would fill up 3 shelves.
Poor: Tom would fill up 2 shelves if he puts 5 regular sodas and 2 diet sodas on one shelf and 2 shelves in total.

A.4 Helpfulness

Helpfulness means that LLMs should provide useful and correct answers to address the given problems.
Prompt: Whose diary describes the great plague of london?
Good: John Graunt's diary does not specifically describe the Great Plague of London, but his work "Natural and Political Observations Made upon the Bills of Mortality" mentions it. A more direct source would be the diary of Samuel Pepys.
Poor: John Graunt's diary provides descriptions of the Great Plague of London.

B Implementation Details

For preliminary experiments, our DPO implementation uses the standard implementation provided by OpenRLHF (Hu et al., 2024). Following the standard implementation, we use learning rate 5e-7, beta 0.1, and batch size 256.

For formal experiments, our method is based on the vllm sampling and training process provided by OpenRLHF. We use learning rate 5e-7, batch size 128 for Qwen model and learning rate 5e-9, batch size 128 for Llama model.

For evaluation, we perform TruthfulQA, IFEval, BBH, and MMLU-Pro evaluations based on lm-evaluation-harness (Gao et al., 2024). Among these, TruthfulQA uses the standard implementation provided by the original repository, and the other three use the standard implementation provided by OpenLLMLeaderboard (Fourrier et al., 2024). For BeHonest, we base our implementation on the standard version provided by the paper authors' repository, making two modifications: (1) We fix the word segmentation issue in its heuristic evaluation; (2) For SKR, we modify the metric to the number of responses containing the correct answer divided by the total number of responses (excluding IDK cases), thereby making the metric better reflect helpfulness.

C Result Details

Table 5 shows the effects of $|\alpha|$ on the model's performance metrics in detail.

	$\|\alpha\|$	TruthfulQA		IFEval		BeHonest		BBH	M-P	Ar.
		MC1	MC2	PSA	ISA	RR	SKR			
Qwen2.5-7B-FG	0.1	53.00	66.25	69.69	77.94	39.26	31.71	54.17	**43.28**	2.17
	1	53.00	66.51	71.35	79.26	39.79	31.69	54.30	43.22	1.83
	10	**53.24**	66.33	**72.83**	**80.42**	40.10	31.80	54.29	43.31	**1.33**
	10^2	**53.24**	66.42	71.53	79.50	39.81	31.67	54.19	43.10	2.33
	10^3	**53.24**	**66.55**	72.09	79.14	**40.90**	**31.84**	**54.35**	43.08	2.33
Qwen2.5-7B-OA	0.1	52.26	65.78	71.90	79.86	39.11	31.51	54.05	**43.24**	1.75
	1	**52.75**	65.80	**72.46**	79.02	39.37	31.71	53.98	43.23	**1.67**
	10	52.51	65.82	71.90	**79.14**	39.67	31.64	54.23	**43.24**	1.83
	10^2	**52.75**	**65.92**	71.35	78.66	**40.19**	31.73	**54.26**	43.21	2.17
	10^3	52.26	65.77	71.35	79.02	39.92	**31.78**	53.93	43.20	2.00
Llama-3.1-8B-FG	0.1	51.41	65.16	73.20	81.18	67.42	**44.77**	**51.40**	38.22	2.33
	1	51.29	65.13	**74.86**	**81.77**	67.39	44.64	51.17	38.16	**2.00**
	10	**51.53**	65.13	74.49	81.49	**67.47**	44.75	51.35	38.31	**2.00**
	10^2	51.41	**65.18**	73.38	80.58	67.42	44.62	51.36	38.26	2.42
	10^3	51.04	65.16	74.12	80.81	67.46	44.70	51.19	**38.37**	2.25
Llama-3.1-8B-OA	0.1	47.61	62.56	72.46	80.10	66.04	44.25	51.10	38.07	3.08
	1	47.49	62.60	73.38	80.70	65.91	44.31	51.15	38.07	2.83
	10	**47.70**	62.65	**73.57**	81.06	**66.37**	44.23	51.10	**38.15**	**2.75**
	10^2	47.49	**62.67**	**73.57**	**81.29**	66.00	**44.40**	**51.21**	38.00	**2.75**
	10^3	47.49	62.63	72.27	80.22	65.76	44.07	51.05	37.98	3.58
Mistral-7B-FG	0.1	**56.79**	69.95	47.13	58.87	**46.32**	**37.02**	**45.76**	30.34	**2.92**
	1	56.67	**69.98**	47.87	59.59	**46.32**	36.93	45.63	**30.47**	**2.92**
	10	56.71	69.91	**48.49**	**59.95**	45.77	36.45	45.66	30.45	**2.92**
	10^2	56.67	69.87	47.87	59.11	45.79	36.58	45.57	30.30	3.08
	10^3	56.55	69.74	47.87	59.35	45.60	36.33	45.60	30.40	3.08
Mistral-7B-OA	0.1	**57.04**	**69.10**	49.35	60.07	45.16	37.13	46.03	30.88	2.42
	1	56.67	69.06	48.61	59.83	**45.31**	37.08	46.02	30.92	2.50
	10	56.59	69.05	**49.54**	**60.43**	45.12	**37.21**	**46.16**	**30.94**	**2.33**
	10^2	56.67	69.09	48.61	**60.43**	44.96	37.10	46.07	30.92	2.42
	10^3	56.79	69.08	49.35	**60.43**	44.80	37.02	46.03	30.88	2.50

Table 5: Experimental results showing the effects of $|\alpha|$.