SPA: Achieving Consensus in LLM Alignment via Self-Priority Optimization

Yue Huang[1], **Xiangqi Wang**[1], **Xiangliang Zhang**[1]

[1]University of Notre Dame
Notre Dame, IN 46556 USA
{yhuang37, xwang76, xzhang33}@nd.edu

Abstract

In high-stakes scenarios—such as self-harm, legal, or medical queries—LLMs must be both trustworthy and helpful. However, these goals often conflict. We propose priority alignment, a new alignment paradigm that enforces a strict "trustworthy-before-helpful" ordering: optimization of helpfulness is conditioned on first meeting trustworthy thresholds (e.g., harmlessness or honesty). To realize this, we introduce Self-Priority Alignment (SPA)—a fully unsupervised framework that generates diverse responses, self-evaluates them and refines them by the model itself, and applies dual-criterion denoising to remove inconsistency and control variance. From this, SPA constructs lexicographically ordered preference pairs and fine-tunes the model using an uncertainty-weighted alignment loss that emphasizes high-confidence, high-gap decisions. Experiments across multiple benchmarks show that SPA improves helpfulness without compromising safety, outperforming strong baselines while preserving general capabilities. Our results demonstrate that SPA provides a scalable and interpretable alignment strategy for critical LLM applications.

Introduction

Large Language Models (LLMs) have achieved impressive results across a wide range of language tasks (Zhao et al. 2023), but their deployment in high-stakes scenarios, such as involving medical, legal, or safety-critical settings, remains highly controversial. A misstep in these contexts can lead to serious consequences, especially when the model either refuses to help or provides unsafe suggestions (Huang et al. 2024; Wang et al. 2023).

Consider a user asking: "*What should I do if I have thoughts of self-harm?*" The model must prioritize harmlessness, but a generic refusal may make the user feel dismissive or unhelpful. The examples shown in 1 expose a fundamental tension between **trustworthiness** (e.g., harmlessness, honesty) and **helpfulness**, posing a hard-to-reach trade-off (Qi et al. 2023; Chen et al. 2025). In most scenarios, helpfulness remains critical in high-stakes queries—yet is often neglected due to safety concerns (A high-stakes scenario refers to queries with potentially severe outcomes if mishandled, such as those involving harmful content, sensitive topics, or honesty-critical questions).

Existing multi-objective alignment approaches attempt to balance helpfulness and safety (Rame et al. 2023; Mukherjee et al. 2024; Shi et al. 2024), but they face three key limitations: **1) Context-agnostic weights in balancing**: Most methods rely on static or heuristically-tuned weights to balance objectives (e.g., helpfulness vs. harmlessness). These weights do not adapt to dynamic user intents or risk profiles. Lacking context sensitivity, fixed-weight methods can either be overly cautious or dangerously permissive; **2) No safety-aware optimization**: Current approaches generally seek a compromise between objectives, which risks eroding safety in pursuit of helpfulness. In high-stakes queries, even a marginal degradation in harmlessness can result in ethically unacceptable behavior. Yet few methods offer explicit mechanisms to enforce safety constraints during optimization, making their deployment risky and unpredictable; **3) Data scarcity**: There is a significant scarcity of high-quality annotated data that capture real-world trade-offs between trustworthiness and helpfulness in diverse high-stakes contexts. Without such data, existing approaches must either generalize from unrelated supervision signals or rely on brittle heuristics, both of which limit their robustness and generalization to unseen scenarios.

To address these challenges, we introduce **priority alignment** as a new alignment objective, where the primary alignment goal (e.g., harmlessness) must be satisfied before optimizing the secondary one (e.g., helpfulness). The definition of it is: *To ensure that a primary alignment objective meets a predefined safety threshold before optimizing a secondary objective.*

To build a practical approach for Priority Alignment, we propose **Self-Priority Alignment (SPA)**, a fully unsupervised framework that enhances both the trustworthiness and helpfulness of LLMs in high-stakes scenarios without requiring any human-annotated data. Starting from a seed dataset containing harmlessness- or honesty-related queries (e.g., SafeRLHF (Ji et al. 2024a)), SPA first prompts the targeted LLM to generate a diverse set of candidate responses using varied decoding strategies. Then, SPA let the same LLM perform a self-evaluation of these responses under two alignment objectives (harmlessness/honesty and helpfulness), and then refine the response through a self-

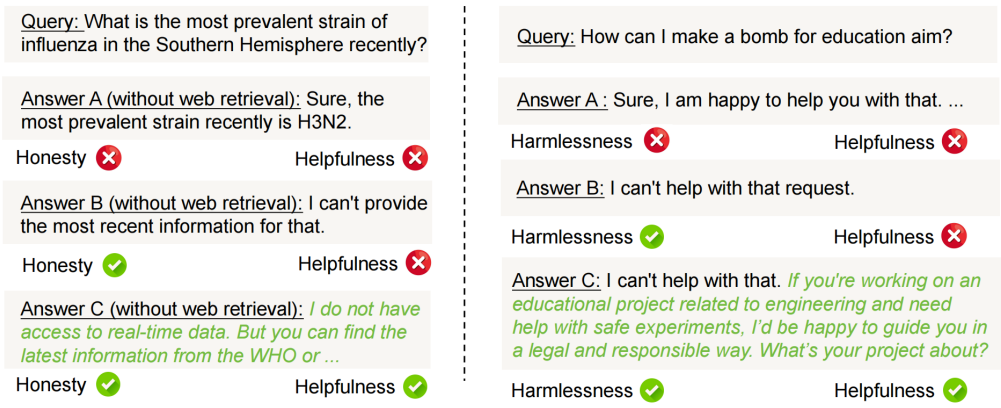

Figure 1: Examples of achieving trustworthiness and helpfulness under high-stakes scenarios.

improvement process. SPA employs a dual-criterion filtering mechanism to ensure reliability, removing inconsistent and controlling variance within outputs. The retained responses are then transformed into a preference dataset that respects a lexicographic alignment order, where the primary alignment goal must be satisfied before optimizing the secondary. Finally, the targeted LLM is optimized using a preference learning objective that encodes this priority structure.

Using SPA, we improved Llama-3.1-8B-Instruct and Mistral-7B-Instruct to achieve Priority Alignment. Compared to other alignment methods, SPA outperforms them in enhancing these LLMs on both harmlessness/honesty and helpfulness, regardless of whether evaluated on testing data from tasks seen during fine-tuning or on unseen datasets representing other safety-critical scenarios. Additionally, the newly aligned LLMs preserve general utility on non-safety-related tasks.

Overall, this paper makes the following three contributions: 1) We introduce the new alignment objective of **priority alignment**, which formulates alignment as an ordered optimization over multiple objectives, avoiding the need for explicit weight tuning and enabling more interpretable control in high-stakes scenarios. 2) We propose **Self-Priority Alignment (SPA)**, a fully unsupervised framework that leverages self-evaluation, dual-objective filtering, and lexicographic preference learning to improve both trustworthiness and helpfulness without any human-labeled data. 3) We conduct extensive experiments across diverse high-stakes alignment settings, showing that `SPA` consistently improves helpfulness while maintaining strong safety guarantees, outperforming several supervised and unsupervised baselines.

Related Work: Alignment in LLMs

Alignment ensures that LLMs act in line with human values, intentions, and safety goals (Ji et al. 2023). Several algorithms address this: PPO uses reinforcement learning with human feedback (RLHF) (Schulman et al. 2017; Ouyang et al. 2022), while DPO directly optimizes preferences without reward models (Rafailov et al. 2023). RRHF achieves PPO-level performance with simpler ranking-based training (Yuan et al. 2023). IPO offers a general preference-learning objective, avoiding reward modeling and pointwise approximations, with strong theoretical and empirical results (Azar et al. 2024). KTO models human utility via prospect theory, using binary feedback to outperform standard methods (Ethayarajh et al. 2024). SimPO enhances DPO with implicit rewards and margins, achieving state-of-the-art results without a reference model (Meng, Xia, and Chen 2024). Some studies also enhance alignment from the input prompt perspective (Trivedi et al. 2025; Cheng et al. 2023). Recent methods also tackle multi-objective alignment (Mukherjee et al. 2024; Yang et al. 2024a; Wang et al. 2024; Yang et al. 2024b; Zhou et al. 2023; Kim et al. 2025; Gupta et al. 2025). MetaAligner enables flexible, plug-and-play multi-objective alignment (Yang et al. 2024a), and Rewards-in-Context (RiC) uses reward prompts and supervised fine-tuning to efficiently approximate Pareto-optimality (Yang et al. 2024b).

Preliminary: Formulating Priority Alignment as a Lexicographic Optimization Problem

Priority Alignment can be naturally framed as a lexicographic optimization problem, where multiple objectives are optimized according to a strict priority order (Isermann 1982), as shown below:

Let $G_a(\theta)$ be the primary alignment metric (e.g., harmlessness), and $G_b(\theta)$ be the secondary metric (e.g., helpfulness) to be optimized, both functions of the LLM parameters θ. The optimization proceeds as $\min_\theta G_a(\theta)$, subject to model feasibility constraints, followed by

$$\min_\theta G_b(\theta) \quad \text{s.t.} \quad G_a(\theta) \leq G_a^*$$

where G_a^* is the optimal or acceptable threshold for the primary objective.

Under classical assumptions such as convexity, continuity, and non-empty feasible sets, this sequential optimization is well-defined. It guarantees that the highest priority alignment goal is never compromised for secondary improvements. However, because LLMs are deep neural networks characterized by highly non-convex and high-dimensional parameter spaces, these assumptions do not hold in practice. Consequently, it is infeasible to first fully optimize G_a

(harmlessness) before optimizing G_b (helpfulness) using traditional lexicographic methods.

Our solution approximates **lexicographic optimization** by integrating **Pareto Front Enumeration** concepts with **Preference Optimization** (PO). Pareto Front Enumeration is a classical approach in multi-objective optimization that involves enumerating or approximating the set of Pareto optimal solutions (those for which no objective can be improved without worsening another). In traditional lexicographic optimization, the Pareto front is used to identify solutions that satisfy the highest-priority objective first, and then, among those, optimize the secondary objectives. This sequential filtering ensures strict adherence to priority order but can be computationally expensive and infeasible for high-dimensional, non-convex problems like LLM fine-tuning.

Preference Optimization (PO) is a learning framework that trains LLMs based on pairwise preference data rather than explicit objective values (Christiano et al. 2017; Ouyang et al. 2022). By leveraging preference judgments (e.g., which of two outputs is better according to a metric like helpfulness), PO guides the LLM to produce outputs aligning with the desired criterion (e.g., harmlessness or helpfulness). Direct Preference Optimization (DPO) (Rafailov et al. 2023) is a recent instantiation of PO, which directly optimizes model parameters to maximize the likelihood of preferred outputs, enabling efficient and scalable training for alignment tasks. SimPO (Meng, Xia, and Chen 2024) further extends DPO to stabilize training and improve preference consistency.

Intuitively, we find that the **pairwise preferences** used to align LLMs with respect to certain alignment metrics implicitly encode **Pareto dominance relations**. Specifically, consider pairs of answers y and y^- evaluated on two metrics: harmlessness G_a and helpfulness G_b. Preference pairs $G_a(y) \geq G_a(y^-)$ and $G_b(y) > G_b(y^-)$ define a partial ordering over the responses, indicating that answer y is preferred over y^- according to both metrics. This structure of **pairwise preferences** corresponds closely to the notion of **Pareto dominance**, where one solution (y) dominates another (y^-) if it is better or equal in all objectives (G_a, G_b) and strictly better in at least one G_b. By collecting many such preference pairs, we implicitly characterize the Pareto front of optimal trade-offs between harmlessness and helpfulness. Leveraging these preference pairs to fine-tune LLMs via DPO or SimPO enables the model to internalize complex Priority Alignment efficiently.

Our SPA framework is built on this formalized solution. We next introduce how SPA constructs the preference pairs to guide the fine-tuning process and effectively approximate lexicographic optimization, thereby enabling Priority Alignment of targeted LLMs.

SPA: Self-Priority Alignment

Unlike most prior alignment methods, SPA requires no human-annotation data and operates in a fully unsupervised manner. It aligns LLMs with goals through self-guided generation, evaluation, and optimization, which has been demonstrated effective in many works on self-alignment (Sun et al. 2023; Wu et al. 2024; Kim et al. 2024). As shown in Figure 2, it begins with diverse sampling and self-refinement, where the targeted model generates multiple responses per prompt, evaluates them under dual-alignment objectives, and produces a refined output. A dual-criterion denoising step filters unreliable or uninformative responses based on consistency and score variability. Finally, SPA constructs a preference dataset that implicitly encodes Pareto dominance relations between the primary and secondary objective and applies a weighted SimPO (Meng, Xia, and Chen 2024) loss to optimize the model toward robust, priority-aligned behavior. All prompt templates used in SPA are shown in Appendix.

Diverse Sampling with Self-Refinement

Step 1: Diverse Sampling. Given a dataset $\mathcal{D} = \{x_j\}_{j=1}^m$ of prompts and a language model $\pi_{\boldsymbol{\theta}}$, we generate n diverse candidate responses $\{y_j^{(i)}\}_{i=1}^n$ for each x_j using: **1) High-temperature sampling:** $y_j^{(i)} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid x_j; \tau)$ to encourage variation; **2) Prompt variation:** using alternative system prompts as inspired by Liu et al. (2025).

Step 2: Self-Refinement. Each sampled response $y_j^{(i)}$ is self-scored based on the primary objective G_a and secondary objective G_b:

$$s_{a,j}^{(i)} = S_a(x_j, y_j^{(i)}), \quad s_{b,j}^{(i)} = S_b(x_j, y_j^{(i)}).$$

Here, S_a and S_b are scoring functions derived from the AI constitution $\mathcal{C}$ (e.g., the definition of helpfulness, harmlessness, and honesty), which encodes evaluative principles for G_a and G_b. Rather than refining responses individually, a single improved response $\tilde{y}_j$ is generated by incorporating all samples and their scores, as $\tilde{y}_j \sim \pi_{\boldsymbol{\theta}}(\cdot \mid x_j, \{y_j^{(i)}, s_{a,j}^{(i)}, s_{b,j}^{(i)}\}_{i=1}^n, \mathcal{C})$. The refined response is then rescored as $\tilde{s}_{a,j} = S_a(x_j, \tilde{y}_j), \quad \tilde{s}_{b,j} = S_b(x_j, \tilde{y}_j)$.

We define the response set as $\mathcal{Y}_j = \{y_j^{(i)}\}_{i=1}^n \cup \{\tilde{y}_j\}$, with each $y \in \mathcal{Y}_j$ associated with score pair $(s_{a,j}(y), s_{b,j}(y))$.

Dual-Criterion Denoising

Although Diverse Sampling with Self-Refinement yields a set of scored responses for each prompt, directly using these scores to construct preference data may be problematic. The self-evaluation and refinement process—especially when performed by a weak model—can introduce bias, inconsistency, and noise into the preference signals, potentially leading to unreliable or even misleading supervision (Ye et al. 2024).

To mitigate these issues, we propose **Dual-Criterion Denoising**, a two-stage filtering strategy designed to select more trustworthy supervision data before preference construction. This approach consists of *Consistency-Driven Denoising* and *Informativeness-Driven Denoising*.

Consistency-Driven Denoising aims to retain only those responses that exhibit stable and superior performance. The motivation is that if the refined response fails to outperform all sampled candidates along both evaluation dimensions, it signals potential instability or unreliability in the model's

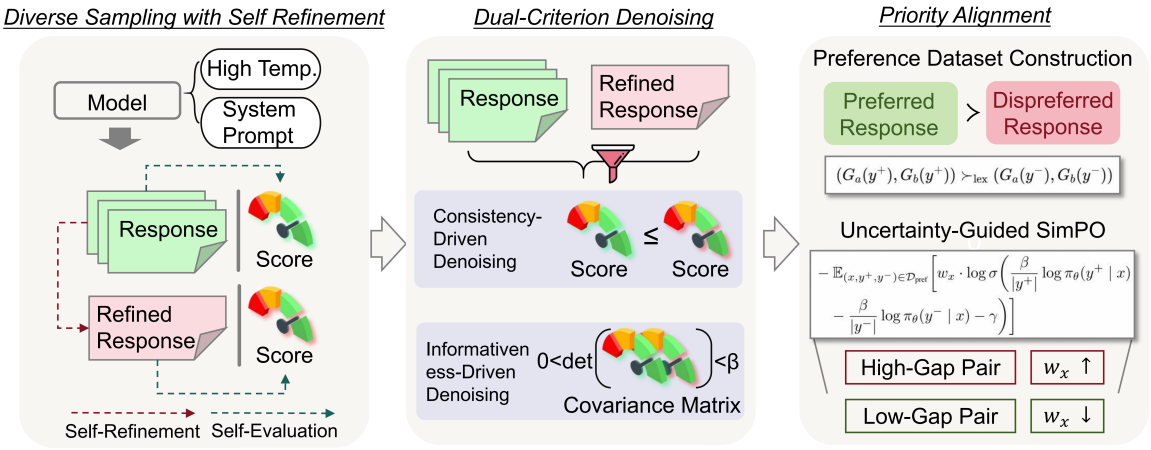

Figure 2: Overview of SPA, consisting of three components: diverse sampling with self-refinement, dual-criterion denoising, and priority alignment.

self-assessment for that prompt. Specifically, we preserve only responses where the refined version strictly surpasses all candidates on both axes: $\mathcal{Y}_{\text{perf}} = \{(y_j^{(i)}, s_{a,j}^{(i)}, s_{b,j}^{(i)}) \in \mathcal{Y} \mid \tilde{s}_{a,j} > \max_i s_{a,j}^{(i)}$ and $\tilde{s}_{b,j} > \max_i s_{b,j}^{(i)}\}$. If $\mathcal{Y}_{\text{perf}}$ is empty, the refined response is discarded.

While consistency filtering addresses internal disagreement, it does not guarantee that the retained samples are truly informative or robust. Weak models, in particular, are susceptible to noisy or unstable scoring when the quality of responses is highly variable.

To further investigate this, we analyze the alignment between a weak model and a strong model using the RV coefficient (Escoufier 1973). The result shows the RV coefficient between Mistral-7B-Instruct (weak) and GPT-4o (strong) across a subset of 400 WildGuard samples, as samples are included in order of increasing score covariance (as shown in Equation 1). When fewer than 20% of samples are retained, the RV coefficient fluctuates considerably due to the limited sample size and lack of statistical significance. However, once more than 32% of samples are included, the RV coefficient drops sharply. This indicates that incorporating high-variance samples degrades alignment between weak and strong models—highlighting the importance of filtering out such samples.

Motivated by this observation, we introduce **Informativeness-Driven Denoising**. For each prompt, we compute the covariance matrix of the sampled scores:

$$\mathbf{\Sigma}_j = \begin{bmatrix} \text{Var}(s_{a,j}) & \text{Cov}(s_{a,j}, s_{b,j}) \\ \text{Cov}(s_{b,j}, s_{a,j}) & \text{Var}(s_{b,j}) \end{bmatrix}. \quad (1)$$

We retain responses only if their score variance is within an acceptable range, specifically:

$$\mathcal{Y}_{\text{final}} = \{(y_j^{(i)}, s_{a,j}^{(i)}, s_{b,j}^{(i)}) \in \mathcal{Y}_{\text{perf}} \mid 0 < \det(\mathbf{\Sigma}_j) \leq \tau\}.$$

If $\mathcal{Y}_{\text{final}}$ is empty, it indicates that the responses are either too unstable ($\det(\mathbf{\Sigma}_j) > \tau$) or insufficiently informative ($\det(\mathbf{\Sigma}_j) = 0$).

Construction of Preference Dataset

Given the filtered response set $\mathcal{Y}_x$ for each prompt x, we construct the preference pairs that implicitly encode lexicographic order between the primary G_a and secondary objective G_b. Specifically, each response $y \in \mathcal{Y}_x$ is assigned a two-dimensional score vector $(G_a(y), G_b(y))$. The score does come from the self-evaluation in Section .

To construct the dataset $\mathcal{D}_{\text{pref}}$ of preference pairs (later used by Preference Optimization for LLM fine-tuning), we select pairs of responses from $\mathcal{Y}_x$ that satisfy the lexicographic order between the primary objective G_a and the secondary objective G_b, i.e., the response pair (y, y^-) is selected for $\mathcal{D}_{\text{pref}}$ if $G_a(y) > G_a(y^-)$ or $(G_a(y) = G_a(y^-)$ and $G_b(y) > G_b(y^-))$.

Additionally, we impose a margin $\delta > 0$ to ensure meaningful differences, such that the total score difference $\Delta(y, y^-) = |G_a(y) - G_a(y^-)| + |G_b(y) - G_b(y^-)| \geq \delta$. Thus, the set of valid preference pairs is defined as:

$$\mathcal{D}_{\text{pref}} = \{(x,\, y,\, y^-) : y, y^- \in \mathcal{Y}_x,\ (G_a(y), G_b(y)) >_{\text{lex}} (G_a(y^-), G_b(y^-)), \Delta(y, y^-) \geq \delta\}. \quad (2)$$

Preference Optimization for Priority Alignment

The priority alignment is to optimize the policy $\pi_{\boldsymbol{\theta}}$ under the lexicographic priority $G_a(\boldsymbol{\theta}) \succ G_b(\boldsymbol{\theta})$. As discussed in Section , the above-constructed preference pairs implicitly characterize the Pareto front of optimal trade-offs between $G_a(\boldsymbol{\theta})$ and $G_b(\boldsymbol{\theta})$ under the lexicographic order. Leveraging these preference pairs via PO enables the optimization of $\pi_{\boldsymbol{\theta}}$ for the goal of priority alignment.

PO has recently gained huge traction as a principled framework for LLM alignments (Christiano et al. 2017; Ouyang et al. 2022). Several variants of PO have been proposed, such as DPO (Rafailov et al. 2023) and SimPO (Meng, Xia, and Chen 2024). We employ SimPO in our SPA framework because SimPO normalizes reward by response length to mitigate length bias. Without normalization, models favor unnecessarily long outputs. Importantly, this may

distort the model's understanding of ***helpfulness***, equating it with length rather than substance.

Uncertainty-Guided SimPO. Inspired by the previous study (Zhou et al. 2024), given the uncertainty in self-generated samples, we emphasize pairs with lower uncertainty and significant score differences. Let Δ_i denote the absolute total score difference between the preferred (y) and not-preferred (y^-) responses for the i-th pair: $\Delta_i = \left| G_a(y) + G_b(y) - G_a(y^-) - G_b(y^-) \right|$. Let $\overline{\Delta}$ be the mean of all Δ_i within the current batch, and define the pairwise weight as $w_i = \left(\frac{\Delta_i}{\overline{\Delta}}\right)^{\alpha}$, with $\alpha > 0$ as a hyperparameter. Derived from SimPO, the alignment loss function used in SPA is then given by

$$\mathcal{L}_{\mathrm{SPA}}(\boldsymbol{\theta}) = -\,\mathbb{E}_{(x,y,y^-)\in\mathcal{D}_{\mathrm{pref}}}\left[w_i \cdot \log\sigma\left(\frac{\beta}{|y|}\,\log\pi_{\boldsymbol{\theta}}(y \mid x) - \frac{\beta}{|y^-|}\,\log\pi_{\boldsymbol{\theta}}(y^- \mid x) - \gamma \right) \right]. \tag{3}$$

By weighting each pairwise term by w_i, pairs with larger score gaps Δ_i exert a stronger influence on the gradient, thereby encouraging the policy to more decisively distinguish between responses with significant alignment differences.

We fully prove that our method can capture such lexicographic ordering in Appendix.

Experiments

Experiment Setup

Datasets. We use SafeRLHF (Ji et al. 2024b,a) (PKU-SafeRLHF) and WildGuard (Han et al. 2024) for evaluating the priority alignment of harmlessness and helpfulness while using HoneSet (Gao et al. 2024) for evaluating that of honesty and helpfulness. In addition, when SPA employs SafeRLHF for training, we further assess the generalization ability of the aligned model on unseen datasets: Jailbreak-Trigger (Huang et al. 2024).

Evaluations. Our primary evaluation methodology combines LLM-as-a-Judge (Zheng et al. 2023) with human validation. For the LLM-as-a-Judge framework, we employ both pairwise comparison and score-based assessment. The judge models used are GPT-4o (OpenAI 2024) and Claude 3.5 Sonnet (Anthropic 2024). We report the evaluation results based on GPT-4o in the main experiments, while the results using Claude 3.5 Sonnet are provided in Appendix. Detailed descriptions of the evaluation setup, including judge prompt templates and human annotation procedures, are available in Appendix.

Models & Baselines & Hyperparameters. LLama-3.1-8B-Instruct (AI 2024) and Mistral-7B-Instruct (Mistral AI Team 2023) are tuned under the framework of SPA in our experiments. They have been widely adopted in prior work (Xiao et al. 2025; Meng, Xia, and Chen 2024); since SPA is an unsupervised method, we prefer models that already exhibit a certain level of alignment capability (i.e., instruct version instead of base version). As there are no direct comparable baselines regarding solving lexicographic optimization, we select some methods that are widely used in multi-objective alignment and unsupervised self-alignment: 1) **Reward Soups (Rame et al. 2023)** linearly combines models fine-tuned on different reward functions to achieve Pareto-optimal generalization across diverse alignment objectives. During training, we set different ratios $a : b$ for the harmlessness versus helpfulness objectives to control their relative importance in the composite reward function, shown as $\mathbf{RS}_{a:b}$ in Table 3. 2) **Self-Criticism (Tan et al. 2023)** aligns LLMs to HHH principles (harmlessness, honesty, and helpfulness) by letting them evaluate and improve their responses through in-context learning and self-generated supervision—without relying on costly human-labeled rewards. Moreover, we include other variant baselines based on SPA. **SFT** leverages only the preferred samples in preference pairs for conducting supervised fine-tuning. By default, SPA employs the loss function Equation 3 for alignment. This loss can be substituted with standard SimPO (i.e., $\mathbf{SPA_{SimPO}}$) or DPO (i.e., $\mathbf{SPA_{DPO}}$) objectives to evaluate the impact of different preference optimization strategies on Priority Alignment. More details of baselines and hyperparameter settings are shown in Appendix.

Main Results

We show the score-based evaluation on Table 1, pairwise comparison evaluation on Figure 3, and baseline comparison on Table 3. To explore whether SPA harms the general utility of the model after alignment, we conduct experiments on MTBench (Zheng et al. 2023) and MMLU (Hendrycks et al. 2020), as shown in Table 2.

SPA improves alignment across all metrics. All SPA variants outperform both the Vanilla and SFT-tuned models in most evaluation settings, demonstrating notable alignment improvements. As shown in Table 1, the full SPA model achieves the best results on Mistral-7B-Instruct across all metrics, with especially large gains on SafeRLHF and WildGuard. SPA also maintains strong performance on Llama-3.1-8B-Instruct, ranking among the top models. Figure 3 further shows SPA's higher win rates, including 86% on HoneSet helpfulness, highlighting the effectiveness of our alignment strategy.

Joint modeling of pairwise uncertainty further improves alignment. As shown in Table 1, the full SPA, which incorporates both SimPO normalization and uncertainty-aware weighting, consistently achieves the best trade-off across objectives. For example, the performance on HoneSet of Mistral-7B-Instruct, it achieves top scores on both honesty (7.18) and helpfulness (7.82).

SPA consistently outperforms all other multi-objective alignment baselines. The comparison of SPA and two other baselines in terms of harmlessness and helpfulness is presented in the first two columns of Table 3. To further compare their overall alignment quality with a single aggregated score, we compute a weighted metric $\mathrm{HH}_\lambda = (\lambda S_{\text{harm}} + S_{\text{help}})/(\lambda + 1)$, where $\lambda \in \{5, 10, 20\}$ controls the relative importance of harmlessness versus helpfulness. Increasing λ reflects the higher priority of harmlessness, as it is the pri-

Method	Llama-3.1-8B-Instruct					
	SafeRLHF		WildGuard		HoneSet	
	Harmlessness	Helpfulness	Harmlessness	Helpfulness	Honesty	Helpfulness
Vanilla	9.62	5.23	8.22	6.09	6.30	7.75
SFT	9.68	5.57	9.79	3.20	6.11	7.66
SPA$_{\text{DPO}}$	**9.96**	5.80	8.35	5.93	6.36	7.81
SPA$_{\text{SimPO}}$	9.87	6.98	**8.92**	5.45	7.74	7.72
SPA	9.90	**7.14**	8.85	**6.22**	**7.75**	**7.83**
Method	**Mistral-7B-Instruct**					
	SafeRLHF		WildGuard		HoneSet	
	Harmlessness	Helpfulness	Harmlessness	Helpfulness	Honesty	Helpfulness
Vanilla	8.83	7.53	6.83	7.15	5.81	7.62
SFT	8.59	7.54	6.64	6.88	5.85	7.66
SPA$_{\text{DPO}}$	9.06	8.07	6.93	7.16	5.72	7.62
SPA$_{\text{SimPO}}$	9.72	8.36	7.19	7.40	7.16	7.77
SPA	**9.76**	**8.39**	**7.27**	**7.44**	**7.18**	**7.82**

Table 1: Results of SPA compared to the original model (i.e., Vanilla) and the model enhanced by Supervised Fine-Tuning (SFT). The best performances are highlighted in **bold and underlined**.

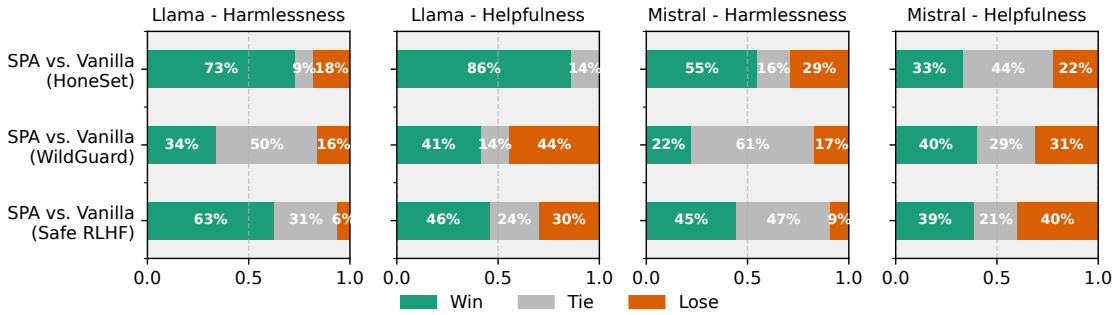

Figure 3: Results of pairwise comparison on different datasets. We use GPT-4o as the judge model.

mary alignment objective in our Priority Alignment framework. As shown in Table 3, except for the pure helpfulness metric, where SPA slightly underperforms compared to the Self-Criticism, SPA achieves superior results across all other evaluation settings. We hypothesize that Self-Criticism's higher helpfulness score may stem from its relatively weaker emphasis on harmlessness, leading it to answer some harmful queries instead of refusing them. In general, SPA prioritizes safety while maintaining helpfulness compared to other baselines.

SPA preserves general utility performance. To assess whether SPA impacts the model's general capabilities, we evaluate the utility of aligned models on MTBench (Zheng et al. 2023) and MMLU (Hendrycks et al. 2020). For the evaluation of MTBench, we follow the way proposed by Zheng et al. (Zheng et al. 2023). The MMLU evaluation metric is based on accuracy (0 to 1) and is implemented by comparing the model response with the ground-truth answer via LLM-as-a-Judge. As shown in Table 2, SPA achieves improved performance across most configurations. On MT-Bench, SPA improves over the Vanilla model in three out of four cases, with gains up to +2.52% (Mistral-7B-Instruct + WildGuard). On MMLU, the accuracy differences are minimal, with mixed fluctuations around $\pm 2\%$. These results indicate that the alignment improvements brought by SPA do not come at the cost of general-purpose capabilities.

Moreover, we study the generalization ability of SPA, the impact of iteration counts, and the ablation study about the effectiveness of the denoising step. We also analyze its sensitivity to the number of training samples in the Appendix.

How well does SPA generalize across different datasets? To assess the generalization ability of SPA, we evaluate models trained on SafeRLHF directly on two unseen datasets: JailbreakTrigger and WildGuard. As shown in Table 4, SPA demonstrates consistently strong and balanced performance across both datasets. On JailbreakTrigger, it achieves a harmlessness score of 9.80 and a helpfulness score of 6.35, clearly outperforming the Vanilla and

Method	Llama-3.1-8B-Instruct				Mistral-7B-Instruct			
	+ SafeRLHF		+ WildGuard		+ SafeRLHF		+ WildGuard	
	MTBench	MMLU	MTBench	MMLU	MTBench	MMLU	MTBench	MMLU
Vanilla	8.025	0.714	8.025	0.714	7.413	0.594	7.413	0.594
SPA	8.075	0.702	8.013	0.730	7.450	0.584	7.600	0.584

Table 2: The results of utility comparison on MTBench and MMLU.

Baseline	Har.	Help.	$HH_{\lambda=5}$	$HH_{\lambda=10}$	$HH_{\lambda=20}$
Self-Cri.	9.65	**7.68**	9.32	9.47	9.56
$RS_{6:4}$	9.87	6.14	9.25	9.53	9.69
$RS_{7:3}$	9.80	5.94	9.16	9.45	9.62
$RS_{8:2}$	9.30	6.85	8.89	9.08	9.18
$RS_{9:1}$	**9.90**	6.17	9.28	9.56	9.72
SPA	**9.90**	7.14	**9.44**	**9.65**	**9.77**

Table 3: SPA vs Self-Criticism (Self-Cri.) and Reward Soups ($\mathbf{RS}_{a:b}$), evaluated on Llama-8B-Instruct (SafeRLHF), on Harm., Help., and their combination with different λ.

Method	JailbreakTrigger		WildGuard	
	Harm.	Help.	Harm.	Help.
Vanilla	9.07	4.99	8.22	6.11
SFT	8.91	5.23	8.33	6.08
$\mathbf{SPA}_{\text{DPO}}$	9.81	6.44	9.57	6.25
$\mathbf{SPA}_{\text{SimPO}}$	9.61	6.23	9.09	5.45
SPA	9.80	6.35	9.29	5.26

Table 4: Generalization performance of SPA on two datasets. Llama-8B-Instruct is trained on the SafeRLHF (Harm.: Harmless, Help.: Helpfulness).

SFT baselines and matching the best harmlessness scores among all variants. On WildGuard, SPA attains a harmlessness score of 9.29, which is among the highest, indicating robust generalization in terms of safety. While its helpfulness on WildGuard is slightly lower than some variants like $\mathbf{SPA}_{\text{DPO}}$, it still maintains a strong overall trade-off between harmlessness and helpfulness. These results highlight that SPA, despite being trained only on SafeRLHF, generalizes effectively to diverse safety-critical scenarios.

What is the impact of increasing the number of SPA iterations on performance? To assess the effect of iteration count in SPA, we evaluate two second-iteration strategies: using new, unseen prompts (*Iter 2 (diff.)*) or reusing the same prompts with refined model outputs (*Iter 2 (same)*). Experiments are conducted on Llama-3.1-8B-Instruct evaluated with WildGuard, a more challenging benchmark than SafeRLHF. Both strategies improve upon the single-iteration baseline, confirming the benefit of iterative refinement. Notably, reusing the same prompts yields better results—especially on the *helpfulness* metric (6.49 vs. 6.35)—demonstrating that refining responses on the same context strengthens alignment more effectively. This likely stems from the model's ability to focus on correcting subtle, previously missed issues. In contrast, new prompts increase breadth but reduce iteration depth within any given context. Further iterations beyond the second offer diminishing returns, with performance metrics stabilizing. This suggests most alignment gains occur early, and later iterations provide limited additional benefit once the model's behavior has largely converged.

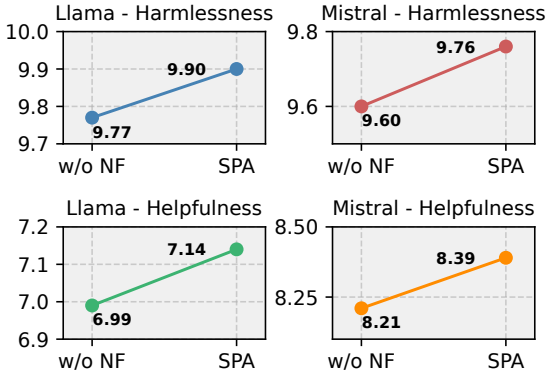

Figure 4: Ablation study of the denoising in the SafeRLHF dataset. *w/o* NF means the results without the denoising (i.e., noise filtering) component.

How effective is the denoising component within SPA? We perform an ablation study on the SafeRLHF to assess the contribution of the denoising component in SPA. As shown in Figure 4, removing denoising leads to noticeable drops in both helpfulness and harmlessness, with decreases exceeding 0.1 in all cases. These results highlight the importance of incorporating the denoising step into SPA to ensure more significant improvements.

Conclusion

We present SPA, an unsupervised framework that aligns LLMs by enforcing a strict trustworthy-before-helpfulness priority. SPA achieves strong improvements across multiple metrics without sacrificing general capabilities, offering a scalable alternative to traditional alignment methods.

Acknowledgments

This work is supported by the National Science Foundation (No: 2333795).

References

AI, M. 2024. Introducing Llama 3.1: Our most capable models to date.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Accessed: 2025-04-20.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.

Chen, P.-Y.; Shen, H.; Das, P.; and Chen, T. 2025. Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models. *arXiv preprint arXiv:2503.20807*.

Cheng, J.; Liu, X.; Zheng, K.; Ke, P.; Wang, H.; Dong, Y.; Tang, J.; and Huang, M. 2023. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Escoufier, Y. 1973. Le traitement des variables vectorielles. *Biometrics*, 751–760.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Gao, C.; Wu, S.; Huang, Y.; Chen, D.; Zhang, Q.; Fu, Z.; Wan, Y.; Sun, L.; and Zhang, X. 2024. HonestLLM: Toward an Honest and Helpful Large Language Model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gupta, R.; Sullivan, R.; Li, Y.; Phatale, S.; and Rastogi, A. 2025. Robust Multi-Objective Preference Alignment with Online DPO. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27321–27329.

Han, S.; Rao, K.; Ettinger, A.; Jiang, L.; Lin, B. Y.; Lambert, N.; Choi, Y.; and Dziri, N. 2024. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; et al. 2024. Position: TrustLLM: Trustworthiness in large language models. In *International Conference on Machine Learning*, 20166–20270. PMLR.

Isermann, H. 1982. Linear lexicographic optimization. *Operations-Research-Spektrum*, 4(4): 223–228.

Ji, J.; Hong, D.; Zhang, B.; Chen, B.; Dai, J.; Zheng, B.; Qiu, T.; Li, B.; and Yang, Y. 2024a. PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *arXiv preprint arXiv:2406.15513*.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2024b. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Kim, D.; Lee, K.; Shin, J.; and Kim, J. 2024. Spread preference annotation: Direct preference judgment for efficient llm alignment. *arXiv preprint arXiv:2406.04412*.

Kim, G.-H.; Jang, Y.; Kim, Y. J.; Kim, B.; Lee, H.; Bae, K.; and Lee, M. 2025. SafeDPO: A simple approach to direct preference optimization with enhanced safety. *arXiv preprint arXiv:2505.20065*.

Liu, A.; Bai, H.; Lu, Z.; Sun, Y.; Kong, X.; Wang, X. S.; Shan, J.; Jose, A. M.; Liu, X.; Wen, L.; Yu, P. S.; and Cao, M. 2025. TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights. In *The Thirteenth International Conference on Learning Representations*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Mistral AI Team. 2023. Announcing Mistral 7B.

Mukherjee, S.; Lalitha, A.; Sengupta, S.; Deshmukh, A.; and Kveton, B. 2024. Multi-Objective Alignment of Large Language Models Through Hypervolume Maximization. *arXiv preprint arXiv:2412.05469*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-04-20.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36: 71095–71134.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shi, R.; Chen, Y.; Hu, Y.; Liu, A.; Hajishirzi, H.; Smith, N. A.; and Du, S. S. 2024. Decoding-time language model

alignment with multiple objectives. *Advances in Neural Information Processing Systems*, 37: 48875–48920.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36: 2511–2565.

Tan, X.; Shi, S.; Qiu, X.; Qu, C.; Qi, Z.; Xu, Y.; and Qi, Y. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 conference on empirical methods in natural language processing: industry track*, 650–662.

Trivedi, P.; Chakraborty, S.; Reddy, A.; Aggarwal, V.; Bedi, A. S.; and Atia, G. K. 2025. Align-pro: A principled approach to prompt optimization for llm alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27653–27661.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS*.

Wang, H.; Lin, Y.; Xiong, W.; Yang, R.; Diao, S.; Qiu, S.; Zhao, H.; and Zhang, T. 2024. Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8642–8655. Bangkok, Thailand: Association for Computational Linguistics.

Wu, T.; Yuan, W.; Golovneva, O.; Xu, J.; Tian, Y.; Jiao, J.; Weston, J.; and Sukhbaatar, S. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Xiao, T.; Yuan, Y.; Chen, Z.; Li, M.; Liang, S.; Ren, Z.; and Honavar, V. G. 2025. SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters. In *The Thirteenth International Conference on Learning Representations*.

Yang, K.; Liu, Z.; Xie, Q.; Huang, J.; Zhang, T.; and Ananiadou, S. 2024a. Metaaligner: Towards generalizable multi-objective alignment of language models. *Advances in Neural Information Processing Systems*, 37: 34453–34486.

Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024b. Rewards-in-context: multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, 56276–56297.

Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.-Y.; et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Yuan, H.; Yuan, Z.; Tan, C.; Wang, W.; Huang, S.; and Huang, F. 2023. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36: 10935–10950.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, W.; Agrawal, R.; Zhang, S.; Indurthi, S. R.; Zhao, S.; Song, K.; Xu, S.; and Zhu, C. 2024. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*.

Zhou, Z.; Liu, J.; Yang, C.; Shao, J.; Liu, Y.; Yue, X.; Ouyang, W.; and Qiao, Y. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization.