MPAW: MULTI-PREFERENCE ALIGNMENT THROUGH WEAK MODEL COLLABORATION FOR EFFICIENT AND FLEXIBLE LLM DECODING

Nuo Chen
National University of Singapore

Guojun Xiong
Harvard University

Bingsheng He
National University of Singapore

ABSTRACT

Aligning large language models (LLMs) with diverse and competing human preferences remains a critical challenge for safe and effective deployment. While recent work demonstrates that decoding-time alignment via weak preference models achieves strong performance with minimal compute, existing methods optimize for single objectives, severely limiting their adaptability to real-world scenarios requiring multifaceted trade-offs (e.g., safety vs. helpfulness). We propose Multi-Preference Alignment through Weak Model Collaboration (MPAW), a scalable framework that aggregates guidance from heterogeneous weak preference models-smaller LLMs aligned to distinct objectives-into a unified decoding strategy. By dynamically integrating signals from specialized proxies (e.g., safety classifiers, conciseness scorers), MPAW preserves the generalization capabilities of large base models while enabling zero-shot adaptation to arbitrary preference weightings. Empirical results demonstrate reliable alignment quality and nearly matching the performance of computationally expensive multi-objective RLHF fine-tuning. Our findings establish weak model collaboration as a promising pathway for efficient, flexible LLM alignment without retraining.

1 INTRODUCTION

Recent advances in weak-to-strong decoding-time alignment have demonstrated the remarkable potential for small, aligned models to guide larger language models (LLMs) towards human preferences with minimal computational overhead, effectively bypassing the need for costly fine-tuning of the strong model (Burns et al., 2023). This paradigm—leveraging weak preference models as proxies for human feedback—unlocks significant efficiency gains while preserving the generalization capabilities of large foundation models. However, existing weak-to-strong methods operate under a critical limitation: they rely on guidance from a single weak model (Rafailov et al., 2024a; Zhou et al., 2024c;a; Mitchell et al., 2023; Liu et al., 2024a; Huang et al., 2024) , which inherently restricts their adaptability to real-world scenarios where users demand alignment across multiple, often competing objectives (e.g., harmlessness, helpfulness, and personalization).

While multi-objective alignment frameworks—including *post-hoc interpolation* (e.g., Reward Soup (Rame et al., 2023), MOD (Shi et al., 2024), MORLHF (Wu et al., 2023), and *dynamic preference optimization* MODPO (Zhou et al., 2024b) and RiC (Yang et al., 2024))—aim to address diverse preferences, these methods face two critical limitations. First, existing approaches require training or fine-tuning separate models for each preference P weighting configuration, incurring $\mathcal{O}(P)$ complexity where P grows combinatorially with the number of users or preference dimensions (Wu et al., 2023). Second, parameter fusion techniques (e.g., Reward Soup (Rame et al., 2023), MODPO (Zhou et al., 2024b)) mandate structural homogeneity (identical architectures, shared initialization), precluding the integration of heterogeneous weak models or black-box APIs encountered in real-world deployments.

Aligning large language models with multiple objectives requires more efficient and adaptable methods. In a scenario with M distinct reward functions, a naive approach would involve training a separate LLM for each combination of user preferences (e.g., safety, conciseness, creativity), leading to a large number of models, which becomes computationally unfeasible. Instead, we propose using

M pre-trained smaller models, each aligned with a specific reward function, to guide a larger LLM towards the desired behavior without further fine-tuning.

Additionally, we can enhance alignment by adjusting the output probabilities at each decoding step using logits tuning. Building on previous work that used a single weak model (Liu et al., 2024a), we extend this approach to use multiple weak models to target different aspects of human preference.

In this work, we propose a novel multi-weak model collaborative framework (`MPAW`) that synergizes the efficiency of weak-to-strong decoding with the expressiveness of multi-objective alignment. Our method dynamically aggregates guidance from diverse weak models—each specialized in distinct preference dimensions (e.g., safety, conciseness, creativity)—through a unified decoding-time scoring mechanism. This enables:

▷ **Zero-Shot Adaptability.** Our method enables seamless adjustment to arbitrary user preference weightings without retraining – achieving $\mathcal{O}(M)$ complexity for M fixed objectives compared to the $\mathcal{O}(P)$ scaling of traditional multi-objective alignment approaches, where P grows combinatorially with user-defined preference combinations.

▷ **Heterogeneous Model Compatibility.** `MPAW` can integrate a diverse set of weak models, including those that are black-box APIs or have different architectures. This makes it highly scalable and compatible with real-world deployment constraints.

▷ **Efficient Multi-Objective Alignment.** By leveraging multiple weak preference models, `MPAW` eliminates the need for computationally expensive fine-tuning of large models for each preference configuration. The complexity of our method is reduced to $O(\text{Models})$ instead of $O(\text{Preferences})$.

▷ **Enhanced Control with Logits Tuning.** We extend the concept of logit-level tuning by leveraging multiple small preference models, each specializing in different aspects of human preference. This approach provides granular control over the generated text, achieving a more nuanced and robust alignment across diverse criteria, reaching close to full fine-tuning of strong models.

2 PRELIMINARIES

Modeling of Reward Function Let π_{ref} be a reference LLM to align with human feedback. For each sampled question x, π_{ref} is prompted to produce pairs of answers (y_1, y_2); in particular, y_1 is preferred than y_2, the information of which is provided by human feedback. This procedure yields a preference dataset $\mathcal{D} = \{(x, y_1, y_2)\}$. The underlying preference is assumed to follow a Bradley-Terry (Bradley & Terry, 1952). Precisely,

$$\mathcal{R}^* = \arg\min_{\mathcal{R}_\phi} \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}[\sigma(\mathcal{R}(y_1|x) - \mathcal{R}(y_2|x))], \tag{1}$$

where σ refers to sigma function and $\mathcal{R}^*$ is considered a ground-truth reward function.

Reinforcement Learning from Human Feedback (Bai et al., 2022; Schulman et al., 2017) After obtaining $\mathcal{R}^*$, the task of alignment is performed by optimizing

$$\arg\max_{\pi_\theta} \mathbb{E}_{(x,y)\sim\pi_\theta}[\mathcal{R}^*(y|x)] - \beta D_f(\pi_\theta, \pi_{\text{ref}}), \tag{2}$$

where π_θ is initialized as π_{ref} and D_f is any f-divergence defined as

$$D_f(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y)\sim\pi_{\text{ref}}} f\left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right). \tag{3}$$

Recall that $f : [0, \infty) \to \mathbb{R}$ is assumed to be convex and $f(1) = 0$. If $f(x) = x\log x - x + 1$,[1] which corresponds to the reverse KL divergence, $D_f(\pi_\theta, \pi_{\text{ref}}) = D_{KL}(\pi_\theta|\pi_{\text{ref}})$.

Generalized Direct Preference Optimization (Rafailov et al., 2024b; Wang et al., 2023) Assume that $f'(z) \to -\infty$ as $z \to 0$ and f is strongly convex, the problem Eq. 2 can be solved analytically as

$$\pi^*(y|x) = \pi_{\text{ref}}(y|x)(f')^{-1}\left(\frac{\mathcal{R}^*(y|x) - Z(x, \mathcal{R}^*(\cdot|x))}{\beta}\right), \tag{4}$$

[1]Note that $x\log x - x + 1$ and $x\log x$ produce the same divergence.

where $Z(x, \mathcal{R}^*(\cdot|x))$ is a normalization factor ensuring that

$$\sum_y \pi_{\text{ref}}(y|x)(f')^{-1}\left(\frac{\mathcal{R}^*(y|x) - Z(x, \mathcal{R}^*(\cdot|x))}{\beta}\right) = 1. \tag{5}$$

In particular, we have

$$\mathcal{R}^*(y|x) = \beta f'\left(\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)}\right) + Z(x, \mathcal{R}^*(\cdot|x)). \tag{6}$$

Substituting this equality into the problem Eq. (1), an objective that is considered more stable to optimize can be expressed as

$$\arg\min_{\pi_\theta} \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}\sigma}\left(\beta f'\left(\frac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right) - \beta f'\left(\frac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)\right). \tag{7}$$

Multi-Objective Alignment (Wu et al., 2023) In this scenario, there are M reward functions $(\mathcal{R}_1^*, \mathcal{R}_2^*, \ldots, \mathcal{R}_M^*)$, each of them is built upon a preference dataset denoted by $\mathcal{D}_i$. Each user may have their own preferences $w \in \Delta^{M-1} = \{w \in \mathbb{R}^M \mid \sum_i w_i = 1 \text{ and } w_i \geq 0\}$.

3 METHODS

3.1 PROBLEM

Let a user preference $w \in \Delta^{M-1}$ be given, a naive approach to align models using multiple reward functions is: i) $\mathcal{R}_w^* \leftarrow \sum_i w_i \mathcal{R}_i^*$, and then replace $\mathcal{R}^*$ by $\mathcal{R}_w^*$ in the problem equation 2. However, the drawback is clear: if there are P user preferences (typically, $P \gg M$), we have to finetune models P times. Note that the most time-consuming step in RLHF is the finetuning process. To be more computationally efficient, we focus on the scenario where there are only M finetuned small/weak models, each of them was aligned using individual reward functions, and our goal is to align large/strong models in a finetuning-free manner. Precisely, we perform alignment while generating answers to any given prompt x. Therefore, the complexity of our method is $O(M)$ instead of $O(P)$.

Suppose we have a set of policies $\{(\pi_i^*, \pi_i^{\text{ref}})\}_{i=1}^M$. Each aligned policy π_i^* is initialized as π_i^{ref} and the alignment was done using a reward function $\mathcal{R}_i^*$ under f-divergence regularization. These aligned policies are supposed to be small and thus weak. Our goal is to align a **stronger** model π_{base} using $\{(\pi_i^*, \pi_i^{\text{ref}})\}_{i=1}^M$ for every user preference $w \in \Delta^{M-1}$ to optimize Eq. (2). Given a user preference w, a weighted reward is defined as $\mathcal{R}_w^* = \sum_i w_i \mathcal{R}_i^*$. Inspired by Shi et al. (2024), integrating Eqs. (6) into (4), we have

$$\begin{aligned}
\pi_w^*(y|x) &= \pi_{\text{base}}(y|x)(f')^{-1}\left(-\frac{Z_w^*(x, \mathcal{R}_w^*(\cdot|x))}{\beta} + \frac{1}{\beta}\sum_{i=1}^M w_i \mathcal{R}_i^*\right) \\
&= \pi_{\text{base}}(y|x)(f')^{-1}\left(-\frac{Z(x, w, \{\mathcal{R}_i^*(\cdot|x)\}_{i=1}^M)}{\beta} + \sum_{i=1}^M w_i f'\left(\frac{\pi_i^*(y|x)}{\pi_i^{\text{ref}}(y|x)}\right)\right),
\end{aligned} \tag{8}$$

where

$$Z(x, w, \{\mathcal{R}_i^*(\cdot|x)\}_{i=1}^M) = Z_w^*(x, \mathcal{R}_w^*(\cdot|x)) - \sum_{i=1}^M w_i Z_i^*(x, \mathcal{R}_i^*(\cdot|x)). \tag{9}$$

3.2 DECODING ALGORITHM

Based on beam search, we employ π_{base} to generate a list of candidate answers for each question x, where $\pi_w^*(y|x)$ in (8) is treated as a score function that select the top-K candidates at each iteration. In general, $Z(x, w, \{\mathcal{R}_i^*(\cdot|x)\}_{i=1}^M)$ is intractable to evaluate. Nevertheless, such an issue can be

Algorithm 1: `MPAW`

Input: input x, beam width W, expansion factor κ, chunk length L, model to align π_{base}, alignment pairs $\{(\pi_i^*, \pi_i^{\text{ref}})\}_{i=1}^M$, weights $\{w_i\}_{i=1}^M$

Output: Optimal answer y to x

Initialize beam $\mathcal{B} \leftarrow \{(x, y' = \emptyset)_i\}_{i=1}^W$

while $\exists (x, y') \in \mathcal{B}$ *incomplete* **do**

  $\mathcal{C} \leftarrow \emptyset$

  foreach $(x, y') \in \mathcal{B}$ *incomplete* **do**

    $\mathcal{Y} \leftarrow \text{Sample}_\kappa(y_L)_i \overset{\text{i.i.d.}}{\sim} \pi_{\text{base}}(\cdot|x, y')$ ($|y_L| = L$, independent and identically distributed)

    $\mathcal{C} \leftarrow \mathcal{C} \cup \{(x, \text{cat}(y', y_L) : y_L \in \mathcal{Y}\}$

  $\mathcal{B} \leftarrow \underset{(x,y')\in\mathcal{C}}{\text{SelectTop-}W} \left[\pi_{\text{base}}(y'|x) \exp\left(\sum_{i=1}^M w_i \log \frac{\pi_i^*(y'|x)}{\pi_i^{\text{ref}}(y'|x)}\right)\right]$

return $\arg\max_{(x,y)\in\mathcal{H}} \pi_{\text{base}}(y|x) \exp\left(\sum_{i=1}^M w_i \log \frac{\pi_i^*(y'|x)}{\pi_i^{\text{ref}}(y|x)}\right)$

Table 1: Overall comparison with other baselines. The number of user preferences (same amount as weights) is much larger than the number of objectives.

Algorithms	Free from RM	Number of trained models	Requirement
MORLHF (Wu et al., 2023)	✗	# preferences	
MODPO (Zhou et al., 2024b)	✔	# preferences	
RS (Rame et al., 2023; Jang et al., 2023)	✔	# objectives	same arch. & init.
DPA (Wang et al., 2024a), CPO (Guo et al., 2024), RiC (Yang et al., 2024)	✗	1	SFT
MOD (Shi et al., 2024)	✔	# objectives	same tokenizer
`MPAW`	✔	# objectives	

circumvented if we use KL divergence, which corresponds to $f(x) = x \log x - x + 1$. In this case,

$$\begin{aligned}\pi_w^*(y|x) &= \pi_{\text{base}}(y|x) \exp\left(-Z(x, w, \{\mathcal{R}_i^*(\cdot|x)\}_{i=1}^M) + \sum_{i=1}^M w_i \log \frac{\pi_i^*(y|x)}{\pi_i^{\text{ref}}(y|x)}\right) \\ &\propto \pi_{\text{base}}(y|x) \exp\left(\sum_{i=1}^M w_i \log \frac{\pi_i^*(y|x)}{\pi_i^{\text{ref}}(y|x)}\right).\end{aligned} \tag{10}$$

Therefore, sticking to KL divergence, we can discard $Z(x, w, \{\mathcal{R}_i^*(\cdot|x)\}_{i=1}^M)$ without losing any information in decoding π_{base}. Our proposed method is summarized in Algorithm **??**. Plus, we outline the comparison with the baselines in Table 1.

Proxy Decoding: Logits-based Alignment While the previous section focused on optimizing the entire generation process at the trunk level, we can further refine alignment by adjusting the output probabilities at each decoding step, known as logits-based alignment. This offers granular control over text generation. Building on the observation that even a single small preference model can effectively guide a large language model via logits tuning (Liu et al., 2024a), we explore leveraging multiple small preference models for enhanced multi-objective alignment.

Take model alignment as a model state transfer from π_{ref} to $\pi_{\text{ref}*}$. Let's explore to study the pattern for another model π_{base}. From Eq. (4), we can get

$$\underbrace{\pi_{\text{ref}*}(y|x)}_{\text{transferred}} = \pi_{\text{ref}}(y|x)(f')^{-1}\left(\frac{\mathcal{R}^*(y|x)}{\beta} - \frac{Z(x, \mathcal{R}^*(\cdot|x))}{\beta}\right),$$

To transfer this alignment to π_{base} (or multi-objective π_w), we can apply a similar transformation:

$$\pi_{\text{base}^*} = \underbrace{\pi_{\text{base}} \frac{\pi_{\text{ref}^*}}{\pi_{\text{ref}}}}_{\text{transferred}} = \pi_{\text{base}} (f')^{-1} \left(\frac{\mathcal{R}^*(y|x)}{\beta} - \frac{Z(x, \mathcal{R}^*(\cdot|x))}{\beta} \right), \tag{11}$$
$$\pi_{w*} \propto \pi_{\text{base}} (\frac{\sum_i w_i \mathcal{R}_i^*}{\beta}) = \pi_{\text{base}} \prod_i \left(\frac{\pi^i_{\text{ref}*}}{\pi^i_{\text{ref}}} \right)^{w_i},$$

Take softmax as the activation function, $\pi(y_i|x) = \text{softmax}(\text{logits}_i) = \frac{\exp(\text{logits}_i)}{\sum_j \exp(\text{logits}_j)}$. Then

$$\text{logits}_{\text{w}^*} = \text{logits}_{\text{base}} + \beta(\sum_i w_i (\text{logits}_{\text{ref}^*} - \text{logits}_{\text{ref}})), \tag{12}$$

Specifically, when using KL divergence (where $f(x) = x \log x - x + 1$), the equation can be simplified to:

$$\pi_{\text{ref}*}(y|x) = \pi_{\text{ref}}(y|x) \exp\left(\frac{\sum_i w_i \mathcal{R}_i^*}{\beta} \right) \cdot \left(\sum_y (\pi_{\text{ref}(y|x)} \exp(\frac{\sum_i w_i \mathcal{R}_i^*}{\beta})) \right). \tag{13}$$

During decoding, $\pi_{\text{ref}(y|x)}$, $\mathcal{R}^*(y|x)$, β are fixed for a given context and preference model. This allows the validity of logits-based alignment.

4 EXPERIMENT

Following the protocol established in Shi et al. (2024), we evaluate the proposed logits-based alignment method on two distinct reasoning tasks: GSM8K-COT (denoted as *math*) and Codex@1 (denoted as *code*). Our optimization adopts a weighting mechanism where w_{code} : w_{math} ratios provide the alignment focus.

As shown in Figure 1, our method achieves 70.58% mathematical accuracy under 0.2:0.8 weighting configuration, surpassing the proxy tuning baseline (69.60%) while approaching full fine-tuning performance (72.02%).

Method	Code pass@1	Math Acc.
Base	46.31	1.52
Directly Tuned	65.7	72.02
`MPAW` (1,0)	43.69	67.32
`MPAW` (0.5,0.5)	45.49	68.99
`MPAW` (0.2,0.8)	-	70.58
`MPAW` (0.8,0.2)	44.91	-
`MPAW` (0,1)	46.10	69.60

Table 1: Multi-task alignment performance with varying code:math weight ratios. The weak model is Qwen2-0.5B (Qwen, 2024) fine-tuned on code/math. The strong model is Qwen2-7B.

Discussion `MPAW` achieves multi-preference/multi-domain generalization through dynamic integration of multiple reward models during the decoding phase, and can naturally extend to accommodate any number of optimization objectives. It is worth noticed that `MPAW`, a decoding-time method, can closely approach the performance of computationally intensive full fine-tuning methods is particularly noteworthy. This highlights the efficiency of weak model collaboration in guiding strong models towards desired behaviors without the need for extensive retraining. Furthermore, the observed trade-off between code and math performance as weights are adjusted provides empirical evidence for `MPAW`'s ability to navigate and optimize for competing objectives, a crucial aspect for real-world applications where user preferences are often multifaceted and complex.

5 RELATED WORK

Multi-objective LMs alignment In order to meet diverse human needs, various approaches have been proposed to simultaneously align language models (Bai et al., 2022) with multiple objectives, addressing trade-offs between different dimensions (Rame et al., 2023; Jang et al., 2023; Ji et al., 2023; Wang et al., 2024b; Badrinath et al., 2024; Khaki et al., 2024; Lee et al., 2024; Han et al., 2024). *Learning-based* algorithms align through gradient descent and optimize model parameters. MORLHF (Wu et al., 2023) use fine-grained reward models to personalize language models for diverse user needs. MODPO (Zhou et al., 2024b) introduces a novel multi-objective alignment framework by integrating learned reward representations, enabling the model to align with multiple

objectives on the initial preference dataset. Additionally, parameter merging offers a training-free solution, such as reward soup (Lin et al., 2024; Rame et al., 2023; Jang et al., 2023), which achieves multi-objective alignment by linearly combining model weights trained for individual objectives. Preference-conditioned *prompting* (Yang et al., 2024; Wang et al., 2024a; Guo et al., 2024), which directly incorporates preference weightings into prompts after a fine-tuning process, and *search-based* algorithms (Cui et al., 2024; Dong et al., 2023; Gao et al., 2022; Guo et al., 2024; Huang et al., 2024), which use graph-based methods, optimize multiple objectives at *inference* time. Search-based method can neglect KL diverge in Eq. (2) because the results of search only determines by order. While these methods are relatively efficient, they often rely on reducing the mis-specification hypothesis (Rame et al., 2023) or struggle with out-of-distribution generalization ability (Zhou et al., 2024d), posing challenges in interpretability and robustness.

Controllable Decoding Controllable decoding focuses on steering language model outputs in a direction that aligns with specific goals or constraints, including token-level and response-level control techniques. *Response-level* methods often treat the entire generated text as a sample from a probability distribution. To solve this, energy-based optimization approaches (Qin et al., 2022; Kumar et al., 2022) continuously optimize LLMs through gradients. *Token-level* methods, fine-tune output generation at each timestep to increase control and ensure the generated sequence aligns with desired outcomes. (Mudgal et al., 2024; Liu et al., 2024b) align through value models, while (Khanov et al., 2024; Zhou et al., 2024c) treat it as a search problem. (Liu et al., 2024c) introduce a distribution approximation per token. (Huang et al., 2024; Liu et al., 2024a; Zhao et al., 2024) operate token logits to control decoding.

Weak-to-Strong Generalization Weak-to-strong generalization (Burns et al., 2023) showcase a phenomenon that naively finetune superintelligence models by a weak model, they consistently perform better than their weak supervisors. For example, (Rafailov et al., 2024a; Zhou et al., 2024c) propose token-level weak-to-strong alignment. (Zhou et al., 2024a; Mitchell et al., 2023; Liu et al., 2024a; Huang et al., 2024) raise proxy or emulated fine-tuning strategies, which use the distributional differences between small tuned and untuned models to adjust the output of large language models, or even black-box ones. However, these *learning-based* approaches rely on shared vocabularies between small and large models, limiting their practical applications. *Searching-based* approach (Zhou et al., 2024c; Zhao et al., 2024) are not restricted by vocabularies.

6 Conclusion

We introduced `MPAW`, a framework for multi-objective alignment of LLMs. By leveraging weak models aligned with distinct reward functions and integrating them through a collaborative decoding strategy, `MPAW` efficiently adapts to user-defined preferences without the need for costly retraining. The ability to integrate heterogeneous weak models, coupled with the zero-shot adaptability to different preference weightings, offers significant advantages over to `MPAW`. Our experiments show that `MPAW` achieves high-quality alignment with multiple objectives while maintaining the generalization power of the base model. Future work could explore further optimizations to the framework, as well as its applicability to even more complex, real-world scenarios. The code is released at `https://github.com/NuoJohnChen/MPAW`.

Acknowledgement

This research / project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative, the AI Singapore Programme (AISG Award No: AISG2-TC-2021-002), and the MOE Academic Research Fund (AcRF) Tier 1 Grant in Singapore (Grant No. T1 251RES2315). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We thank Weida Li, Anningzhe Gao and Fei Yu for valuable mathematical suggestions and discussions that helped improve this work.

REFERENCES

Anirudhan Badrinath, Prabhat Agarwal, and Jiajing Xu. Hybrid preference optimization: Augmenting direct preference optimization with auxiliary objectives. *CoRR*, abs/2405.17956, 2024. doi: 10.48550/ARXIV.2405.17956. URL https://doi.org/10.48550/arXiv.2405.17956.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL https://arxiv.org/abs/2312.09390.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2024. URL https://openreview.net/forum?id=pNkOx3IVWI.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11275–11288, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.754. URL https://aclanthology.org/2023.findings-emnlp.754.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment, 2024. URL https://arxiv.org/abs/2402.19085.

Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. Value augmented sampling for language model alignment and personalization, 2024. URL https://arxiv.org/abs/2405.06639.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models, 2024. URL https://arxiv.org/abs/2404.11045.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023. URL https://arxiv.org/abs/2310.11564.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023. URL https://arxiv.org/abs/2307.04657.

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. *ArXiv*, abs/2402.10038, 2024. URL https://api.semanticscholar.org/CorpusID:267682216.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *ArXiv*, abs/2402.01694, 2024. URL https://api.semanticscholar.org/CorpusID:267411977.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-based constrained sampling from language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 2251–2277. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.144. URL https://doi.org/10.18653/v1/2022.emnlp-main.144.

Kyungjae Lee, Dasol Hwang, Sunghyun Park, Youngsoo Jang, and Moontae Lee. Reinforcement learning from reflective feedback (rlrf): Aligning and improving llms via fine-grained self-reflection. *ArXiv*, abs/2403.14238, 2024. URL https://api.semanticscholar.org/CorpusID:268553509.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of rlhf, 2024. URL https://arxiv.org/abs/2309.06256.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy, 2024a. URL https://arxiv.org/abs/2401.08565.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding, 2024b. URL https://arxiv.org/abs/2309.15028.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models, 2024c. URL https://arxiv.org/abs/2402.02992.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D. Manning. An emulator for fine-tuning large language models using small language models, 2023. URL https://arxiv.org/abs/2310.12962.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. Controlled decoding from language models, 2024. URL https://arxiv.org/abs/2310.17022.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. COLD decoding: Energy-based constrained text generation with langevin dynamics. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/3e25d1aff47964c8409fd5c8dc0438d7-Abstract-Conference.html.

Qwen. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024a. URL https://arxiv.org/abs/2404.12358.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024b. URL https://arxiv.org/abs/2305.18290.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lSbbC2VyCu.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hannaneh Hajishirzi, Noah A. Smith, and Simon Du. Decoding-time language model alignment with multiple objectives, 2024. URL https://arxiv.org/abs/2406.18853.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints, 2023. URL https://arxiv.org/abs/2309.16240.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.468. URL https://aclanthology.org/2024.acl-long.468.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Oluwasanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. *ArXiv*, abs/2402.00742, 2024b. URL https://api.semanticscholar.org/CorpusID:267365201.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=CSbGXyCswu.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *CoRR*, abs/2402.10207, 2024. doi: 10.48550/ARXIV.2402.10207. URL https://doi.org/10.48550/arXiv.2402.10207.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. Weak-to-strong jailbreaking on large language models, 2024. URL https://arxiv.org/abs/2401.17256.

Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang, Wanli Ouyang, and Yu Qiao. Emulated disalignment: Safety alignment for large language models may backfire!, 2024a. URL https://arxiv.org/abs/2402.12343.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10586–10613, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.630. URL https://aclanthology.org/2024.findings-acl.630.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. Weak-to-strong search: Align large language models via searching over small language models, 2024c. URL https://arxiv.org/abs/2405.19262.

Zhaoyi Zhou, Chuning Zhu, Runlong Zhou, Qiwen Cui, Abhishek Gupta, and Simon Shaolei Du. Free from bellman completeness: Trajectory stitching via model-based return-conditioned supervised learning. In *The Twelfth International Conference on Learning Representations*, 2024d. URL https://openreview.net/forum?id=7zY781bMDO.