
Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging

Joel Jang^{1,2} Seungone Kim³
Bill Yuchen Lin² Yizhong Wang¹ Jack Hessel² Luke Zettlemoyer¹
Hannaneh Hajishirzi^{1,2} Yejin Choi^{1,2} Prithviraj Ammanabrolu⁴

¹University of Washington ²Allen Institute for AI
³Carnegie Mellon University ⁴UC San Diego
joeljang@cs.washington.edu

Abstract

While Reinforcement Learning from Human Feedback (RLHF) aligns Large Language Models (LLMs) with general, aggregate human preferences, it is suboptimal for learning diverse, individual perspectives. In this work, we study Reinforcement Learning from Personalized Human Feedback (RLPHF) problem, wherein LLMs are aligned to multiple (sometimes conflicting) preferences by modeling alignment as a Multi-Objective Reinforcement Learning (MORL) problem. Compared to strong single-objective baselines, we show that we can achieve **personalized alignment** by decomposing preferences into multiple dimensions. These dimensions are defined based on personalizations that are declared as desirable by the user. In this work, we show that they can be efficiently trained independently in a distributed manner and combined effectively post-hoc through parameter merging.¹

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) [2, 3, 7, 23, 25] typically optimizes a policy model that receives training signals from a single reward model that aims to capture the *general* preferences of a population. In this work, we instead propose Reinforcement Learning from *Personalized* Human Feedback (RLPHF), a new, multi-objective formulation of the human preference alignment problem, where Large Language Models (LLMs) are trained to be efficiently aligned with a range of different, potentially personalized combinations of human preferences.

We model RLPHF as a Multi-Objective Reinforcement Learning (MORL) problem, which allows training the policy model with multiple, *conflicting* objectives since it aims to vary the importance of each objective during inference. In existing RLHF formulations, pairwise human feedback is collected by asking human annotators to choose which model response is *generally* better and is used to train a general reward model. This makes *implicit* assumptions that may not hold for everyone. For example, recent work has shown that LLMs aligned with RLHF prefer verbose output generations [7, 35, 45, 55]. We aim to support a wider range of multifaceted preferences that are explicitly declared as desirable by the user—giving the user control over the facets of output text they want to see as well as the personal data they wish to reveal to the model. We collect *personalized* human feedback corresponding to multiple such dimensions, noting that they may also be conflicting.

We first implement a strong MORL baseline called PROMPTED-MORL where there are multiple reward signals for each of the objectives (preferences) given via prompts during RL training. Next,

¹Code: <https://github.com/joeljang/RLPHF>

we propose PERSONALIZED SOUPS, a method that circumvents simultaneously optimizing multiple preferences by first optimizing multiple policy models each with distinct preferences with Proximal Policy Optimization (PPO) and merging the parameters of the policy models whose preferences we want to composite together on the fly during inference. This *modular* approach significantly reduces the computational complexity from exponential to linear in relation to the total number of unique preferences. Furthermore, since PERSONALIZED SOUPS does not have to be trained in a multitask fashion, it does not require *re-training* the policy every time a novel preference (objective) is added.

We empirically show that by transforming the problem of aligning LLMs to human preferences into a MORL problem, we are able to have **personalized alignment** that provides a deeper level of adaptation to individual users that supervised fine-tuning, RLHF, and prompting cannot attain. We also emphasize the modularity of PERSONALIZED SOUPS by performing experiments in a scenario where the user additionally writes novel preferences that they want to integrate with existing preferences. We show that in this scenario, PERSONALIZED SOUPS still performs competitively to PROMPTED-MORL while being exponentially more efficient through parameter merging.

2 Reinforcement Learning from *Personalized* Human Feedback

The current RLHF can be denoted as optimizing policy π :

$$\pi^* = \arg \max_{\pi} (R(x, \pi(x))) \tag{1}$$

where R is the reward model trained on *general* human feedback. As pointed out in Silver et al. [34], the following optimization may implicitly be occurring under the hood:

$$\pi^* = \arg \max_{\pi} (r_1 + r_2 + \dots + r_n) \tag{2}$$

where r_i represents rewards from objectives that human annotators may *generally* consider positive objectives (e.g., informative, helpful, kind, etc.) and n is the total number of unique ‘dimensions’ of these positive objectives. This formulation does not allow modeling *conflicting* objectives, which may occur in real-world scenarios. For example, some people may prefer concise and unpretentious responses in contrast to informative, polite responses.

In this section, we formalize RLPHF where we allow modeling conflicting preferences during alignment. We explain how we collect *conflicting* feedback in Section 2.1. In Appendix A, we explain how we convert the current RLHF formulation into a MORL problem. Lastly, we explain the details of our evaluation in Appendix E.1.

2.1 Personalized Reward and Feedback

Collecting *Conflicting* Pairwise Feedback We utilize Tulu-7B LM [45], a model that uses LLaMA-7B [41] as a base model and is instruction tuned on a mixture of open-source instruction-tuning datasets, as the base model for our experiments. We utilize 10k prompt instances from GPT4-Alpaca [27], one of the datasets used to train Tulu-7B, as our instruction dataset D_{train} to generate rollouts and collect pairwise feedback data. We also use the same D_{train} during Proximal Policy Optimization (PPO) training [32] of Tulu-7B.

Following previous work, we simulate human annotators with GPT-4 for collecting large-scale pairwise feedback data [3, 7]—but note that our evaluations are validated with (smaller-scale) human preference data collected from crowdworkers. While Dubois et al. [7] mostly simulates GPT-4 and other LLMs to choose which is *generally* a better response between two candidate responses, we provide GPT-4 with a single preference (full list shown in Table ??) to decide which is a better response. We also provide the same preference criteria via additional prompts during the rollout generation of the two candidate responses; we use Tulu-30B for the rollout generation while the actual policy model we train is Tulu-7B for our main experimental setup, making our experimental setting an off-policy training set-up.

Reward Model Training While we have feedback on which of the two model responses is more aligned with a single preference via GPT-4 annotation, utilizing only two *positive* pairs during reward model training was empirically shown to be less robust during the PPO training. Instead,

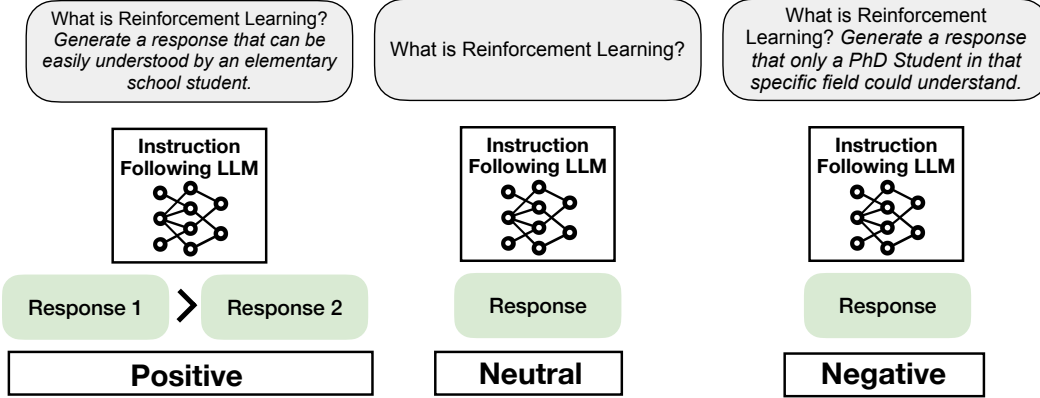


Figure 1: Reward model training with conflicting preferences.

we train our reward model on multiple comparisons [15, 36] by including a *neutral* response and a *negative* response as shown in Figure 1. Specifically, the reward model is provided with four different comparisons for a single prompt during training: positive 1 > positive 2 (decided by GPT-4), positive > neutral, positive > negative, and neutral > negative. The positive response when compared with the neutral and the negative response is chosen randomly. This allows the reward model to be exposed to different granularity of the specific preference and give scores accordingly during PPO training. We explore (1) training a single reward model in a multitask fashion that leverages the preference prompts during inference to give distinct rewards according to each preference and (2) training multiple reward models, each tailored to the distinct preference.

2.2 Multi-Objective Reinforcement Learning (MORL)

The MORL problem can be denoted as:

$$\pi^* = \arg \max_{\pi} (w_1 r_1 + w_2 r_2 + \dots + w_n r_n) \quad (3)$$

where π^* is the policy model that aims to maximize multiple objectives from the rewards r_1, r_2, \dots, r_n and w_1, w_2, \dots, w_n is the importance placed on each objective. If w_1, w_2, \dots, w_n are constants during PPO training, this problem essentially becomes a single objective problem, maximizing towards a single, *general* objective. In our setup, we have *conflicting* preferences which require dynamically varying w_i in a binary manner with respect to the conflicting preference during training and inference.

PROMPTED-MORL (P-MORL) First, we introduce a strong baseline that varies w_i during MORL through *prompts*. While w_i are given as inputs directly to the policy model in traditional RL settings using PPO with MORL, there is no straightforward way of integrating different w_i as an input to LLMs. Instead, we utilize the preference prompts as binary signals for w_i .

We append the unique preference combination (shown in Table ??) with a training prompt t from D_{train} $\{t + P1 + P2 + P3\}$ ² before feeding it to our initial policy model and getting the output response. Then, we gather reward signals for each of the preferences by feeding $\{t + P1/P2/P3 + \text{output}\}$ into a single reward model (doing three forward passes)³ to get the reward signal specific to the individual preference and averaging the three reward values to get a single scalar reward used for PPO training. We multitask train the policy model across the eight different unique combinations of preferences, which essentially results in varying w_i .

While PROMPTED-MORL can be a clear baseline for converting the alignment problem into a MORL problem, we propose another approach that does not have to see all existing preference combinations

² $P1, P2, P3$ each represents preference prompts from each preference dimension in Table ?. For one example, one unique combination might be P1A + P2B + P3A (ABA) where the combined objective for the response needs to be elementary level, informative, and friendly.

³Empirically, utilizing a single reward model instead of multiple reward models led to better performance. We hypothesize this is due to the problem of normalizing signals from different reward models [10], which is known to be a nontrivial problem in MORL.

during training thus allowing increasing the total number of distinct preferences at scale, which is required for true personalization.

PERSONALIZED SOUPS (P-SOUPS) We decompose the MORL problem into multiple single-objective problems:

$$\pi_1^* = \arg \max_{\pi} (r_1), \pi_2^* = \arg \max_{\pi} (r_2), \dots, \pi_n^* = \arg \max_{\pi} (r_n) \quad (4)$$

where we optimize each policy π_n individually. Then during inference, we pick and choose the policy models whose objective we want to maximize together and perform a weighted sum of the parameters on the fly:

$$\pi^* = w_1 \pi_1^* + w_2 \pi_2^* + \dots + w_n \pi_n^* \quad (5)$$

where $\sum_1^n w_n = 1$. This means that even though the exact preference combinations haven't been seen during training, we are still able to composite them on the fly during inference. While the total computational complexity increases exponentially when we are required to observe all possible combinations, optimizing individual objectives separately only increases the complexity in linear space. This also means that multitask training is not necessary and allows efficient integration of novel preferences. Since *personalization* also entails that there can be an infinite number of new preference dimensions, we assert that PERSONALIZED SOUPS makes tackling RLPHF feasible.

2.3 Multifaceted Evaluation

Due to space limits, we explain about our evaluation method in Appendix E

3 Experiments

3.1 Baseline Methods

Method	Reward Model	Policy Model	Training
Vanilla Baseline	-	single	-
Traditional RLHF	single	single	single
Preference Prompting	-	single	-
Multi-task Training	single	single	multitask
Prompted MORL	single/multi	single	multitask
PERSONALIZED SOUPS	multi	multi	single

Table 1: Components of different methods.

The summary of the key differences of baselines against our proposed methods is provided in Table 1.

Vanilla Baseline (VB) As the most simple baseline, we simply utilize the base Tulu-7B model to generate responses without providing it any notion of preferences. During the evaluation, we use the same response to evaluate on the 8 different preference combinations.

Reinforcement Learning from Human Feedback (RLHF) We perform RLHF in the traditional manner where GPT-4 labels which response is *generally* better, train a reward model using the pairwise feedback data, and use the reward model to adapt the policy model with PPO training. The same 10k instances from D_{train} are used for RLHF.

Preference Prompting (PP) We observe how far the instruction-tuned base LM can integrate multiple preference combinations by simply prompting for the preferences without any additional training.

Multi-task Training (MT) For a competitive baseline, we utilize the positive candidate selected by GPT-4 as the output for imitation learning, which is essentially performing rejection sampling [23] that uses GPT-4 as the reward model for selecting golden responses from the distribution of responses. We append the individual preference prompt with instances from D_{train} and multitask train the Tulu-7B model across all six individual preferences. This method also allows *distilling* the outputs of Tulu-30B for training Tulu-7B.

Method	VB	RLHF	PP	MT	P-MORL	P-SOUPS	Avg.
VB	-	44.52	45.95	43.00	46.58	40.14	44.04
RLHF	55.48	-	37.81	38.98	40.48	45.09	43.57
PP	54.05	62.19	-	48.08	49.09	45.00	51.68
MT	57.00	61.02	51.92	-	48.37	46.64	<u>52.99</u>
P-MORL	53.42	59.52	50.91	51.63	-	46.96	52.49
P-SOUPS	59.86	54.91	55.00	53.36	53.04	-	55.23

Table 2: Simulated pairwise win rate (%) across all methods using GPT-4.

Method	VB	RLHF	PP	MT	P-MORL	P-SOUPS	Avg.
VB	-	46.56	47.16	41.40	37.29	39.39	42.36
RLHF	53.44	-	52.73	44.22	38.18	44.97	46.71
PP	52.84	47.27	-	42.33	43.75	37.50	44.74
MT	58.60	55.78	57.67	-	43.48	45.45	52.20
P-MORL	62.71	61.82	56.25	56.52	-	56.21	58.70
P-SOUPS	60.61	55.03	62.50	54.55	43.79	-	<u>55.29</u>

Table 3: Pairwise win rate (%) across all methods through Human Evaluation.

3.2 Main Results

Table 2 and 3 show the results of doing all possible pairwise comparisons across the methods using GPT-4 and humans as judges, respectively. Note that the win rate of each battle is calculated using the aggregated win rate explained in Appendix E.1. Each individual preference combination results are shown in Appendix G. We also show the average *criteria-wise* win rate instead of the **aggregated** win rate across all of the methods in Appendix E.3.

Notably, there is a limitation to the extent prompting (PP) can integrate multiple preferences. This means that specific training for integrating the multiple preferences is necessary to composite them together. Next, we can see that supervised fine-tuning (RS) underperforms MORL-based methods, which is consistent with prior work that also showed the advantage of RL-based approaches when aligning LLMs with human feedback compared to its supervised-finetuning counterpart. Finally, while P-MORL and P-SOUPS both outperform other methods on average, there exists a discrepancy between the simulated and human evaluation; P-SOUPS has the highest average win rate in GPT-4 evaluation while P-MORL has the highest in human evaluation. Nonetheless, P-SOUPS is able to show superior performance in comparison to baseline methods and competitive performance to P-MORL. In previous parameter merging literature, multi-task fine-tuning (RS) used to be considered the upper bound for compositional abilities through parameter merging [6]. However, in our scenario, we can see that parameter merging (P-SOUPS) is able to *outperform* multitask fine-tuning. Due to space limits, we added our extended analyses of P-MORL and P-SOUPS regarding the **Trade-off between general helpfulness and personalization and scalability to new preferences** in Appendix B.

4 Conclusion

Previous work has shown that adapting LLMs with RLHF helps them generate outputs that are preferred by humans over the supervised fine-tuned counterpart. However, recent work has also pointed out that simply training LLMs to abide by the preference of the general may result in ignoring individual preferences and values. In this work, we provide the first steps to tackle this issue by proposing Reinforcement Learning from *Personalized* Human Feedback as a multi-objective problem so that LLMs can be aligned to follow *conflicting* preferences. We propose a promising method called P-SOUPS that is able to composite models trained on single objectives on the fly during inference. We also highlight the scalability of P-SOUPS by showing that it scales linearly, instead of exponentially like the MORL baseline, with regards to the number of new preferences, which is required to provide true personalization to individual users.

References

- [1] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.7. URL <https://aclanthology.org/2021.acl-long.7>.
- [2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John 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 B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL <https://api.semanticscholar.org/CorpusID:248118878>.
- [3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [4] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.
- [5] Deborah Cohen, Moonkyung Ryu, Yinlam Chow, Orgad Keller, Ido Greenberg, Avinatan Hassidim, Michael Fink, Yossi Matias, Idan Szpektor, Craig Boutilier, et al. Dynamic planning in open-ended dialogue using reinforcement learning. *arXiv preprint arXiv:2208.02294*, 2022.
- [6] Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. Cold fusion: Collaborative descent for distributed multitask finetuning. *arXiv preprint arXiv:2212.01378*, 2022.
- [7] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.
- [8] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, 2023.
- [9] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. *Blog post, April*, 1, 2023.
- [10] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.
- [11] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [12] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *ArXiv*, abs/2307.13269, 2023. URL <https://api.semanticscholar.org/CorpusID:260155012>.
- [13] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *ICLR*, 2022.

- [14] Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Exploring the benefits of training expert language models over instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14702–14729. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/jang23a.html>.
- [15] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. *arXiv preprint arXiv:2305.13735*, 2023.
- [16] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.
- [17] Noriyuki Kojima, Alane Suhr, and Yoav Artzi. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9:1303–1319, 2021.
- [18] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022.
- [19] Pan Li and Alexander Tuzhilin. Towards controllable and personalized review generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3237–3245, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1319. URL <https://aclanthology.org/D19-1319>.
- [20] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5976–5982, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1613. URL <https://aclanthology.org/D19-1613>.
- [21] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1298. URL <https://aclanthology.org/D18-1298>.
- [22] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.
- [23] Reiichiro Nakano, Jacob Hilton, S. Arun Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332, 2021. URL <https://api.semanticscholar.org/CorpusID:245329531>.
- [24] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis

- Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL <https://api.semanticscholar.org/CorpusID:246426909>.
- [26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [27] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- [28] Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*, 2022.
- [29] Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*, 2023.
- [30] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.
- [31] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [33] Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.
- [34] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.
- [35] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.
- [36] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.
- [37] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- [38] Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback. *arXiv preprint arXiv:2212.09710*, 2022.
- [39] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [40] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [42] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20:27–40, 2018.

- [43] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.
- [44] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pages 191–199. IEEE, 2013.
- [45] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023.
- [46] Mitchell Wortsman, Suchin Gururangan, Shen Li, Ali Farhadi, Ludwig Schmidt, Michael Rabbat, and Ari S Morcos. lo-fi: distributed fine-tuning without communication. *arXiv preprint arXiv:2210.11948*, 2022.
- [47] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.
- [48] Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.
- [49] Yuwei Wu, Xuezhe Ma, and Diyi Yang. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, 2021.
- [50] 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. *arXiv preprint arXiv:2306.01693*, 2023.
- [51] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pages 10607–10616. PMLR, 2020.
- [52] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL <https://aclanthology.org/2022.acl-long.356>.
- [53] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.
- [54] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.
- [55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.
- [56] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*, 2019.
- [57] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Related Work

Aligning Language Models To Human Preferences Incorporating human preference feedback into a reward model, and subsequently optimizing a language model to output text that reward model scores highly with an RL algorithm, has been shown to result in language models that generate outputs humans generally prefer [26]. This process has been applied to summarization [37, 48, 57], answering questions with long-form answers using text retrieved from the web [22, 24], generating engaging responses in a dialogue settings [5, 40] and following human instructions [15, 17, 38].

However, the standard RLHF setup commonly addressed in prior work assumes a reward model that accounts only for *average* annotator preference, i.e., the fact that different users may desire different outputs, even for the same prompt, is ignored Casper et al. [4]. Individual preferences can vary not only on aesthetic axes, but also on semantics. For example, Santurkar et al. [31] use public opinion polling to show that “default” LLM preferences vary in their degree of expressed-opinion alignment with different average opinions among demographic groups.⁴ Kirk et al. [16] defines a taxonomy and policy framework for the alignment of LLMs with personalized feedback. While Wu et al. [50] performs fine-grained RLHF which is very similar in spirit and allows personalization, our work develops MORL algorithms for scenarios where there are *conflicting* preferences, not only orthogonal objectives.

Multi-objective Reinforcement Learning (MORL) In this work, we propose formulating LLM personalization as a MORL problem, which was typically studied in decision-making tasks [10] that aims to tackle the problem of simply optimizing by a single, scalar, additive reward function [39], which possesses many limitations such as (1) suboptimal solutions due to lack of representation [10], (2) lack of explainability of distinct objectives, and (3) ensuring fair outcomes for multiple participants [33, 42].

Previous work has aimed to alleviate these problems through novel MORL methods [43, 44, 51, 53]. Other work aims to solve complex problems such as water management, military purchasing, wind farm control, etc. [10] by converting the single-objective RL problem into a MORL problem. In this work, we convert the problem of aligning LLMs to human preferences into a MORL problem to (1) provide a more optimal solution for each individual, (2) allow users to dynamically choose the distinct objectives they want to optimize, and (3) ensure fairness by allowing preferences that may be in the long-tail to be integrated.

Personalization in Natural Language Processing Personalization in Natural Language Processing (NLP) has mainly been focused on creating personalized dialogue agents [21, 49, 52, 54, 56], where the task is to create chitchat agents that are engaging with distinct personas based on user profile (e.g. gender, age, residence, etc.) or past user history data (e.g. Reddit posts, etc.). Another line of work [30] leverages personalized information to boost performance on specific tasks such as review generation [19], recipe generation [20], and headline generation [1]. This line of work requires model providers to make better models utilizing the personal information of the user. In our work, we propose a framework that allows users to choose which preference the language model should prefer, essentially giving control to the user.

Parameter Merging Recent work has shown that performing weighted linear interpolation of model parameters leads to the composition of each model ability [6, 12, 18, 46, 47]. This line of work has led to many interesting applications of model merging such as composing the abilities of expert models that perform different tasks [13, 14] and introducing language-specific modules for growing the total capacity of multilingual LMs [28].

Most recently, Rame et al. [29] proposed to merge policy models that were trained to perform specific tasks such as question answering and summarization using proxy reward models. While they mostly deal with reward models trained on the same data, our proposed MORL methods are an extension of this work that actually deals with diverse reward models trained on multifaceted human feedback to show compositional abilities through parameter merging rather than just ensembling.

⁴Feng et al. [8] suggests that “default” LLM expressed opinions stem directly from the pretraining data.

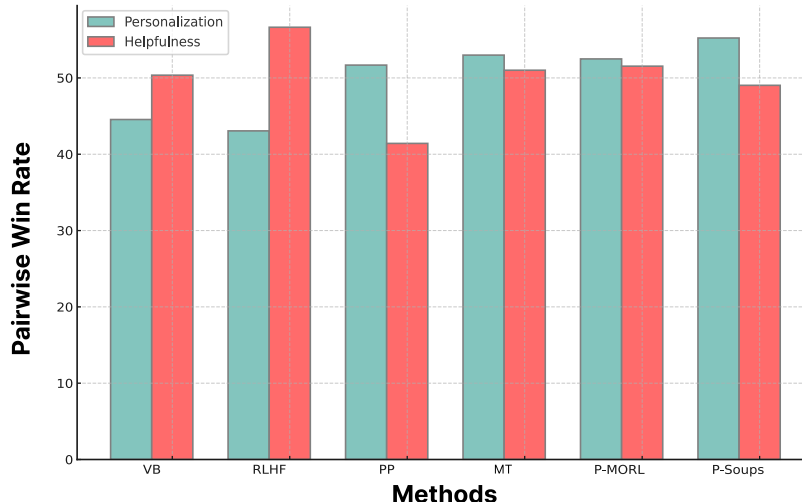


Figure 2: Average simulated pairwise win rate (%) across all the methods using personalization and helpfulness as the evaluation criteria.

B Extended: Experimental Results

B.1 Trade-off Between General *Helpfulness*

One might still wonder about the general ‘helpfulness’ capabilities of models that are trained to be tailored to multiple preferences. In Figure 2, we first show the average pairwise win rate from Table 2 in green. Next, we instead ONLY perform pairwise comparisons with the unseen objective ‘helpfulness’ (ask GPT-4 which model response they *generally* prefer better) and report the average pairwise win rate in red.

RLHF performs the best in this scenario, which shows that there is *no free lunch*; while the objective of RLHF was to provide model responses that are *generally* preferred (highly correlated with ‘helpfulness’), the other methods were prompted/trained to be optimized towards the personalized preference aspects, possibly deviating away from general helpfulness. While RS, P-MORL, and P-SOUPS are able to retain similar performance in terms of helpfulness compared to the initial instruction-tuned model (VB), we observe that prompting (PP) significantly underperforms compared to other methods which also highlights the limitation of simply prompting base/instruction-tuned models for personalized preferences and shows the need for specialized training methods for personalization.

B.2 Scaling to New Preferences

While we explore 6 distinct preferences in this work, we are still limited in doing ‘declarative’ personalization; that is, the individual preferences have been pre-defined to measure the performance of different methodologies. However, in the real world, individuals may not be bound by pre-defined preferences. Furthermore, people’s preferences might change over time, which requires continual learning of new preferences. This means that we may be required to train *infinite* numbers of preferences to be truly personalized to individuals’ preferences. Considering this aspect, the *scalability* of methods becomes a critical factor in implementing RLPHF in real-world scenarios.

In order to compare the scalability of P-MORL and P-SOUPS, we add two new preferences (in addition to the ones in Table ?? to the STYLE dimensions: (P3C) “Generate/Choose a response (that answers) in a sassy manner.” and (P3D) “Generate/Choose a response (that answers) in a sarcastic manner.”, which results in a total of 16 ($2 \times 2 \times 4$) unique preference combinations. We re-train P-MORL on the 16 new preference combinations and only train two new policy models for integrating P-SOUPS. The simulated win-rate between P-MORL and P-SOUPS on each of the original preference combinations (53.04% win rate of P-SOUP over P-MORL in Table 2 decomposed into each preference combinations) and the 16 new preference combinations are shown in Figure 3.

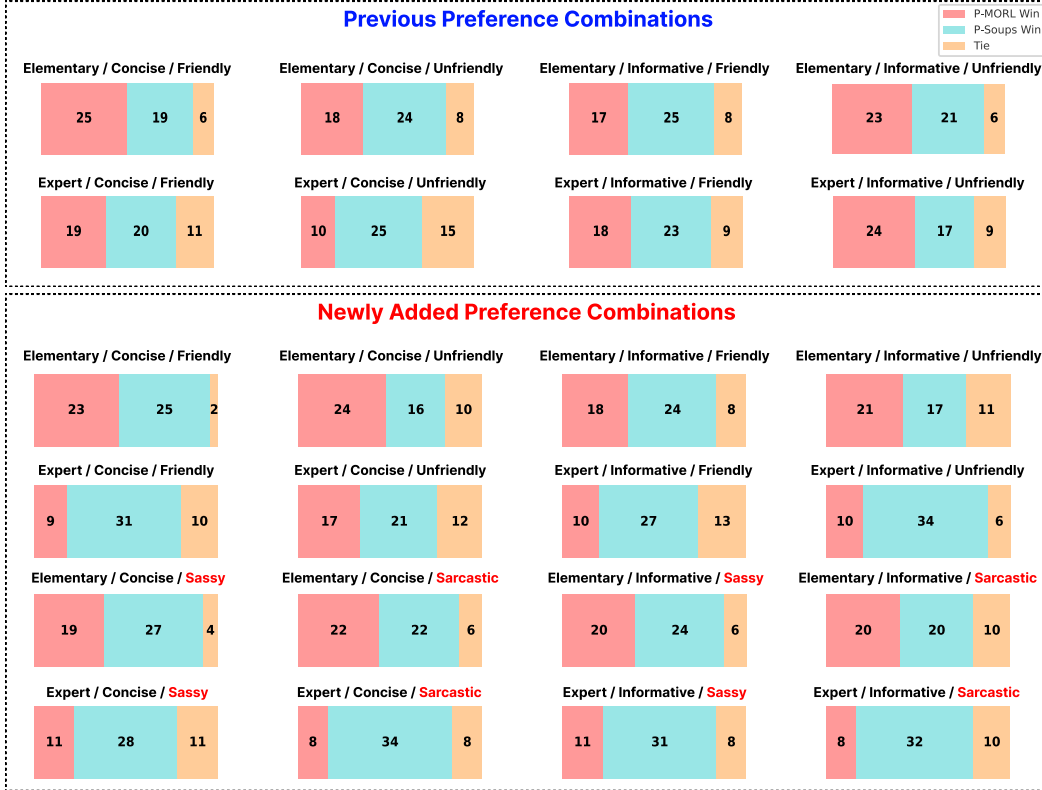


Figure 3: The *win / lose / tie* statistics for each of the *original* and *new* preference combination using GPT-4 as the evaluator. Getting the total $\frac{\#Wins}{\#Wins+\#Loss}$ for the 16 preference combinations results in **59.06%** win rate for P-SOUPS over P-MORL.

As shown in the figure, P-SOUPS shows competitive performance compared to P-MORL while being much more efficient considering that it (1) did not have to observe all 16 possible preference combinations and (2) did not have to re-train on the previous preferences, but just train two new policies each for the new preference in a modular manner and merge their parameters on-the-fly during inference. Considering that P-MORL is bounded by $O(2^n)$ while P-SOUPS is bounded by $O(n)$ where n is the total number of preferences (assuming there are two unique preferences for each dimension), we assert that P-SOUPS allows tackling RLPHF to be feasible.

C Extended: Experimental Details

C.1 Experimental Details

For both the reward model and policy training, we limit ourselves to going through D_{train} only once (1 epoch). In the initial exploration stage, the end performance for the policy model did not improve even if we trained the reward model for longer epochs. For policy model training, we utilize our evaluation dataset (50 prompts) to get the average reward and chose the policy model checkpoint that showed the highest average reward on the evaluation set for our final evaluation. We utilize LoRA [11] for both the reward model and policy model training. The detailed hyperparameters for the reward model and policy model training are provided in our github repository ⁵.

⁵<https://anonymous.4open.science/r/RLPHF-3F3F/>

Dimension	Preference Prompt	Symbol
Expertise	Generate/Choose a response that can be easily understood by an elementary school student.	P1A
	Generate/Choose a response that only a PhD Student in that specific field could understand.	P1B
Informativeness	Generate/Choose a response that is concise and to the point, without being verbose.	P2A
	Generate/Choose a response that is very informative, without missing any background information.	P2B
Style	Generate/Choose a response that is friendly, witty, funny, and humorous, like a close friend.	P3A
	Generate/Choose a response (that answers) in an unfriendly manner.	P3B

Table 4: List of 6 conflicting preferences divided into 3 distinct dimensions that are used for our experiments. We simulate a scenario where a person has a preference from each dimension, resulting in 8 unique combinations of preferences: AAA, AAB, ABA, ABB, BAA, BAB, BBA, and BBB.

D Additional Figure and Table Illustrating RLPHF and Conflicting Preferences

Please refer to Figure 4 and Table 4

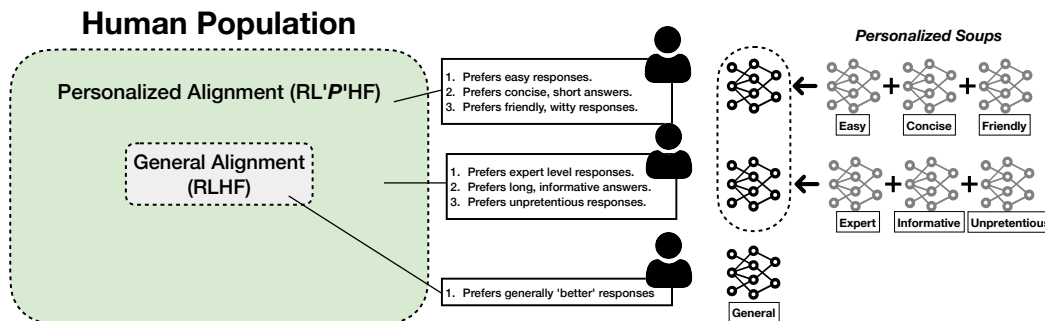


Figure 4: Current RLHF only tackles general alignment while RLPHF is able to take into account multifaceted human preferences, thus providing a more personalized form of model-to-human alignment.

E Details of Evaluation Setup

E.1 Multifaceted Evaluation

Evaluation For evaluation, we manually filter out 50 instances from the Koala evaluation [9] that require open-ended generations. We also modified some of the prompts so that the evaluation prompts do not contain any elements requiring individual preferences (e.g., removing the phrase asking for an elementary-level response from the original prompt since we want to test the LLM to generate an expert-level response). The full list of evaluation prompts used for our experiments is shown in Appendix F. In our evaluation setup, we simulate users to have a unique combination of preferences, each from the three preference dimensions (Expertise, Informativeness, Style) in Table 4, which equates to 8 unique preference combinations (examples shown in Figure 4). We get the average win rate across the simulated 8 preference combinations for our final evaluation. We use a variant of the AlpacaFarm evaluation framework for simulated (GPT4) evaluation and hire 24 crowdworkers for human evaluation. Details of human evaluation are provided in Appendix E.

Aggregated Win Rate When given an evaluation prompt p , we first generate responses from each model to get the outputs $\theta_A(p) = o_A, \theta_B(p) = o_B$ where θ_A is model A and θ_B is model B. The common approach is to get $H(p, o_A, o_B) = \{\text{WIN}, \text{TIE}, \text{LOSE}\}$ by asking the human which model response is *generally* preferred.

In our evaluation setup, we first assign scores to each of the possible feedback: WIN = 1, TIE = 0, LOSE = -1. Next, we iterate through the different preference dimensions and get an aggregated score value: $\sum_{i=1}^n H_i(p, o_A, o_B) = \text{score}$. Finally, we have WIN if $\text{score} > 0$, TIE if $\text{score} = 0$,

and LOSE if $score < 0$. To get the final win rate between θ_A vs. θ_B , we iterate through the entire evaluation set (50 prompts) \times the unique preference combinations (8 combinations) and get the total $\frac{\#Wins}{\#Wins+\#Loss}$ as the final win rate, while disregarding the total number of ties.

E.2 Simulated pairwise evaluation

We use a modified version of the GPT4 annotation prompt used by Dubois et al. [7]. We modify the criteria to perform the pairwise evaluation from *general* to a single preference dimension. We also provide 3 demonstrations: one scenario where there is a tie because both responses do not contain any notion of the preference (e.g. both responses do not show any signs of friendliness), one scenario where there is a clear winner, and one scenario where they are both good, but one is better than the other.

We recruited 24 crowd workers for our human evaluation. Figure 5 shows the interface used for human evaluation. We consider both the ‘Tie’ and ‘Both are bad’ options to be Ties.

E.3 Criteria-wise Evaluation

The criteria-wise win rate (%) across all the methods are shown in Figure 6. The criteria-wise win rate was calculated by getting the average win-rate of the preference combinations that contained the specific preference dimension. For example, when calculating the criteria-wise win rate of ‘Elementary’, we got the average win-rate of the preference combinations that contained the ‘Elementary’ preference, which includes AAA, AAB, ABA, and ABB.

F The full list of evaluation prompts

The full list of evaluation prompts used in our experiments are provided in Table 5.

G Detailed results for Human Evaluation and GPT-4 Evaluation

We provide detailed results (*win / loss / tie*) for each of the preference combinations for our main experimental results. Table 6 shows the GPT4 evaluation and Table 7 shows the human evaluation results.

H Examples of P-SOUPS text generations

Table 8 shows empirical examples of the text generated from each preference combination of the 16 preference combination experiments for the same prompt.

[b]

Instruction

List the layers of the TCP/IP model and for each layer identify one protocol.
<assistant>

ResponseA

The layers of TCP/IP model are as follows:

- 1- Application layer
- 2- Transport layer
- 3- Internet layer
- 4- Network layer
- 5- Data link layer

ResponseB

The TCP/IP model breaks down the internet into seven layers, like a cake! The first layer is the Physical Layer, which deals with the cables and electronics that physically connect things together. The second layer is called the Data Link Layer and makes sure that signals arrive at the right place and in the right order. The third is called Network Layered Layer. It's like a postman on a bike, who brings an envelope from A to B; here, the postman has to decide the best way to deliver the package and also how long it will take him to deliver. The fourth is called Transport Layer in charge of how data is packaged so that, basically, packages can be sent. The fifth is called Session Layer who decides how to keep all kinds of communication happening at the same time, like all kinds, songs and movies. Using different channels and asynchronous, so there can be many talks at once. The sixth is called Presentation Layer like being asked to dance by your crush. It makes sure the data is nice to look at and easy to understand, like if we can understand the dance without too much effort. Last is the Application Layer where we go to a fancy restaurant. So everyone is dancing, eating, and drinking. So, the final question, if you want to see what your dancing kid made, the best plan would be to see it from here, as high as you can.

[b]

Choose a response that can be easily understood by an elementary school student.

- A is better
- B is better
- Tie
- Both are Bad

Choose a response that is concise and to the point, without being verbose.

- A is better
- B is better
- Tie
- Both are Bad

Choose a response that is friendly, witty, funny, and humorous, like a close friend.

- A is better
- B is better
- Tie
- Both are Bad

Choose a response that is helpful (Make sure to put less focus on the actual factualness).

- A is better
- B is better
- Tie
- Both are Bad

Figure 5: Interface for human evaluation

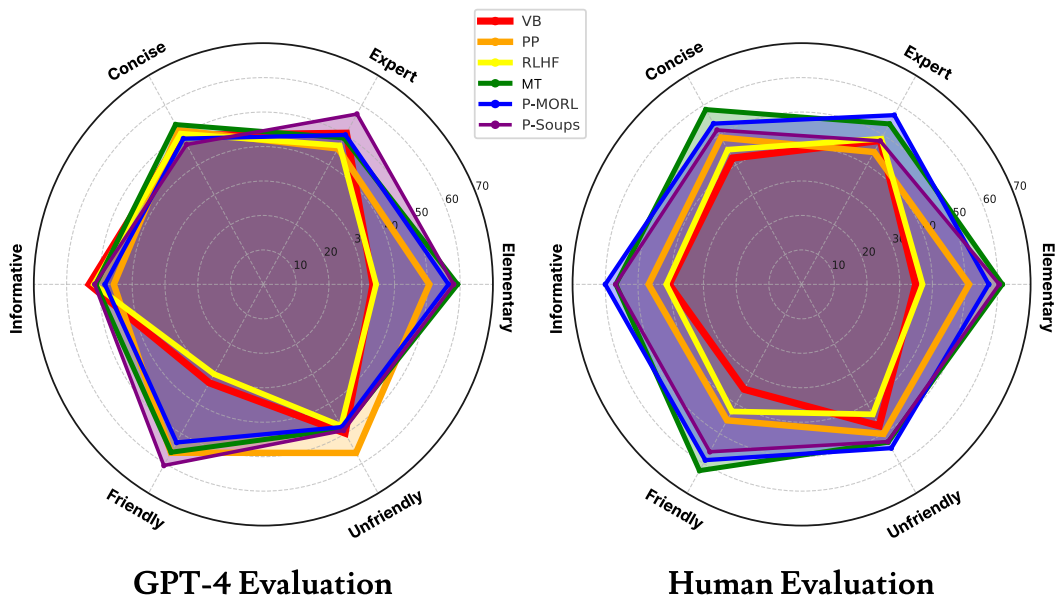


Figure 6: The criteria-wise win rate (%) across all methods through Human Evaluation and GPT-4 Evaluation.

ID	Prompt
1	Why can't bank use cash as capital as a buffer for potential losses?
2	Can you tell my a story about nuclear physics?
3	Create a message for an offer letter for an internship at Apple as a AI research intern.
4	Explain sarcoidosis to me.
5	"Give me a sample 5 day itinary for a switzerland holiday, starting from Basel"
6	Explain The Death of Ivan Ilych
7	"Why is it that only proteins, carbohydrates and fats are deemed to have caloric value?"
8	"Rank the following companies by how pro-consumer they are:Microsoft, Google, Nintendo, Sony, EA."
9	I'm planning to visit Okinawa Japan from April 7th to April 10th. Do you have any recommendation on what to do while I'm there?
10	How to improve instance segmentation AP metrics with human knowledge?
11	What is needed for self-sufficient living spaces?
12	Expand on the relation between inconsistency and ideological thinking.
13	Why do people give Reddit Gold to admins?
14	What does Nassim Nicholas Taleb say about market efficiency?
15	Can a boy and girl be just be best friends only ever?
16	What would be the reason for the popularity of youtube compared to other video sharing websites?
17	"Do you know something about the book "the art of thinking clearly" wrote by Rolf Dobelli?"
18	Antennas that transmit an equal amount of energy in the horizontal direction are called
19	"Hi, I have a question about MFCC (mel frequency cepstral coefficients). Are they the same thing as a MEL-spectrogram, or is there a difference?"
20	Why is it a bad idea to give a mouse a cookie?
21	How can anti-deressants make people think or act suicidally?
22	Create a lesson plan in two different levels: in CEFR A1 and A2 on the topic of friendship. Add a rubric.
23	Is online casino legal in India?
24	"How much of a threat is climate change in the coming years, and what should we do to stop it?"
25	Explain the basics of area and perimeter
26	What are the possible performance issues in a learning program on cross cultural communication?
27	Write description for 925 sterling silver miami cuban link chain.
28	What if people only valued and took pride in the things that were true about themselves?
29	I need to learn English could you help me and make a exercise?
30	Why does warmth make you feel tired?
31	Explain to me the Finite Element Method.
32	Introduce the "financial markets and institutions" by Frederic S. Mishkin
33	When are hops added to the brewing process?
34	Can a Liebherr LTM 11200-9.1 hypothetically lift Mount everest?
35	What are five important topics for game design?
36	What language does argentina people speak?
37	Is queue an ADT or a data structure?
38	What are some basics of nutrition that i should be aware of?
39	Can a qualifying manager work remotely abroad for US employer for an year be eligible for Eb1-C while on h1-b?
40	"I have competencies in remote sensing, machine learning, and water resource knowledge, what are the possible jobs I can occupy? What are the possible projects I can do? What companies I can work at?"
41	Is a banana a fruit or a herb
42	What are african governments doing to improve air traffic connectivity?
43	I want to open the developer tools in chrome with ctrl + shift + i on this website: https://mnsw.pro/ It doesnt work. works on other websites. even here. what is wrong?
44	"Consider this situation. Customer is ready to pay \$13 for my product but we are not able to raise invoice for less than \$20. So in this situation, act like a top sales person; closing expert; give me an email copy to upsell this customer with another \$13 product so that it can create a win win situation."
45	What are the important points for brand promotion and brand marketing?
46	What niches are not being fulfilled on the YouTube platform?
47	Explain TypeScript and Duck Typing.
49	How are carbon fibers used in buildings?
50	List the layers of the TCP/IP model and for each layer identify one protocol.

Table 5: The full list of 50 prompts used for evaluation.

Method	AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB	Total.	Avg Win-rate
VB vs RLHF	15/29/6	17/27/6	12/21/17	15/17/18	18/20/12	21/22/7	20/14/16	16/17/17	134/167/99	44.52%
VB vs PP	17/30/3	17/30/3	19/16/15	11/13/26	22/17/11	19/23/8	19/16/15	18/22/10	142/167/91	45.95%
VB vs MT	11/35/4	15/30/5	14/20/16	17/11/22	25/19/7	24/21/5	9/16/25	14/19/17	129/171/100	43.00%
VB vs P-MORL	20/30/0	19/30/11	14/18/12	12/17/21	18/23/9	25/20/5	20/11/19	15/15/20	143/164/93	46.73%
VB vs P-SOUPS	18/30/2	18/27/5	9/23/18	7/20/23	16/24/10	19/24/7	13/12/25	16/13/21	116/173/111	40.14%
RLHF vs PP	11/38/1	15/32/3	9/24/15	15/20/15	17/23/10	23/25/2	18/15/17	13/22/15	121/199/80	37.81%
RLHF vs MT	10/38/2	16/28/6	6/22/12	14/11/25	22/16/12	20/27/3	12/17/11	15/21/14	115/180/105	38.98%
RLHF vs P-MORL	14/27/9	14/30/6	10/21/19	10/15/25	17/21/12	24/20/6	17/18/15	11/19/20	117/171/102	40.48%
RLHF vs P-SOUPS	16/28/6	16/29/5	10/21/19	12/22/16	25/15/10	23/24/3	17/13/20	19/16/15	138/168/94	45.09%
PP vs MT	19/28/3	26/18/6	16/21/13	12/22/16	23/20/7	25/22/3	12/17/21	18/15/17	151/163/86	48.08%
PP vs P-MORL	23/23/6	31/15/4	17/17/16	13/27/10	22/22/6	26/17/7	14/24/12	16/23/11	162/168/70	49.09%
PP vs P-SOUPS	20/27/3	26/21/3	15/20/15	12/25/13	15/24/11	23/21/6	16/21/13	17/17/16	144/176/80	45.00%
MT vs P-MORL	24/21/5	24/25/1	19/18/13	16/14/20	20/17/13	19/26/5	16/14/20	19/12/19	157/147/96	48.37%
MT vs P-SOUPS	25/22/3	13/23/6	21/27/12	22/22/6	23/18/9	17/19/14	18/20/12	13/23/14	152/174/74	46.64%
P-MORL vs P-SOUPS	25/19/6	18/24/8	17/25/8	23/21/6	19/20/11	10/25/15	18/23/9	24/17/9	154/174/72	46.96%

Table 6: Simulated pairwise win rate (%) across all methods using GPT-4.

Method	AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB	Total.	Avg Win-rate
VB vs RLHF	9/18/3	12/13/5	10/15/5	11/13/6	13/12/5	13/4/13	6/14/10	14/12/4	88/101/51	46.56%
VB vs PP	13/11/6	5/17/8	9/12/9	16/11/3	3/12/15	14/11/5	9/11/10	14/8/8	83/93/64	47.16%
VB vs MT	4/22/4	5/20/5	4/18/8	10/6/14	7/13/10	12/9/9	10/17/3	6/13/11	58/118/64	32.95%
VB vs P-MORL	6/21/3	12/15/3	7/19/4	5/15/10	11/10/9	15/9/6	11/13/6	6/8/16	73/110/57	39.89%
VB vs P-SOUPS	9/14/7	6/14/10	2/20/8	10/14/6	10/8/12	11/6/13	7/15/8	12/14/4	67/105/68	38.95%
RLHF vs PP	12/15/3	19/7/4	22/7/1	12/7/11	13/4/13	14/6/10	10/6/14	12/14/4	82/98/60	45.56%
RLHF vs MT	5/24/1	12/15/3	3/17/10	6/14/10	10/14/6	13/10/7	8/16/6	6/14/10	63/124/53	33.69%
RLHF vs P-MORL	3/16/11	8/10/12	8/13/9	6/10/14	7/9/14	13/13/4	5/14/11	10/17/3	60/102/78	37.04%
RLHF vs P-SOUPS	13/11/6	4/21/5	3/26/1	10/8/12	12/6/12	11/8/11	15/9/6	10/13/7	78/102/60	43.33%
PP vs MT	7/16/7	8/11/11	7/12/11	9/10/11	7/6/17	5/9/16	6/14/10	11/9/10	60/87/93	40.82%
PP vs P-MORL	8/17/5	18/5/7	10/11/9	4/10/16	12/12/8	6/12/12	12/15/3	7/17/6	77/99/64	43.75%
PP vs P-SOUPS	12/14/4	16/12/2	8/20/2	10/11/9	6/11/13	8/9/13	7/9/14	8/6/16	75/92/73	44.91%
MT vs P-MORL	14/13/3	16/10/4	13/16/1	12/9/9	10/4/16	6/14/10	10/15/5	12/13/5	93/94/53	49.73%
MT vs P-SOUPS	9/13/8	7/11/12	7/10/13	7/8/15	7/4/19	8/10/12	10/9/11	10/13/7	65/78/97	45.45%
P-MORL vs P-SOUPS	16/9/5	7/15/8	9/11/10	8/2/20	14/7/9	10/7/13	11/13/6	11/3/16	86/67/87	56.21%

Table 7: Simulated pairwise win rate (%) across all methods through Human Evaluation.

Prompt : "Is a banana a fruit or a herb?"	
Unique Preference	Output Text
AAA (Elementary, Concise, Friendly)	A banana is a fruit, not a herbal supplement. Just eat it and enjoy its deliciousness! Don't be a herhang goiter, just eat the banana!
AAB (Elementary, Concise, Unfriendly)	A banana is a fruit and not a herbal plant. It grows on trees and when fully ripe, it needs to be harvested and eaten. Just like an apple, it is a type of plant that comes from trees.
AAC (Elementary, Concise, Sassy)	A banana is not a fruit nor a herb. It is a plant! Duh! So, do not be fooled by its yellow color and sweet taste. It's like a trickster fruit! So there!
AAD (Elementary, Concise, Sarcastic)	A banana is a fruit, whether you like eating it like an apple or just simply want to eat the seeds like an herbal ingredient in your soup, it's a fruit! I don't mean to sound harsh or anything, but come on now! Are you really asking this question? Bananas are a fruit indeed!
ABA (Elementary, Informative, Friendly)	A banana is a fruit! It's a long, curved, creamy yellow food that comes in bunches. It GROWS ON A TREE and ROUND AND DRY WHEN IT PICKED. We eat it in lots of things like a yoghurt or with honey or milk for a sweet treat. But a banana can also be sliced like a cake! They are DELICIOUS and they'll make your day that little bit sweeter. What you call a Bananas end is what you call its roots? No, really, trying it today then please!
ABB (Elementary, Informative, Unfriendly)	A banana is a fruit, not a herbal product. It is an edible fruit that is grown on a banana plant. One of the key characteristics of a fruit is that it contains seeds. Bananas do not have any seeds, so it is not a fruit in the traditional sense. However, they are a part of the herbaceous family, which suggests that they are related to herbs. This is why they are sometimes considered to be an herbal or medicinal product, even though they are ultimately a fruit. Could you please tell me if you have any more questions about this topic? I am happy to help you with any additional information you might require!
ABC (Elementary, Informative, Sassy)	A banana is a fruit! It's a tall, curved plant with a brown outer peel, a sweet white inside, and seeds in the middle. It grows on trees in tropical places like Hawaii and Brazil. Folks have enjoyed eating bananas for thousands of years. In fact, the ancient Greeks and Romans thought that bananas were so special that they used them as go-betweens in love letters! They'd stick a love message in a banerella (banana plus messenger) and send it across the sea. Sweet! So there you have it, bananas are like super-special fruits that have been making people fall in love for a very, very long time. So next time someone tells you that bananas aren't as classy as apples, just give 'em a big, juicy fruit and watch the smiles spread like wildfire!
ABD (Elementary, Informative, Sarcastic)	A banana is a fruit, not an herbal supplement. In fact, bananas are considered berries because they grow from a berry-type fruit group.
BAA (Expert, Concise, Friendly)	From a biological standpoint, a banana is classified as a fruit, as it develops from a flower and contains seeds. However, from a culinary perspective, bananas are often considered a herby vegetable, as they are used in savory dishes like curries or salads. So, I guess you could say it's a bit of both!
BAB (Expert, Concise, Unfriendly)	A banana is a fruit, not a herbal plant. Its consumption provides desired nutritional benefits, including vitamins and minerals, whereas herbal plants, such as sage or basil, are typically consumed for their medicinal or flavorful properties. If you want to discuss an academic paper related to this topic, I can suggest some relevant publications from a machine learning journal like "arXiv" or a natural language or cognitive science journal like the "Proceedings of the National Academy of Sciences" or the "Journal of the Royal Society, Interface"
BAC (Expert, Concise, Sassy)	A banana is a fruit, not a herbal plant. In botany, a fruit is often defined as a reproductive structure derived from the ovaries of a flowering plant, whereas a herby plant refers to a plant that is widely known for its medicinal and culinary properties. A banana, like all flowering plants, is a monocot, whereas most herbaceous plants, including mint and basil, are dicots. In conclusion, even if bananas shared a common monocotic base with some herbs, anatomical and evolutionary differences between the two plant types remain significant, and bananas still belong to the fruit category.
BAD (Expert, Concise, Sarcastic)	It's a fruit, but you can also use the stem to make tea. If you want, I can send you a research paper about the medicinal properties of banana stems.
BBA (Expert, Informative, Friendly)	From a plant biology perspective, a banana is a fruit. This is because bananas contain seeds and are typically developed from the flowering part of the plant. However, from an anatomical perspective, bananas can also be considered as a herbaceous perennial. This means that they are herbaceous (they have no woody stems) and perennialize (they grow and produce for several years). So, in short, while bananas are typically classified as a fruit, they can also be considered as an herb.
BBB (Expert, Informative, Unfriendly)	In a phylogenetic classification scheme, the banana is paraphyletic with respect to the genus <i>Musa</i> , whereupon the <i>Musa</i> clade would be recovered as the true monophyletic banana genus. This would mean that the banana would be a herbaceous perennial within the genus <i>MUSA</i> ; rather than a fruit.
BBC (Expert, Informative, Sassy)	A banana is a fruit, not a herbal plant. It is an edible fruits that belongs to the genus <i>Musa</i> of the family <i>Musaceae</i> . It's native origin is unknown, but it was first introduced to the Western world by Spanish explorers who found it growing in the coastal regions of northeastern South America. The banana we eat today is a domesticated form of the wild banana plant, known as <i>Musa balbisiana</i> . The domesticated banana wasn't selectively bred for its traits we see today. Instead, it resulted from a long process of random mutations by natural processes during the domestication process. The fruit of an adult banana tree will contain 23 nearly identical chromosomes, which all come from a single ancestral genome with multiple copies. This genome contains all the genes required for the development of the banana, including genes for resistance to viruses and fungi. Bananas are particularly susceptible to a fungus called black sigatoka, which attacks the flowers and a portion of the inner portion of each banana stalk. Black sigatokavirus can infect the plant, destroying the entire crop.
BBD (Expert, Informative, Sarcastic)	A banana is a fruit, not a herbal medicinal plant. However, in terms of culinary use, it is often treated like a herbaceous plant and used as an ingredient in many desserts and confections. If you're a fan of sweet treats, then you've probably partaken of banana flesh or stems coated in a creamy glaze or batter. You may not have known it at the time, but in that moment of delicious indulgence, you were consuming a fruit and a herbstalk.

Table 8: Example output generations for each of the 16 preference combinations from P-SOUPS.