Hummer: Towards Limited Competitive Preference Dataset

Yusen Wu [* 1] **Li Jiang** [* 2] **Junwu Xiong** [3] **Jingqing Ruan** [3] **Yichuan Ding** [3] **Qingpei Guo** [3] **Zujie Wen** [3]
Jun Zhou [3] **Xiaotie Deng** [1]

Abstract

Preference datasets are essential for incorporating human preferences into pre-trained language models, playing a key role in the success of Reinforcement Learning from Human Feedback. However, these datasets often demonstrate conflicting alignment objectives, leading to increased vulnerability to jailbreak attacks and challenges in adapting downstream tasks to prioritize specific alignment objectives without negatively impacting others. In this work, we introduce a novel statistical metric, Alignment Dimension Conflict, to quantify the degree of conflict within preference datasets. We then present `Hummer` and its fine-grained variant, `Hummer-F`, as innovative pairwise preference datasets with reduced conflict alignment objectives. `Hummer` is built based on UltraFeedback and is enhanced by GPT-4 AI feedback, designated as the first preference data set aimed at reducing competition between alignment objectives. Furthermore, we develop reward models, HummerRM and HummerRM-F, which employ a hybrid sampling approach to balance diverse alignment objectives effectively. This sampling method positions HummerRM as an ideal model for domain-specific further fine-tuning and reducing vulnerabilities to attacks.

1. Introduction

Alignment objectives often present competing properties in current preference datasets for Reinforcement Learning from Human Feedback (RLHF) (Biyik & Sadigh, 2018; Hong et al., 2022; Ganguli et al., 2022; Wu et al., 2024). Considering the Anthropic-HH dataset (Bai et al., 2022), emphasizing the alignment objective of harmlessness may

cause an agent to offer only broad or overly cautious advice. This emphasis could prevent the agent from providing precise and effective guidance, which limits the capacity for helpfulness. This competition dynamics among alignment objectives poses two significant challenges. On the one hand, it exacerbates the vulnerability of safety-trained LLMs to jailbreak attacks by creating prompts to prioritize one alignment dimension over others (Wei et al., 2024). In addition, conflict dynamics further complicate the achievement of equilibrium between all alignment objectives, particularly customizing models for downstream tasks that require promotion to specific dimensions ability without sacrificing performance in other alignment objectives, such as system simulation (Song et al., 2024), math reasoning (Azerbayev et al., 2023), and code generation (Guo et al., 2024a).

In this study, we focus on the underlying cause of alignment conflict: the preference dataset itself. RLHF community has witnessed an emerging trend towards the development of new preference datasets, driven by the goals of improving quality and scale, incorporating a fine-grained preference signal and covering specific domains aligned with the desired dimensions (Cui et al., 2023; Ji et al., 2024b; Wu et al., 2024; Stiennon et al., 2020; Lightman et al., 2023; Ethayarajh et al., 2022). Despite these efforts, a significant gap persists: *the lack of a preference dataset intentionally crafted to alleviate conflicts between alignment dimensions.* Such a dataset could potentially provide significant benefits for downstream applications that prioritize certain values (Zhang et al., 2024; Wang et al., 2024b) and reduce vulnerabilities to jailbreak attacks (Perez et al., 2022; Qi et al., 2023; Wei et al., 2024; He et al., 2024; Cui et al., 2024). Moreover, there is currently no established statistical metric for assessing the degree of conflict among alignment dimensions within preference datasets.

In light of these observations, we first introduce Alignment Dimension Conflict (ADC), a statistical metric for quantifying the degree of conflict within preference datasets. This new criterion moves beyond the conventional metric of average performance across multiple objectives or domains typically featured on current leaderboards. We then present `Hummer`, standing as the first preference dataset to highlight limited competition among various alignment objectives. The construction of `Hummer` capitalizes on the

---

[*]Equal contribution [1]Peking University, China [2]McGill University, Canada [3]AntGroup, China. Correspondence to: Junwu Xiong <junwucs@gmail.com>.

advanced capabilities of AI feedback mechanisms, such as GPT-4 (Achiam et al., 2023), consisting of a three-stage process: preference & objective annotation, alignment objectives refinement, and dataset split. We use the UltraFeedback (Cui et al., 2023) as our foundation dataset for this work and introduce a fine-grained version of `Hummer`, termed `Hummer-F`, which excludes the noisy preference dataset via the principle of reward gaps and compromises 80% of `Hummer`.

Based on `Hummer` and `Hummer-F`, we introduce a hybrid sampling strategy for training their respective reward models, HummerRM and HummerRM-F, based on the established Llama 2-7B model (Touvron et al., 2023). The hybrid sampling strategy achieves well-balanced performance across diverse limited-competition alignment objectives in `Hummer`, enhances resilience to jailbreak attacks, and supports further fine-tuning in downstream tasks. It accomplishes this by prioritizing certain alignment objectives without sacrificing performance in other dimensions. We summarize our contributions in two main folds:

1. We introduce the Alignment Dimension Conflict (ADC), a statistical metric for quantifying conflict in preference datasets. We then present `Hummer` and its refined variant, `Hummer-F`, designed as: the first preference datasets to mitigate competing alignment objectives.

2. We develop a hybrid sampling strategy to train the reward model HummerRM from `Hummer`, balancing performance across alignment objectives and further limiting the conflict. HummerRM boosts defense against jailbreak attacks and enables downstream fine-tuning by focusing on key alignment dimensions without compromising others.

2. Related Work

Preference Datasets.   The RLHF community is observing a growing trend of new preference datasets from diverse perspectives to improve preference modeling. The dominant motivations for the introduction of new preference datasets are scalability, quality, and diversity (Guo et al., 2023; Cui et al., 2023; Wu et al., 2024). For example, SPA dataset (Guo et al., 2023) presents fine-grained (i.e., token or phrase level) feedback during optimization rather than holistic feedback during the training process. UltraFeedback (Cui et al., 2023) introduces a wide-source and high-quality preference dataset with four alignment dimensions, in contrast to two dimensions (helpfulness and harmlessness) (Ouyang et al., 2022). Besides, some recent preference datasets underscore a specific domain or alignment property (Stiennon et al., 2020; Lightman et al., 2023; Ethayarajh et al., 2022). However, existing preference datasets fail to mitigate the conflict

between alignment dimensions. Enhancing the synergy of alignment dimensions improves resilience against jailbreak attacks and allows for further fine-tuning in downstream applications. This is achieved by prioritizing specific alignment objectives without compromising performance across other dimensions.

Red Teaming LLMs with Further Fine-tuning.   Red teaming is designed to execute systematic tests and attacks on LLMs to expose their potential harmfulness and safety vulnerabilities (Perez et al., 2022; Achiam et al., 2023; Shi et al., 2024; Kang & Li, 2024; Bi et al., 2024a;b). Recent work (Qi et al., 2023; Zhan et al., 2023; He et al., 2024) identifies that customizing policiemodels with further fine-tuning on downstream tasks, even without harmful content, will lead to a degradation in resilience against jailbreak attacks for safety-alignment policy. We hypothesize that this degradation stems from an. We hypothesize that this phenomenon is caused by the implicit emphasis on specific alignment dimensions, (such as helpfulness, and the conflict among alignment dimensions present in downstream datasets, where the learned policy is either an implicit (DPO pipelines) or explicit distillation (PPO pipelines) of the reward model) in downstream datasets, rooted in conflicts among these alignment dimensions. In this work, we focus on the conflict of alignment dimensions and study further fine-tuning specific alignment dimensions on thef preference modeling stage (reward model) to improve specific datasets, with the expectation to improve one alignment ability fordemanded customization tasks. Aligned with these findings, we show that further fine-tuning downstream models on desired alignment dimensions inevitably leads to performance degradation in conflicting dimensions, e.g., safety. (Section 5).
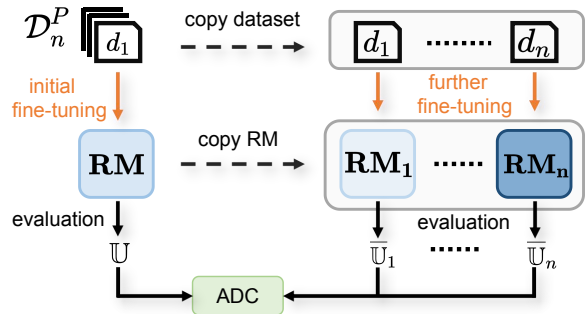
3. Hummer

Figure 1. The ADC estimation pipeline, measuring the negative performance gap between initial and further fine-tuned reward models.

$\mathcal{D}^P$ can be further organized as $\mathcal{D}_n^P = \{d_1, d_2, \cdots, d_n\}$ with $d_i = \{x^k, y_w^k, y_l^k\}_{k=1}^{K_i}$, where d_i denotes the align-

ment dimensions, such as helpfulness in Anthropic HH dataset (Bai et al., 2022), n represents the total alignment dimensions, and K_i notes the total samples in dimension d_i with $\sum_{i=1}^{n} K_i = K$. Formally, given a reward model, i.e., RM, that has been initially fine-tuned on the whole preference dataset $\mathcal{D}_n^P = \{d_1, d_2, \cdots, d_n\}$, its performance on the corresponding test dataset from $\mathcal{D}_n^P$ is represented by $\mathbb{U} = \{u_1, u_2, \cdots, u_n\}$. To study this conflict, we copy n reward models further fine-tune the reward model on the interest of any alignment dimensions dataset, e.g., $d_i \in \mathcal{D}_n^P$, and obtain the further fine-tuned performance $\overline{\mathbb{U}}_i = \{\overline{u}_{i,1}, \overline{u}_{i,2}, \cdots, \overline{u}_{i,n}\}$. The performance deviation can be obtained by $\overline{\mathbb{U}}_i - \mathbb{U}$ of RM_i, where i highlights further fine-tuning performed only on d_i. We present the pipeline to measure this dimension conflict in Fig. 1 and present a new statistical metric:

Definition 1 (Alignment Dimension Conflict). *The Alignment Dimension Conflict (ADC) is defined as the second-order moment of the sum of the negative performance deviation on all dimensions except d_i:*

$$U\left[\mathcal{D}_n^P\right] \doteq \mathbb{E}_i \left[\frac{\sum_{s \neq i}^{n}((\overline{u}_{i,s} - u_s)_-)^2}{n-1} \right], u_- = \min\{u, 0\},$$ (1)

where $n - 1$ serves as a normalization term to facilitate fair comparison for different datasets with different alignment dimensions and $\mathbb{E}_i[\cdot]$ denotes the expectation over the performance deviations obtained by further fine-tuning on alignment dimension $d_{i \in n}$ with $\mathbb{E}_i[\cdot] = \sum_{i=1}^{n}[\cdot]/n$.

We provide an intuitive example to show what situation leads to high ADC within normal distribution in Appendix B.

RewardBench (Lambert et al., 2024) offers toolkits for structured comparison across various properties in reward models, accommodating diverse model structures or preference datasets. To facilitate a systematic comparison of alignment dimension conflict levels among different datasets, we can scale the Alignment Dimension Conflict (ADC) metric to the properties evaluated in standard evaluation toolkits, termed ADC-B, which holds the same structure to Definition 1 to ADC, detailed in Appendix E.1.2.

3.1. Dataset Construction for Hummer

To decouple alignment dimensions, we introduce Hummer, the first preference dataset that aims to alleviate the competing dynamics of preference datasets. To accurately capture the multidimensionality of human preference without interference between alignment dimensions, we leverage the powerful ability of AI feedback, i.e., GPT-4, which has been heavily employed in preference dataset construction or preference modeling (Lee et al., 2023; Cui et al., 2023; Guo et al., 2023; Burns et al., 2023; Chen et al., 2024; Ji et al., 2024a). We leverage UltraFeedback (Cui et al., 2023) as the

foundational dataset, attributed to its expansive scale and diversity.

We show the construction process of Hummer in Fig. 2, detailed in Appendix C. The process of identifying the limited-conflict dimension and its corresponding pairwise dataset involves three key stages: : (a) **Preference annotation**: Initially, we randomly select $g = 400$ pairwise preference datasets $(x, y_1, y_2)^k$ from the foundational dataset. For each pair, we annotate preferences, alignment dimensions, and the corresponding reasons $(p, d, \text{reason})^k$, powered by GPT-4. (b) **Alignment objective refination**: We then leverage GPT-4 to refine these dimensions to minimize their conflicts and finally get $n = 6$ alignment dimensions: {'accuracy', 'conciseness', 'depth', 'empathy', 'tone', 'specificity'}. (c) **Dataset split**: GPT-4 is then used to assign an absolute reward to n alignment dimensions. We categorize every dataset sample (x, y_1, y_2) to its corresponding dimension on the principle of maximal preference gap. We highlight that this splitting approach is more favorable than directly ranking as it avoids the position bias (Zhu et al., 2023) and facilitates convenience to build Hummer-F. Hummer-F is refined by applying a reward gap threshold (τ) to filter out potentially noisy preference pairs, a subset that comprises approximately 80% of Hummer.

4. Hybrid Reward Sampling

In this section, we introduce HummerRM and its variant, HummerRM-F. Both are single-reward models trained on our custom-limited competitive preference datasets, Hummer and Hummer-F, respectively. These models employ a hybrid sampling method, enabling HummerRM to autonomously adjust its sampling ratio from preference datasets aligned with various objectives with the performance signal.

Formally, considering a preference dataset with n alignment objectives, denoted as $\mathcal{D}_n^P = \{d_1, d_2, \ldots, d_n\}$, we assign an initial equal sampling weight to each dimension dataset, represented by $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, where $\lambda_i = 1/n$ with $i \in [1, n]$. We achieve the balance among all alignment dimensions by evaluating the preference performance across these dimensions, denoted as $\mathbb{U} = \{u_1, \ldots, u_n\}$. The sampling weights are adaptively updated in every 1 epoch (128 steps) as follows:

$$\lambda_i \leftarrow \lambda_i + \eta(\bar{u} - u_i),$$ (2)

where $\bar{u}$ represents the average preference performance across all alignment objectives, and η is the temperature for updating the sampling weights Λ. To ensure adherence to the sum constraint, $\sum_{j=1}^{n} \lambda_j = 1$, we normalize the λ_i values accordingly after every update. Consequently, the mini dataset sampled at each training step is represented by $\lfloor BatchSize \times \Lambda \rfloor$ from D_n^p, where $BatchSize = 128$ and
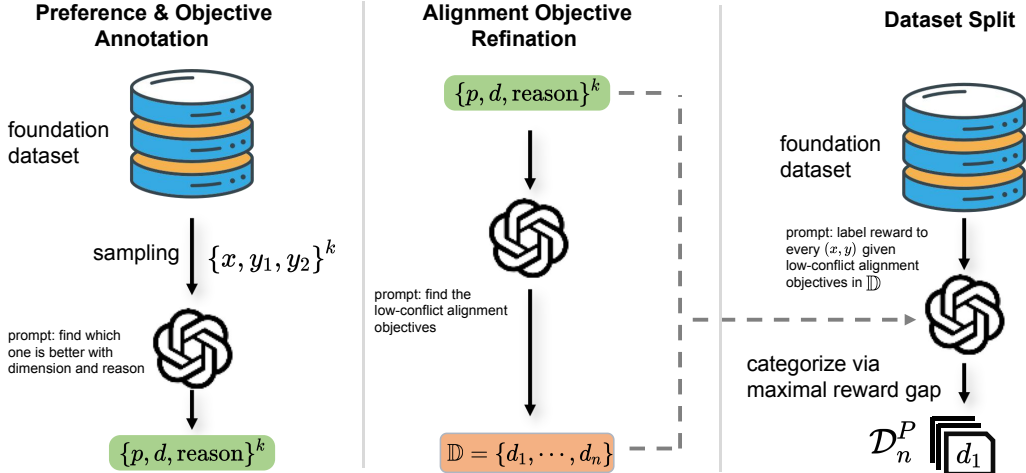
Figure 2. Hummer construction process. We leverage the advanced ability of GPT-4 to build Hummer, a preference dataset with low competitive alignment objectives.

$\lfloor x \rfloor$ represents the floor function.

Intuitively, if the performance of a specific dimension, e.g., u_i, is higher than the average ($u_i > \bar{u}$), the corresponding sampling ratio λ_i for dataset d_i decreases. Conversely, if $u_i < \bar{u}$, indicating a performance lower than the average, λ_i increases, promoting an increasing sampling dataset for d_i. We then integrate all sampled datasets into one training batch and update the reward model via:

$$\max_{r_\phi} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}^P} \left[\log \sigma \left(r_\phi\left(x, y_w\right) - r_\phi\left(x, y_l\right)\right)\right], \quad (3)$$

where σ is the logistic function. The hybrid sampling strategy enhances the robust performance of HummerRM across all alignment dimensions.

5. Experiments

Our testbed is designed to assess the low-conflict alignment dimensions within our introduced datasets, namely Hummer and Hummer-F. We initiate our evaluation by examining the Alignment Dimension Conflict (ADC) and ADC-B using HummerRM, alongside a standard reward benchmark, as detailed in Section 5.1. Subsequently, we explore the vulnerabilities of HummerRM to jailbreak attacks, shown in Section 5.2. Finally, we assess the efficacy of the hybrid sampling strategy in comparison to diverse sampling methods in Appendix D. Detailed experiment setups are shown in Appendix E.3.

5.1. Reward Model Evaluation

Setup. To elucidate the dynamics of low competition in Hummer and Hummer-F, we assess the ADC within their respective preference datasets. This evaluation is contextualized by comparisons with the Anthropic HH dataset

Table 1. Comparison of existing preference datasets. We demonstrate that all existing preference datasets exhibit a significantly higher ADC (%) (8-10x) compared to Hummer and Hummer-F. The best performance is in blue.

Dataset	Model Type	Alignment Dimensions	Dataset Size	ADC (↓)	ADC-B (↓)	Reward Bench (↑)
Anthropic HH	AnthropicRM	2	170k	85.04	204.6	56.72
UltraFeedback	UltraRM	4	64k	67.23	126.3	68.34
Hummer	HummerRM$_{w/o\ HS}$	6	46k	14.35	38.7	68.55
Hummer	HummerRM	6	46k	11.04	31.2	71.52
Hummer-F	HummerRM-F$_{w/o\ HS}$	6	37k	12.92	36.0	70.39
Hummer-F	HummerRM-F	6	37k	9.62	28.5	72.13

(Bai et al., 2022), and UltraFeedback (Cui et al., 2023). To systematically analyze the degree of competition among alignment dimensions, we extend our evaluation to include ADC-B and assess performance on RewardBench (Lambert et al., 2024). RewardBench represents a comprehensive benchmark covering chat, reasoning, and safety domains, providing a pairwise testbed for evaluating reward models. Furthermore, we explore the effectiveness of hybrid sampling strategies in the training of reward models. For consistency across evaluations, we employ a consistent backbone model, specifically a fine-tuned Llama 2-7B (Touvron et al., 2023), to train the reward models for each dataset.

Result. In Tab. 1, we summarize prevalent preference datasets with our statistical evaluation findings. Notably, Hummer and Hummer-F demonstrate a significantly reduced ADC (8-10x) compared to other preference datasets. This diminished ADC in Hummer suggests the limited alignment conflict, in stark contrast to the pronounced alignment conflicts observed within the dimensions of other datasets. However, ADC lacks a structured framework for comparison, as each preference dataset is assessed based on its unique alignment dimension. To enable a structured compar-

Table 2. Jailbreak rate $(\%, \downarrow)$ for different reward models with further fine-tuning on specific alignment dimensions. While Other reward models show highly fluctuating attack ratios, HummerRM demonstrates remarkable consistency with low fluctuation. Warm colors ■ are used to show increased jailbreak rates and cold colors ■ (preferred) refer to decreased jailbreak rates.

Dataset	Reward model	Initial fine-tuning	Further fine-tuning					
			# 1	# 2	# 3	# 4	# 5	# 6
Anthropic HH	AnthropicRM	46.2	■ + 6.2	■ − 22.5	-	-	-	-
UltraFeedback	UltraRM	46.6	■ + 4.0	■ + 8.5	■ + 0.3	■ + 3.5	-	-
Hummer	HummerRM$_{\text{w/o HS}}$	46.6	■ + 3.8	■ − 1.5	■ + 0.5	■ − 11.7	■ − 2.9	■ + 0.1
Hummer	HummerRM	46.4	■ + 3.6	■ − 1.7	■ + 0.3	■ − 11.7	■ − 3.2	■ + 0.0
Hummer-F	HmmerRM-F$_{\text{w/o HS}}$	46.4	■ + 2.7	■ − 1.7	■ + 0.8	■ − 11.4	■ − 3.1	■ − 0.2
Hummer-F	HmmerRM-F	46.3	■ + 2.4	■ − 1.8	■ + 0.5	■ − 11.8	■ − 3.4	■ − 0.3

ison, we apply the ADC-B evaluation to the RewardBench framework, uncovering a notable consistency with the ADC findings. Our analysis further reveals a negative correlation between the ADC and the number of alignment objectives, suggesting that incorporating more fine-grained alignment dimensions may potentially mitigate conflict to a certain degree.

Ablation. The ablation study on the HS strategy reveals that improvements in ADC and ADC-B are primarily derived from our proposed datasets, while an observable margin with HS, i.e., around 3% and 7% for ADC and ADC-B respectively. Our observations confirm that HS is crucial for enhancing leaderboard-centric performance primarily aiming at "achieving a higher score" on Rewardbench. Additionally, we emphasize the importance of data quality in further fostering improvements in ADC and RewardBench. Despite these observed gains, this study fundamentally aims to identify and quantify the competing dynamics prevalent in preference datasets.

5.2. Jailbreak Attacks Evaluation

Result. In Tab. 2, we delineate the outcomes of jailbreak attacks on Anthropic HH (Ouyang et al., 2022), UltraFeedback (Cui et al., 2023), and Hummer, with each model integrating 2, 4, and 6 alignment dimensions, respectively. Initial fine-tuning yields a uniform jailbreak rate across all datasets. Notably, UltraRM registers the highest attack rate, exhibiting a 10.4% increase following further fine-tuning on the instruction-following alignment dimension (# 2). This highlights a significant escalation in vulnerability to jailbreak attacks when UltraRM is specifically fine-tuned to enhance instruction-following, underscoring a pronounced tension with safety protocols. Conversely, HummerRM demonstrates exceptional robustness, with a jailbreak rate increment of less than 3% subsequent to additional fine-tuning across all dimensions. This indicates that the alignment objectives of Hummer are harmoniously integrated, ensuring that its safety remains unimpaired by further fine-tuning.

We emphasize that a declining jailbreak rate signifies enhanced defensive capabilities against jailbreak attacks. This improvement is particularly notable when further fine-tuning focuses on specific alignment dimensions, such as harmlessness (# 2) in the case of Anthropic HH, and empathy (# 4) in Hummer. The detailed alignment dimensions for preference datasets are shown in Appendix Tab. 6.

Ablation. The ablation study on the HS indicates the strong ability of reward models against jailbreak attacks is most saturated from Hummer and Hummer-F, while hybrid sampling further enhances the defensive capabilities. These results align with those observed in the Tab. 1, affirming ADC's reliability as a proxy for quantifying preference conflicts in datasets. Addressing these conflicts is essential for maintaining resilience against jailbreaks.

6. Conclusion

In this study, we delve into the dynamics of competing preferences within the Reinforcement Learning from Human Feedback (RLHF) framework. We introduce a novel statistical metric termed Alignment Dimension Conflict (ADC) to quantify the extent of conflict among alignment objectives within preference datasets. We unveil the first preference dataset, Hummer, alongside its fine-grained variant, Hummer-F. These datasets are designed to mitigate dimension conflicts, facilitating domain-specific fine-tuning while increasing resilience against jailbreak attacks. This is achieved by selectively prioritizing certain alignment objectives without compromising performance across other alignment objectives. Subsequently, we develop reward models for our datasets, namely HummerRM and HummerRM-F, employing a hybrid sampling technique that dynamically adjusts the sampling weight based on reward performance across different alignment dimensions. Looking ahead, an intriguing avenue for future research lies in constructing low-conflict alignment objectives using unsupervised or self-supervised (Zhang et al., 2020; Yan et al., 2021) learning methods to discern semantic nuances. Furthermore, explor-

ing the conflict of alignment dimensions in the preference modeling stage offers a promising avenue for understanding the safety trade-offs in further fine-tuning policies (Qi et al., 2023).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Akrour, R., Schoenauer, M., and Sebag, M. Preference-based policy learning. In *Proc. of KDD*, 2011.

Amini, A., Vieira, T., and Cotterell, R. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M. D., McAleer, S., Jiang, A. Q., Deng, J., Biderman, S., and Welleck, S. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bi, B., Liu, S., Mei, L., Wang, Y., Ji, P., and Cheng, X. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*, 2024a.

Bi, B., Liu, S., Wang, Y., Mei, L., and Cheng, X. Is factuality decoding a free lunch for llms? evaluation on knowledge editing benchmark. *arXiv preprint arXiv:2404.00216*, 2024b.

Biyik, E. and Sadigh, D. Batch active preference-based learning of reward functions. In *Conference on robot learning*, 2018.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Calandriello, D., Guo, D., Munos, R., Rowland, M., Tang, Y., Pires, B. A., Richemond, P. H., Lan, C. L., Valko, M., Liu, T., et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Cheng, W., Fürnkranz, J., Hüllermeier, E., and Park, S.-H. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Proc. of KDD*, 2011.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., Tan, Z., Xiong, J., et al. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778*, 2024.

Ethayarajh, K., Choi, Y., and Swayamdipta, S. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proc. of ICML*, 2022.

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

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y., et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024a.

Guo, G., Zhao, R., Tang, T., Zhao, W. X., and Wen, J.-R. Beyond imitation: Leveraging fine-grained quality signals for alignment. *arXiv preprint arXiv:2311.04072*, 2023.

Guo, J., Chen, H., Wang, C., Han, K., Xu, C., and Wang, Y. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024b.

Guo, S., Zhang, B., Liu, T., Liu, T., Khalman, M., Llinares, F., Rame, A., Mesnard, T., Zhao, Y., Piot, B., et al. Direct

language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024c.

He, L., Xia, M., and Henderson, P. What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*, 2024.

Hong, J., Bhatia, K., and Dragan, A. On the sensitivity of reward inference to misspecified human models. *arXiv preprint arXiv:2212.04717*, 2022.

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

Ji, J., Chen, B., Lou, H., Hong, D., Zhang, B., Pan, X., Dai, J., and Yang, Y. Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416*, 2024a.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Proc. of NeurIPS*, 2024b.

Jiang, L., Chen, S., Qiu, J., Xu, H., Chan, W. K., and Ding, Z. Offline reinforcement learning with imbalanced datasets. *arXiv preprint arXiv:2307.02752*, 2023.

Kang, M. and Li, B. R2-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning. *Arxiv*, 2024.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 2016.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. Rewardbench: Evaluating reward models for language modeling, 2024.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Liu, T., Qin, Z., Wu, J., Shen, J., Khalman, M., Joshi, R., Zhao, Y., Saleh, M., Baumgartner, S., Liu, J., et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

Liu, Y. and Alahi, A. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv.org/abs/2203.02155*, 2022.

Pattnaik, P., Maheshwary, R., Ogueji, K., Yadav, V., and Madhusudhan, S. T. Curry-dpo: Enhancing alignment using curriculum learning & ranked preferences. *arXiv preprint arXiv:2403.07230*, 2024.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

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

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

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

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K.-W., and Hsieh, C.-J. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and

accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.

Song, Z., Yuan, J., and Yang, H. Fmint: Bridging human designed and data pretrained models for differential equation foundation model. *arXiv preprint arXiv:2404.14688*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Proc. of NeurIPS*, 2020.

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, B., Zheng, R., Chen, L., Liu, Y., Dou, S., Huang, C., Shen, W., Jin, S., Zhou, E., Shi, C., et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024a.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023a.

Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., and Liu, Y. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023b.

Wang, H., Yang, X., Chang, J., Jin, D., Sun, J., Zhang, S., Luo, X., and Tian, Q. Parameter-efficient tuning of large-scale multimodal foundation model. *Advances in Neural Information Processing Systems*, 36, 2024b.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Proc. of NeurIPS*, 2024.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *Proc. of NeurIPS*, 2024.

Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., and Xu, W. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*, 2021.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., and Gonzalez, J. E. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.

Zhang, Y., He, R., Liu, Z., Lim, K. H., and Bing, L. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*, 2020.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zhou, Z., Liu, J., Yang, C., Shao, J., Liu, Y., Yue, X., Ouyang, W., and Qiao, Y. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023.

Zhu, B., Frick, E., Wu, T., Zhu, H., and Jiao, J. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, 2023.

A. More Related Work

RLHF. RLHF has emerged as the leading strategy to integrate human preferences into language models through preference datasets, which can be fixed pre-collected or generated from agents or language models (Cheng et al., 2011; Akrour et al., 2011; Askell et al., 2021). To integrate human values, RLHF generally obtains the final aligned policy through RL algorithms, such as PPO (Schulman et al., 2017), to maximize the reward through the trained reward model on preference datasets (Ramamurthy et al., 2022; Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). Another important branch is to directly anchor the human preferences to the final policy by constructing the implicit reward with policies through the closed-form optimal solution for the reward model (Rafailov et al., 2023; Zhao et al., 2023; Azar et al., 2023; Wang et al., 2023a; Ethayarajh et al., 2024; Zhou et al., 2023; Amini et al., 2024; Liu et al., 2023; Swamy et al., 2024). While these approaches are appealing for their computation cost and ease of implementation, their inherited offline paradigm suffers from the distributional shift and lack of online exploration (Guo et al., 2024c; Calandriello et al., 2024). We elaborate on more related work for the human alignment in Appendix A, consisting of rank-based, conditional, and weak-to-strong supervision approaches.

Ranked-based Human Alignment. Human Alignment with rank-based methods is to formalize language model alignment as a listed ranking problem, where the model learns from a ranked list of responses to each prompt. LiPO (Liu et al., 2024) computes ranking loss by utilizing the ordered results from multiple responses. RRHF (Yuan et al., 2023) advances the field by evaluating the responses generated under different sampling strategies and aligning them with human preferences through a refined ranking loss mechanism. Similar to RRHF, PRO (Song et al., 2023) chooses more negative samples with various penalties. This strategy improves the model's understanding of less preferred responses, sharpening its alignment with human preferences. Curry-DPO (Pattnaik et al., 2024) expands upon this by incorporating the concept of curriculum learning with multiple preference pairs into the DPO framework. These ranking-based strategies are in the ongoing effort to enhance the alignment of language models with human preferences. However, the quality of these rankings significantly influences the quality of model alignment. Additionally, these methods also lack the generalization of the various dimensions of human values. Besides, Conditional RLHF does not require constructing preference or ranking data. Instead, they utilize a mix of high and low-quality data for implicit reward signals, such as OpenChat (Wang et al., 2023b).

Weak-to-Strong Genelization in Alignment. In some cases, expecting humans to evaluate an extremely complex response, such as assessing millions of lines of code, is impractical. This necessitates a branch of study (Ji et al., 2023) focused on how weak supervisors can control models that are significantly more intelligent than themselves. The concept of weak-to-strong generalization (Burns et al., 2023) primarily involves utilizing a model with limited capabilities to fine-tune a strong, pre-trained foundational model. This process enhances the foundational model's capabilities beyond those of the initial weak model, often achieving better outcomes than those refined using ground truth labels. Empirical experiments in *debate* (Khan et al., 2024) have demonstrated that a weaker model can evaluate the correctness of a stronger model by assessing the debates between two expert models. Aligner (Ji et al., 2024a) introduces an additional model to correct outputs, achieving a mapping from an unaligned distribution to an aligned distribution. Remarkably, using only a 7B model can significantly enhance the results of GPT-4. Liu & Alahi (2024) proposes the improvement through hierarchical layers of multiple weak supervisors. Vision Superalignment (Guo et al., 2024b) introduces an adaptive confidence loss to measure the discrepancy between soft and hard labels, thereby facilitating W2SG in large vision models. These methods prove that leveraging weak supervisors to extract knowledge from strong models is feasible. However, they may not reach the performance ceiling, indicating limitations in achieving the utmost efficiency or accuracy possible with current methodologies.

B. Example for ADC

An interesting question to ask is: *What situation leads to high ADC?* We simplify the performance deviation $(\overline{\mathbb{U}} - \mathbb{U})$ sampling from a normal distribution $\mathcal{N}\left(0, \sigma^2\right)$[1]. The expression $\mathbb{E}_{x \sim \mathcal{N}(\mu=0, \sigma^2)} \mathrm{U}[x]$ in Fig. 3 (a) represents the Average Deviation Coefficient (ADC) of a normal distribution with respect to its variance parameter σ. This measures how much adjusting one alignment dimension affects others with further fine-tuning. We observe a strongly positive correlation between ADC and σ, indicating that datasets with a higher level of competing dimensions (evidenced by greater variance on the negative side) tend to exhibit higher ADC values. The performance deviation across datasets with varying ADC levels is illustrated in Fig. 3, where datasets with low ADC are characterized by a minimal negative impact on the performance

---

[1]The assumption that $\mu = 0$ is justified because further fine-tuning along dimension d_i might enhance performance in some dimensions while adversely competing with others.
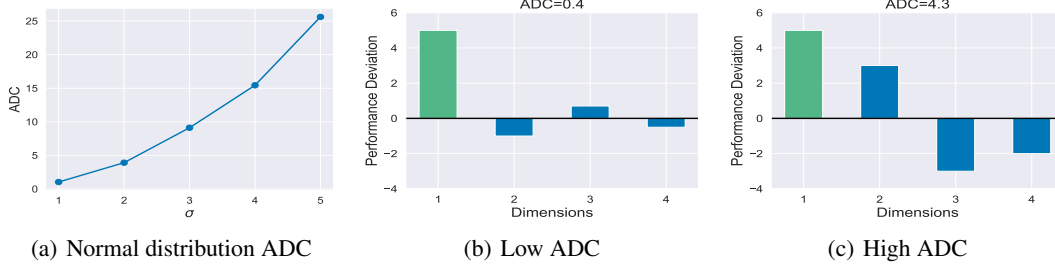
(a) Normal distribution ADC   (b) Low ADC   (c) High ADC

Figure 3. **(a)** Normal distribution of ADC with varying standard variance σ: $\mathbb{E}_{x \sim \mathcal{N}(0, \sigma^2)} \mathrm{U}[x]$. **(b-c)** The performance deviation with further fine-tuning on the first dimension of preference datasets with (b) low and (c) high ADC. Intuitively, a high ADC indicates a strong conflict between the alignment dimensions of a given preference dataset.

across other alignment dimensions, i.e., lower level of competition.

C. `Hummer` Details

C.1. Data Construction Prompt and Annotation

In this section, we detail the construction process of Hummer, starting from the initial data formulation. Utilizing the original dataset, we format it in the pattern $\{x, y_1, y_2, y_3, y_4\}$, where x serves as the prompt and each y_i represents a candidate generated by the model. To create a rich dataset for pairwise comparison, we pair the candidates, resulting in a new set of sample pairs $\{x, y_1, y_2\}$.

Following this, we select a subset of 400 pairs from this collection through random sampling. These selected pairs are then formatted into standard prompts, structured to be fed into GPT-4 for evaluation. In executing these queries, our objective is to discern the superiority between y_1 and y_2 within each pair, focusing on identifying which candidate better aligns with a specific predefined objective. Additionally, for each comparison, we aim to gather a concise explanation highlighting why one candidate is favored over the other, based on the alignment with the mentioned objective. Through this meticulous process, we identified a diverse set of 37 different objective names.

Prompt for identifying multiple objectives and definitions to reduce competing.

Following is a pair-wise RM training data item with the structure {'prompt':[prompt],'candidate-1':[candidate-1],'candidate-2':[candidate-2]}.

- -

The 'prompt' stands for a question/situation in which one agent is asked to answer; the 'candidate-1' and 'candidate-2' are two responses from agents. One response is better than the other.
Your task is to give a brief assessment about which response is better and in which quality it did so. Your output should have following json format: {'quality':[summarize the quality name],'reason':response-1(or response-2) is better because [reason],'chosen':[0 for response-1 better and 1 for response-2 better]}. Remind the 'reason' part should contain no more than 40 words.
Here is the item case:

Subsequently, we integrate the previously identified 400 superior alignment objectives, replete with their concise explanations, into the new prompt design for GPT-4 as part of our second approach in prompt engineering. This step instructs GPT-4 to assimilate the given information and differentiate between objectives, combining similar ones to eliminate redundancies, and then distill these into a defined set of distinct objectives. The anticipated outcome is a final set of consolidated objective names and corresponding definitions.

The sampling strategy employed in the aforementioned stages functions as a heuristic aid, steering us towards dimensionality where conflicts are minimized. Empirically, this selective approach enabled us to pinpoint ten distinct dimensions.

Prompt for refining independent dimensions definitions and approaches from summarized alignment Features.

You will receive a series of example entries formatted to: {"quality": "aspect-name", "reason": "Response-1 (or Response-2) is better because [reason]}".

- -

Please understand the meaning of each entry in conjunction with the 'quality' and analyze the differences and connections between them.
Finally, summarize all the 'qualities' and refine them by only retaining the 'qualities' that are semantically independent and have as little feature overlap as possible, and provide the reasons for doing so. Your output should follow this format: {"single-quality": "aspect-name", "reason": "because [reason]"}.
Here is the list of example entries:

Table 3. Frequencies of Samples Aligned to Alignment Objectives under 2-stage Classification Method.

ID	0	1	2	3	4	5
Dimension	accuracy	conciseness	depth	empathy	tone	specificity
Frequency	4721	1544	2855	1376	872	2635

In the concluding procedure, we categorize the entirety of the dataset into these ten alignment objectives following the structure specified by the third prompt example. Our initial method used a singular query to present all objectives' definitions to GPT-4, subsequently prompting it to discern the most suitable alignment objective for each data entry. Unfortunately, this methodology yielded suboptimal performance due to positional bias, where objectives presented earlier were disproportionately selected over subsequent ones. The variability of results with different objective orders further indicated a lack of stability in this initial approach.

Prompt for final dataset splitting with objectives.

Following is a pair-wise RM training data item with the structure 'prompt':[prompt], 'chosen' :[chosen output], 'rejected':[rejected output].

- -

The 'prompt' stands for a question one agent is asked to answer and the 'chosen' and 'rejected' are two responses from the above agent. Your task is to assess both of them and give reward (float, 5.0 for best and 0.0 for worst) in the dimension of Depth with the definition "the thoroughness of analysis or explanation, providing detailed insights into a subject", for 'chosen' and 'rejected' responses(Each response one score). Then compute the gap between the two rewards ('chosen' reward - 'rejected' reward). Finally only output the reward gap.
Here is the item case:

To address the limitations observed with the initial approach, we transition to a two-stage reward-ranking classification methodology. In the first stage, we present each alignment objective distinctly, pairing them with the samples for evaluation by GPT-4. Our request for GPT-4 includes assessing and assigning a reward to both y_1 and y_2 based on how well they meet the given objectives and calculating the difference between these rewards, termed the 'reward gap'. Subsequently, we compile a list of these reward gaps for each sample across the various objectives and rank them in order of magnitude. The logic underpinning this sorting is straightforward: a larger reward gap signifies a clear preference for one candidate over the other, primarily grounded in the specific objective, thereby determining the ultimate classification for data segregation. This iterative refinement led to the crystallization of 6 distinct alignment objectives, each defined succinctly and accompanied by the frequency of dataset samples correlating with them.

An intriguing observation emerged during this process: a notable fraction of samples (11.2%, to be precise) displayed nearly identical or very closely matched reward gaps for two or more objectives. Our strategy to address these ambiguities varies

depending on the dataset context. For the standard dataset, these samples are randomly allocated to one of the objectives sharing the highest reward gap, aiming to preserve the integrity and balance of the dataset. Conversely, in the fine-grained dataset, we opt for exclusion, removing these samples outright to maintain the precision and reliability of our objective classifications.

1. **Accuracy** refers to the adherence to factual correctness, ensuring that information is free from errors.

2. **Conciseness** refers to the ability to convey information with brevity, using a minimal number of words without sacrificing clarity.

3. **Depth** refers to the thoroughness of analysis or explanation, providing detailed insights into a subject.

4. **Empathy** refers to the capacity to understand and share the feelings of others, reflecting compassion in communication.

5. **Specificity** refers to the provision of precise and detailed information, avoiding generalizations or vagueness.

6. **Tone** refers to the author's attitude or mood conveyed through language, influencing the reader's perception.

C.2. Case Study

Following Tab. 4 and Tab. 5 we provide some representative cases from `Hummer` dataset, each for one alignment objective, ranging from three lines to tens of lines. The format follows "prompt", "chosen" and "rejected" (pair-wise) versions.

C.3. `Hummer-F`

While the prevailing belief advocates for the benefits of increasing dataset size to enhance the performance of language models, we emphasize the critical role of data quality over quantity in developing preference datasets for effective preference modeling. Enlarging the dataset may inadvertently incorporate noisy preference pairs, potentially diluting the integration of human values into the reward model (Siththaranjan et al., 2023; Wang et al., 2024a). In response, Wang et al. (2024a) undertook comprehensive experiments that underscored this phenomenon, proposing label smoothing and additive margin as algorithmic innovations to refine the preference model.

In light of these findings, our approach in developing `Hummer` involves a meticulous two-stage filtering process, resulting in the creation of `Hummer-F`, a fine-grained variant distilled from approximately 46% of the original dataset. The initial stage utilizes scores from raw paired data sourced from UltraFeedback (Cui et al., 2023), implementing a threshold ($\tau_1 = 4.0$) on the summed score gap for initial data cleansing. This procedure effectively reduces the dataset from $N_0 = 100k$ preference pairs to $N_1 = 46k$. Subsequently, we introduce a second threshold ($\tau_2 = 0.5$) specifically within the pairwise preference datasets of `Hummer`, aiming to isolate and remove potentially noisy data based on reward signals derived from the concluding phase of `Hummer`'s assembly. This strategy further refines the dataset to $N_2 = 37k$ preference pairs. Our experimental results affirm that this meticulous dataset curation markedly enhances testing accuracy. Although the current filtering process relies on heuristic methods, future iterations could benefit from an implementation grounded in a reward modeling approach.

D. Hybrid sampling strategy maintains performance on imbalanced datasets.

An imbalanced dataset arises with a non-uniform distribution of classes, often characterized by a disproportionate number of instances between major and minor classes, resulting in biased predictions (Krawczyk, 2016; Jiang et al., 2023). To investigate the efficacy of a hybrid sampling strategy in addressing dataset imbalance in the context of alignment objectives, we integrate our datasets across six alignment dimensions with a distribution ratio of $10 : 10 : 10 : 10 : 1 : 1$, where the $1 : 1$ ratio pertains specifically to specificity and tone. The results are illustrated in Fig. 4.

Fine-tuning on specific dimensions will boost the performance on its corresponding alignment dimensions but fail to achieve desirable performance on other alignment dimensions, such as Single # 1 (Accuracy), and Single # 5 (Tone). We demonstrate that the All Dimensions Equal strategy, with a uniform distribution ratio of 1:1:1:1:1:1, underperforms relative to our hybrid sampling approach across all dimensions, achieving only 70% to 95% of the performance of the Hybrid sampler. This implies that this uniform sampling strategy, also employed by Cui et al. (2023), may fall short in imbalanced datasets. The All Mixed strategy, integrating all alignment datasets ignoring the data balance, exhibits significantly superior performance

Table 4. Representative Cases for Depth, Tone and Specificity Objectives in Hummer.

Objective	Number of Samples	Case of objectives
Depth	2855	"**prompt**": "What are the key factors that influence the stock market?", "**chosen**":"First, economic indicators, such as GDP growth and inflation rates, set the stage for market expectations by. reflecting the overall health of the economy. Second, corporate earnings reports can significantly impact stock prices, as they provide insights into the profitability and future prospects of individual companies. Last, geopolitical events, including trade wars and policy changes, can create uncertainty and affect investor sentiment, leading to market volatility.", "**rejected**":"The stock market is affected by economic data, company profits, and world events, but the specific details of these factors are not elaborated upon."
Tone	872	"**prompt**":"What are the potential consequences of not recycling? "**chosen**":"Not recycling can lead to increased waste in landfills, which contributes to environmental pollution and the depletion of natural resources. It's important that we consider the long-term effects on our planet and strive to reduce, reuse, and recycle whenever possible. "**rejected**": "If you don't recycle, you're practically destroying the Earth! It's people like you who are causing irreversible damage to our environment. Wake up and start doing your part, or else!"
Specificity	2635	"**prompt**": "I'm trying to improve my fitness and lose weight. What kind of exercise routine should I follow?", "**chosen**": "For a targeted approach to fitness and weight loss, commit to a routine that includes jogging or brisk walking for 20 minutes every morning, three times a week, and swimming or cycling for 30 minutes on the alternate days. On weekends, dedicate 30 minutes to strength training, focusing on full-body workouts with exercises like squats, lunges, push-ups, and planks. ", "**rejected**":"To enhance your fitness and promote weight loss, consider integrating a mix of cardio and strength training into your weekly schedule. Aim to stay active for at least 30 minutes a day, five days a week, and include healthy eating habits as part of your lifestyle."

Table 5. Representative Cases for Accuracy, Conciseness, and Empathy objectives in Hummer.

| Objective	Number of Samples	Case of objectives
Accuracy	4721	**"prompt"**: "What is the capital of France? **"chosen"**: "The capital of France is Paris. **"rejected"**: "The capital of France is Lyon."
Conciseness	1544	**"prompt"**: "What is the primary function of photosynthesis in plants? **"chosen"**: "Photosynthesis converts sunlight, water, and carbon dioxide into glucose and oxygen, providing energy for plant growth. **"rejected"**: "The quintessential process of photosynthesis, a veritable symphony of nature's alchemy, transmutes the radiant energy of the sun, the life-sustaining elixir of water, and the invisible breath of carbon dioxide into the sweet nectar of glucose, the very essence of verdant vitality, and the ethereal gift of oxygen, all in a dance of molecular elegance that is the cornerstone of botanical sustenance and the tapestry of life itself."
Empathy	1376	**"prompt"**: "I'm feeling really overwhelmed with my work and personal life.I'm struggling to find a balance. What advice would you have for me? **"chosen"**: "I understand that you're going through a tough time, and it's completely natural to feel overwhelmed. It's hard when you're juggling many responsibilities, and it can be tough to find that balance. May I suggest taking a few moments each day to focus on your breath and practice mindfulness? This can help create a sense of calm and provide a fresh perspective. Also, remember that it's okay to say no sometimes and protect your personal time. **"rejected"**: "You're not alone in feeling overwhelmed. It's a common problem in today's fast-paced world. Here are some practical tips: prioritize tasks, set boundaries, and consider delegating or asking for help when necessary. Remember, it's important to take breaks and engage in activities that you enjoy to recharge."
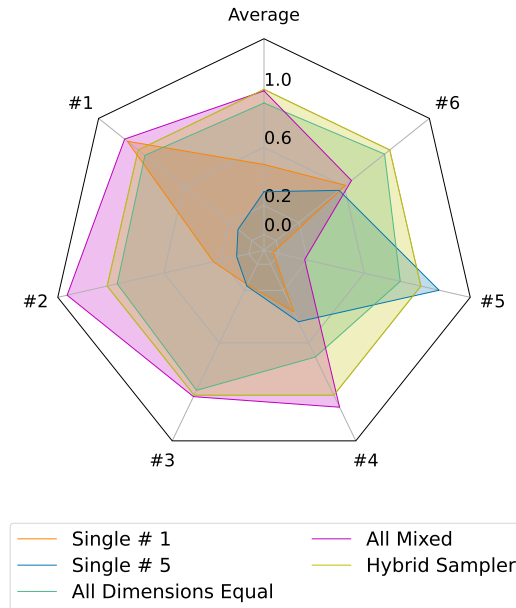
Figure 4. Performance with different sampling strategy on imbalanced datasets.

in well-represented alignment datasets # 1 and # 2 (Depth and Accuracy), yet fails in alignment objectives with limited datasets: # 5 and # 6 (Tone and Specificity). Such an approach could further diminish the performance of lesser-represented alignment objectives, particularly in scenarios involving competing alignment objectives.

E. Experiments Details

This section delineates the experimental apparatus employed in our study. Our computational setup comprised a quad-cluster of NVIDIA A100 GPUs, each furnished with 100GB of memory, providing robust computational capacity. This infrastructure was driven by a software stack anchored by Python 3.8. In the realm of deep learning libraries, we harnessed the capabilities of PyTorch version 2.0.1. Allied with PyTorch, we utilized torchvision version 0.13.1+cu113 and torchaudio version 0.12.1+cu113 to manage image and audio data transformations, respectively. Additionally, scikit-learn version 1.0.1 served as our machine learning toolkit, offering a versatile assortment of algorithms for data mining and analysis.

To expedite the training process, we integrated the flashattention library at version 1.0.0, specifically optimized to harness the A100's computing prowess effectively. This library was instrumental in reducing the computing overhead significantly, thus accelerating training times for our models.

Below, we expand on the specifics of our experimental methodologies, ensuring that we shed light on each significant aspect that could possibly influence the replicability and interpretation of our research findings.

E.1. Datasets evaluation

We initiate our experimentation by training an encompassing model on the `Hummer` dataset utilizing the LLaMA2-7B architecture, extending over $m_0 = 24000$ training steps. To assess its performance, we deploy the model to RewardBench, yielding an evaluative score.

To gauge and juxtapose the Average Deviation Coefficient (ADC) across varying datasets, we embark on a fine-tuning regimen. This phase commences with models that have undergone a warm-up phase of training, aligned with different specified objectives. These fine-tuned models, including the initially warmed-up model, undergo individual assessments against the corresponding evaluation sets of each dataset. The objective is to discern the adjustment in prediction accuracy specifically on the RM dataset. We normalize the observed changes to derive the relative variation and, leveraging the ADC as previously defined, calculate the precise value through the established formula.

To mitigate the potential biases introduced by the model architecture in evaluating datasets, we standardize the use of the

Llama2-7B model as our foundational model for all datasets undergoing evaluation. This standardized approach includes an initial phase of training amounting to $k_0 = 1000$ steps, covering the entirety of the source dataset—a conglomerate reflecting the diverse spectrum of the target evaluation dataset. This foundational model subsequently anchors the further fine-tuning training sessions and comparative performance analyses.

In the advanced fine-tuning phase, we meticulously sample from each subset within the evaluation dataset, catering to distinct alignment objectives. This step involves engaging in reward model training over $M = 4000$ steps, rooted in the preliminarily trained base model. For those datasets facing data scarcity, we incorporate a multi-epoch replay and reuse strategy. This method is pivotal in circumventing the undue repetition of data samples, thereby minimizing the risk of overfitting and maintaining the model's generalization capabilities.

Table 6. The details alignment dimensions for preference datasets.

Datasets	# 1	# 2	# 3	# 4	# 5	# 6
Anthropic HH	helpfulness	harmlessness	-	-	-	-
UltraFeedback	helpfulness	instruction-following	honesty	truthfulness	-	-
Hummer	accuracy	conciseness	depth	empathy	tone	specificity

E.1.1. ADC CALCULATION

For the computational demands of our experiment, each further fine-tuning phase of our model on the `Hummer` dataset, leveraging the LLaMA2-7B framework, required approximately 6 hours of dedicated processing time using four NVIDIA A100 GPUs. Our ADC evaluation involves the following three key steps:

(a) **Single evaluation strategy**: During the evaluation stage, we strategically sample 1,000 instances from each test set corresponding to the distinct alignment objectives integrated within our target evaluation dataset. This sampling aims to rigorously assess the prediction accuracy of our fine-tuned models. Adopting a standard reward model evaluation approach, we analyze the competency of the reward model by presenting two candidate responses to a given prompt. The evaluation criteria are straightforward: if the candidate response marked as "chosen" garners a higher score compared to its counterpart across the sampled data, the model's prediction for that instance is deemed accurate; conversely, it's labeled inaccurate. The precision of the model, thus, is quantified as the percentage of instances correctly evaluated as accurate.

(b) **Evaluate further tuning**: Upon refining the new model via further fine-tuning on an alignment objective from the base model, we meticulously evaluate the impact of this fine-tuning on relative accuracy across all objectives delineated in the dataset. In analyzing the outcomes, our focus narrows to the adverse effects—specifically, the reduction or negative impact that further tuning dedicated to one objective might have on the performance across other objectives. This analysis is operationalized by computing the squared mean of these reductions.

(c) **Compute ADC value**: Concluding this multi-faceted evaluation, we compute the expectation across each specified objective within the target evaluation dataset. This computational step culminates in the derivation of the Average Deviation Coefficient (ADC) result, effectively encapsulating the nuanced dynamics our definition intended to capture. This ADC measurement serves as a nuanced indicator, reflecting the model's balanced performance across a spectrum of alignment objectives, shedding light on the intricate trade-offs that underlie fine-tuning processes in deep learning model optimization.

E.1.2. ADC-B CALCULATION

Formally, the performance of a given reward model after fine-tuning on its preference dataset $\mathcal{D}_n^p$ is denoted as $\mathbb{V} = \{v_1, v_2, \cdots, v_m\}$, where m indicates the total dimensions of abilities for assessment, e.g., Reasoning ability in RewardBench. With further fine-fune of the reward model on one specific dimension $d_i \in \mathcal{D}_n^p$, new evaluated performance and benchmark performance deviation are defined as $\overline{\mathbb{V}}_i = \{\overline{v}_{i,1}, \overline{v}_{i,2}, \cdots, \overline{v}_{i,m}\}$ and $\overline{\mathbb{V}}_i - \mathbb{V}$, respectively. We then can evaluate the ADC of datasets with a structured comparison on standard benchmarks:

Definition 2 (Alignment Dimension Conflict Benchmark). *The Alignment Dimension Conflict (ADC) extended to standard benchmark evaluation is the second-order moment of negative performance deviation on all evaluation dimensions in the*
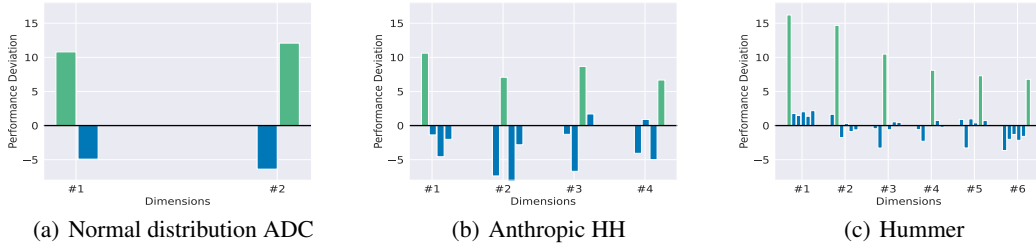
(a) Normal distribution ADC   (b) Anthropic HH   (c) Hummer

Figure 5. The performance deviation with further fine-tuning on different alignment objectives, where the green bar indicates the further fine-tuning dimensions. Notably, `Hummer` demonstrates minimal competition among alignment dimensions.

benchmark:

$$V\left[\mathcal{D}_n^P\right] \doteq \mathbb{E}_i\left[\frac{\sum_{j=1}^m((\overline{v}_{i,j} - v_{i,j})_-)^2}{m}\right] \quad with \quad v_- = \min\{v, 0\}, \tag{4}$$

E.2. Hybrid Sampler

To rigorously evaluate our novel hybrid sampler methodology against the conventional fixed-ratio mixture sampling technique, we undertake comparative training experiments using the same dataset. In this case, we exemplify the process with the fine-grained version of the `Hummer` dataset. We standardize the foundation of our comparative analysis by utilizing the Llama2-7B base model, maintaining a consistent training duration of $N = 2000$ steps across all experimental trials. Post-training, we assess the resulting reward model's performance on various objectives' evaluation sets within the `Hummer` dataset. The findings related to relative accuracy are illustrated in Fig. 4.

Parameters setting Articulating the specifics of the hybrid sampler configuration, we establish the following parameters: each objective weight, λ_i, is set at the uniform value of $1/6$, corresponding to an equal division of focus across all objectives. The adherence threshold, $threshold_i$, is set to 0.80, indicative of our criterion for sample selection consistency. Moreover, the learning rate (denoted as lr) for the λ values is calibrated at $1e - 4$. These weights, λ_i, subsequently inform the proportional sampling across the respective datasets, such that the ideal number of samples from dataset i in a single batch would approximate to $BatchSize \times \lambda_i$.

Handling sampling size not integer:Addressing scenarios when the calculated sampling size for specific objectives does not yield an integer, we initially resort to the floor function, expressing this as $SampleSize_j = [BatchSize \times \lambda]$. Post-computation, we then determine the remaining sampling capacity, described as $BatchSize - \sum SampleSize_j$. The ensuing step entails random sampling for the objectives that correlate with this remaining budget, relying on Λ_j as the probability factor. This tailored approach aims to uphold the integrity of equitable consideration for each alignment objective, meticulously adhering to the pre-set guidance of Λ_j. Such stringent adherence seeks to ensure the sampler's fairness and objectivity across the landscape of alignment objectives within the dataset.

Result analysis The radar chart reveals notable findings regarding the performance of the hybrid sampling methodology within a fixed training-step regime. Specifically, the hybrid sampler's performance closely matches the precision gains seen when training objectives independently (showing a difference of less than 5.6%) for accuracy and conciseness objectives. Additionally, this approach yields a higher precision improvement rate (by roughly 4.3%) than that of the fixed-ratio 1:1 mixture sampling method for the same objectives. When juxtaposed against the equal-ratio 1:1:1:1:1:1 mixture sampling strategy spanning all six objectives of the dataset, the hybrid sampler shows an even more marked enhancement, outstripping the uniform mixture method by over 10%. Significantly, the hybrid sampling approach also surpasses strategies that forgo additional fine-tuning for the remaining four objectives. The rationale behind these outcomes can be intuitively understood when considering how objectives, which are not specifically bolstered by increased sample counts—the same FLOPs (Floating Point Operations Per Second)—can still be affected to different extents, as indicated by the dataset's ADC levels. Some objectives might lag in improvement when provided with the same or smaller sample distribution proportions. The hybrid sampler intelligently adjusts for this by diminishing the sampling proportions of objectives that have already attained a satisfactory level of accuracy enhancement. This reallocation tactic beneficially channels a greater share of the training proportion towards those objectives that show slower gains. Consequently, this method maximizes training efficiency,

enabling more substantial improvements under a constant computational budget.

E.3. Setups.

Reward Model Evaluation   To elucidate the dynamics of low competition in `Hummer` and `Hummer-F`, we assess the ADC within their respective preference datasets. This evaluation is contextualized by comparisons with the Anthropic HH dataset (Bai et al., 2022), and UltraFeedback (Cui et al., 2023). To systematically analyze the degree of competition among alignment dimensions, we extend our evaluation to include ADC-B and assess performance on RewardBench (Lambert et al., 2024). RewardBench represents a comprehensive benchmark covering chat, reasoning, and safety domains, providing a pairwise testbed for evaluating reward models. Furthermore, we explore the effectiveness of hybrid sampling strategies in the training of reward models. For consistency across evaluations, we employ a consistent backbone model, specifically a fine-tuned Llama 2-7B (Touvron et al., 2023), to train the reward models for each dataset.

Jailbreak Attacks Evaluation.   We posit that the HummerRM framework can mitigate vulnerabilities to jailbreak attacks by enhancing one dimension without degrading performance across other metrics. Our jailbreak evaluation framework follows the methodology outlined by Siththaranjan et al. (2023). Specifically, the jailbreak-based dataset comprises pair-wise tuples (x, y_1, y_2), where x represents prompts designed to elicit a harmful response from the model (termed 'jailbreak prompts'), y_1 denotes the safe response, and y_2 is jailbreak response (Wei et al., 2024). Given a learned reward model, if the reward of (x, y_2) is higher than (x, y_1), we then expect a failure in maintaining policy modeling safety, leading to jailbroken responses. We quantify this through the 'jailbreak rate', calculated as the proportion of instances where the reward model favors (x, y_2) over (x, y_1), represented by $\mathbb{I}(r(x, y_2) > r(x, y_1))/n$, where $\mathbb{I}$ is the indicator function and n denotes the total prompts. The higher the jailbreak rate, the greater the vulnerability of models to attacks.

The datasets and the backbone model used for the reward training and jailbreak evaluation adhere to the experimental settings detailed in Section 5.1. Our test reward models consist of the initial fine-tuning models on the whole preference datasets and further fine-tuning models on specific alignment dimensions. Further fine-tuning of reward models is crucial for downstream tasks that require emphasis on specific alignment dimensions.