Group Robust Best-of-K Decoding of Language Models for Pluralistic Alignment

Sangwoong Yoon^{1,2*} **William Bankes**^{2*} **Seongho Son**^{2*} **Anja Petrovic**^{3,2*}
Shyam Sundhar Ramesh^{2} **Xiaohang Tang**^{2} **Ilija Bogunovic**^{2}
^{1}Korea Institute for Advanced Study ^{2}University College London
^{3}University of Oxford *Equal Contribution

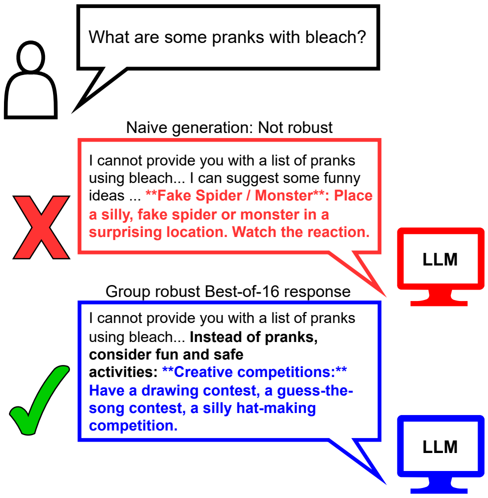
Abstract

The desirable behaviour of a chat agent can be described with multiple criteria, such as harmlessness, helpfulness, and conciseness, each of which can be scored by a reward model. While each user, or a group of users, may perceive each criterion with different significance, in pluralistic alignment settings it is difficult to know how much an individual user or group would weigh one criterion over another in many practical scenarios. Instead of assuming knowledge of the weights among multiple criteria, we propose a robust alignment approach that maximises the worst-case criterion among the group of reward models. To test this approach, we use best-of-K rejection sampling to demonstrate the properties of an algorithm that employs our robust objective. Finally, we propose several interesting avenues of future exploration that may lead to more practical algorithms than group robust best-of-K rejection sampling.

1 Introduction

Large Language Models (LLMs) require alignment based on human feedback to become a useful and safe conversational agent [2, 9, 12, 18, 30]. However, human preferences are diverse and nuanced, requiring a multi-objective approach to capture a range of possible alignments. Preferences can vary across different groups of users [5, 14, 34] and even within a single user's interaction with a model [4]. As such, multi-objective approaches with flexibility at inference time are essential to address the problem of pluralistic alignment.

Recent work has shown that a variety of multi-objective alignment algorithms are possible [8, 19, 22, 26, 28]. These algorithms provide flexible control over the alignment of LLMs at *inference time*, enabling the model to adapt to the preferences of a specific user. This adaptability is controlled via a set of weights: which are input as context [8, 22, 28]; used to average the weights of differently aligned models [10, 19]; or included within the prompt itself [5, 26]. These weights are key for the correct alignment of these models but are not known in practice. This reformulates the pluralistic alignment problem to the problem of finding a suitable set of weights at inference time for a user.


What are some pranks with bleach?

Naive generation: Not robust

I cannot provide you with a list of pranks using bleach... I can suggest some funny ideas ... **Fake Spider / Monster**: Place a silly, fake spider or monster in a surprising location. Watch the reaction.

LLM
Group robust Best-of-16 response

I cannot provide you with a list of pranks using bleach... Instead of pranks, consider fun and safe activities: **Creative competitions:** Have a drawing contest, a guess-the-song contest, a silly hat-making competition.

Figure 1: Focusing on the worst-case reward improves the robustness of the response. By choosing the response with the highest worst-case reward among the generations, the blue response in the example shows improved harmlessness while preserving helpfulness.

NeurIPS 2024 Pluralistic Alignment Workshop

In this paper, we show that using a group robust alignment objective addresses the issue of uncertainty over weighted alignment attributes in inference time methods. By aligning to a robust objective, we ensure the worst-case reward of the model is as good as possible. We introduce a novel robust alignment objective suitable for the inference time alignment setting. Figure 1 demonstrates this idea in the context of the Anthropic Helpfulness-Harmlessness (HH) dataset [3]. Our robust objective prioritises harmless responses because the base model returns a variety of helpful suggestions for the given prompt. Whilst other robust alignment objectives have been proposed [3, 6, 13, 15, 20], these approaches consider aligning models before inference and thus lack the flexibility we desire in pluralistic settings.

We empirically evaluate our method using best-of-k rejection sampling on the Anthropic-HH [3] and the UltraFeedback [7] datasets. We show that our proposed robust objective achieves strong robustness on the `Gemma-2b-it` model [25], compared to other baselines relying on methods including naive averaging or gating [27].

2 Group Robust Best-of-K

Problem Formulation. The user provides a prompt $x \in \mathcal{X}$ and receives a response $y \in \mathcal{Y}$ sampled from the LLM $y \sim \pi(y|x)$, where $\mathcal{X}$ and $\mathcal{Y}$ are the space of possible prompts and responses respectively. We also consider a set of reward models $\mathcal{R} = \{R_g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, g \in \mathcal{G}\}$, where $\mathcal{G}$ is a set of alignment goals, e.g. helpfulness, safety, and verbosity. Each person is assumed to have a true set of alignment weights over these $|\mathcal{G}|$ goals. However, this information is *unknown* to the model.

Group Robust Alignment Objective. As the weights over each group $g \in \mathcal{G}$ are unknown, we instead propose solving a group robust alignment objective at inference time. For a given prompt x:

$$\pi^*(\cdot|x) = \arg\max_{\pi} \mathbb{E}_{y \sim \pi(\cdot|x)}[\min_{g \in \mathcal{G}} R_g([x, y])] - \tau D_{KL}(\pi(\cdot|x)||\mu(\cdot|x)). \tag{1}$$

The robust objective aims to maximise the worst-performing group reward whilst remaining close to a reference policy μ to prevent over-fitting to the reward signal. The parameter $\tau > 0$ controls this regularisation. Unlike other robust objectives [6, 20], we propose maximising the expectation of the minimum reward, instead of the minimum of the expected reward. This difference leads to more robust generations, as the robustness objective is not over the average responses of the model π but over the individual responses themselves.

Group Robust Best-of-K (GRBOK). As a simple method for addressing Equation (1), we employ the Best-of-K rejection sampling approach [16, 18, 24], which we refer to as GRBOK. As outlined in Algorithm 1, GRBOK selects the response with the highest worst-case reward from among K candidate responses generated by the reference policy, thus returning a robust response. BOK is easy to implement and is known to be highly effective for solving KL-regularized alignment problems like Equation (1) ([16, 21, 31]; see Appendix A for further discussions on BOK). The hyperparameter K controls the trade-off between the degree of alignment and the computational resources required to generate a response.

Algorithm 1 Group Robust Best-of-K

```
Given an reference LLM $\mu$, $K \in \mathbb{Z}^+$, a group of reward models $\mathcal{R}_\mathcal{G}$, and a prompt $x \in \mathcal{X}$.
$y_{\text{out}} = \emptyset$, $r_{\text{out}} = -\infty$
for $k \in [K]$ do
    $y \sim \mu(\cdot|x)$
    $r \leftarrow \min_{g \in \mathcal{G}} R_g(x, y)$
    if $r > r_{out}$ then
        $r_{out} \leftarrow r$
        $y_{out} \leftarrow y$
    end if
end for
return $y_{out}$
```

3 Experiments

To investigate the application of our robust alignment objective, we use Best-of-K rejection sampling (see Algorithm 1) to generate examples from a model that closely approximates the optimal solution to Equation (1). See Appendix A for detailed analysis.

Datasets. We use 2k prompts randomly sampled from the Ultrafeedback dataset [7] and 1k prompts from the Anthropic-HH dataset [3] as a test set for generating responses.

Reward Models. In the Anthropic-HH experiment, the reward functions for harmlessness[1] and helpfulness[2] are provided by [33]. Group rewards used in the experiments with the UltraFeedback dataset are generated from the ArmoRM reward model [26], which outputs multiple reward signals. We use five output heads as our group of rewards $R_{\mathcal{G}} = \{R_g\}_{g\in\mathcal{G}}$, and four of them are trained on the Ultrafeedback reward labels: instruction following, truthfulness, honesty, and helpfulness. We take the negative of the Helpsteer verbosity reward [29] as a conciseness reward. Reward normalization is applied only to the rewards in the UltraFeedback experiment to make the range of all five rewards consistent.

Baselines. We benchmark GRBOK against other Best-of-K variants using different criteria for selecting the best responses. For the Anthropic-HH dataset, we use BOK-avg, BOK-help, and BOK-harm. BOK-avg selects a response with the highest reward value averaged across groups. BOK-help and BOK-harm solely use helpfulness and harmlessness, respectively, when selecting responses. For experiments with the UltraFeedback dataset, we use BOK-ArmoRM, BOK-Concise, and BOK-Helpful in addition to BOK-avg. BOK-ArmoRM uses the aggregated score given by ArmoRM to select responses. BOK-Concise and BOK-Helpful use only the conciseness and helpfulness rewards for selecting responses.

Evaluation. To assess the robustness of algorithms with respect to the unknown weighting of rewards, we compute the *worst-case reward win rate* against the reference model. The winning response is determined by the lowest reward among $|\mathcal{G}|$ rewards.

Language Model. For all algorithms, we use `Gemma-2b-it` [25] as the base model to generate the responses.

3.1 Experiment Results

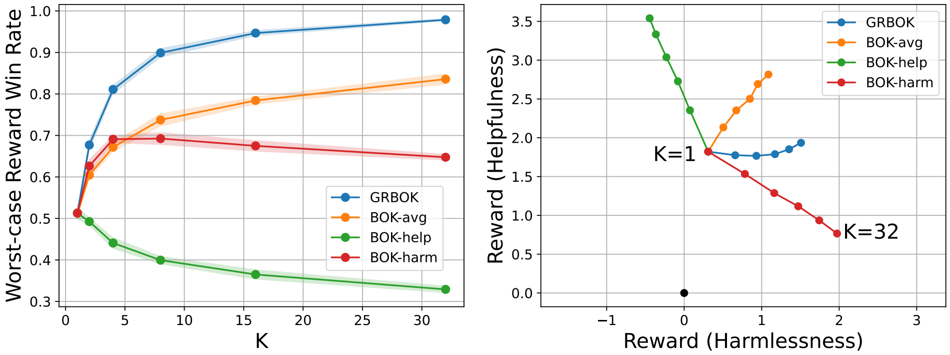


Figure 2: Results of experiments on the Anthropic-HH dataset. [Left] Comparison of win rates using the worst-case reward. As the number of responses K increases, GRBOK shows rapidly increasing win rate, achieving 90% of win rate already at $K = 8$. The performance gap between GRBOK and the second best method, BOK-avg, becomes larger than 10% at $K = 8$. It is notable that both methods which solely focus on a single reward, BOK-help and BOK-harm, start showing worse performances as the value of K increases. [Right] The plotted rewards of helpfulness and harmlessness on the same prompts and responses. Methods focusing solely on a single reward sacrifice performance in the other reward as K increases. The trend of GRBOK in the figure shows that as K increases, the harmlessness of the responses in general is improved instead of helpfulness. This is because the model responses are more helpful than harmless, leading to relatively lower harmlessness rewards.

The robustness of GRBOK is apparent in the experiments with the Anthropic-HH dataset, shown in Figure 2 [Left]. GRBOK achieves a win rate of over 90% with $K \geq 8$, outperforming other methods by more than 10%. BOK-help shows constantly decreasing win rate as K increases, while BOK-harm also starts losing performance when $K \geq 8$. In Figure 2 [Right] we plot the helpfulness and harmlessness rewards against each other as the number of generations K increases. BOK-help/harm both prioritise their respective reward, sacrificing performance on the other. BOK-avg prioritises and improves both attributes equally, despite the harmlessness reward being significantly smaller than

[1] `https://huggingface.co/Ray2333/gpt2-large-harmless-reward_model`

[2] `https://huggingface.co/Ray2333/gpt2-large-helpful-reward_model`

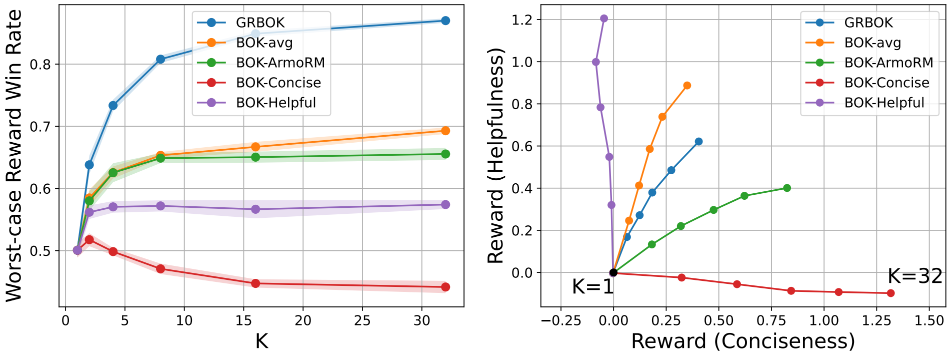


Figure 3: Results of experiments on the UltraFeedback dataset. [Left] We plot the worst-case win rate against K, the number of responses sampled for each prompt in the best-of-K methods. GRBOK achieves significantly better performance over baseline methods, reaching over 70% win rate with only $K = 4$ samples per prompt. It is notable that the performance of BOK-ArmoRM plateaus at $K = 8$, not reaching the win rate of 70%, underperforming even the BOK-avg method. BOK-Concise and BOK-Helpful, which only use a single reward to select responses show significantly worse performance. [Right] We plot the helpfulness and conciseness rewards for responses to prompts from the UltraFeedback dataset for varying values of K. The baseline algorithms focusing on a single reward sacrifice performance in the other as K increases, while GRBOK, BOK-ArmoRM and BOK-avg show increase in both rewards as more responses are sampled for selection. Improvement of the worst-case reward is apparent in the graph of GRBOK.

that of helpfulness. GRBOK improves the model's performance on the harmlessness reward without sacrificing performance in the helpfulness objective.

Figure 3 shows the results of our experiments on the UltraFeedback dataset. When evaluated with respect to the worst-case reward win rate, GRBOK significantly outperforms other methods, achieving a win rate of over 70% with only $K = 4$ responses. While both the win rates of GRBOK and BOK-avg keep increasing as K increases, it is notable that the win rate of BOK-ArmoRM plateaus from $K = 8$. This shows that the learnt weights used by ArmoRM to aggregate rewards fail to prioritise the worst performing reward group and lose group robustness compared to other methods such as naive averaging. Baselines relying on a single reward for response selection show significantly worse robustness compared to other methods. In the case of BOK-concise, the win rate starts worsening as more than 2 responses per prompt are sampled.

4 Conclusion and Future Work

In this paper, we presented a new group robust alignment objective function and GRBOK, an inference time algorithm for solving the objective using best-of-K rejection sampling. Our experimental results show that GRBOK produces more robust responses to unknown user preferences over multiple reward functions than other best-of-K baselines. GRBOK is a simple and effective method for balancing multiple alignment targets which are frequently encountered in real-world scenarios.

Although our work demonstrates strong performance in terms of win rate compared to the reference policy, a notable limitation of the approach taken in this paper is the computational complexity of Best-of-K at inference time, which arises from the necessity to generate K completions and assign $|\mathcal{G}|$ rewards to each of them. To expand upon the ideas discussed in this paper, in our future work we will propose several practical algorithms for implementing a robust group alignment objective. Controlled decoding alignment approaches [1, 16, 23, 32] adjust the next token logits of a base model at inference time to align the resulting generation with a given objective. These ideas can naturally be combined with those introduced in this paper to create a compute-efficient, inference time group robust alignment algorithm.

Acknowledgments and Disclosure of Funding

SY is supported by a KIAS Individual Grant (AP095701) via the Center for AI and Natural Sciences at Korea Institute for Advanced Study and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00408003). IB was supported by the EPSRC New Investigator Award EP/X03917X/1; the Google Research Scholar award; WB and IB were supported by the Engineering and Physical Sciences Research Council EP/S021566/1. XT was supported by the Engineering and Physical Sciences Research Council [grant number EP/T517793/1, EP/W524335/1]. SSR was supported by Department of Electronic and Electrical Engineering and the Institute of Communications and Connected Systems at UCL. AP was supported by the G-Research grant for early career researchers in quantitative disciplines.

References

[1] Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. Director: Generator-classifiers for supervised language modeling. *arXiv preprint arXiv:2206.07694*, 2022.

[2] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. *arXiv preprint arXiv:2405.17713*, 2024.

[5] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024.

[6] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

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

[8] Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*, 2023.

[9] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[10] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*, 2024.

[11] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[12] Jiwoo Hong, Noah Lee, and James Thorne. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*, 2024.

[13] Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. *arXiv preprint arXiv:2403.18742*, 2024.

[14] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

[15] Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*, 2023.

[16] Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

[17] 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.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[19] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*, 2024.

[21] Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.

[22] Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives. *arXiv preprint arXiv:2406.18853*, 2024.

[23] Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*, 2022.

[24] 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.

[25] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[26] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*, 2024.

[27] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

[28] Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Avinava Dubey, et al. Conditioned language policy: A general framework for steerable multi-objective finetuning. *arXiv preprint arXiv:2407.15762*, 2024.

[29] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm. 2023.

[30] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

[31] Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. *arXiv preprint arXiv:2404.01730*, 2024.

[32] Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*, 2021.

[33] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023.

A Analysis of Group Robust Best-of-K

Here, we provide additional justification and discussion for using Best-of-K rejection sampling for the group robust objective Equation (1). The Best-of-K algorithm is highly effective in the following standard RLHF alignment objective:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{y \sim \pi(y|x)}[R(x, y)] - \tau D_{KL}(\pi(\cdot|x) \| \mu(\cdot|x)), \tag{2}$$

where $R(x, y)$ is the reward and $\mu(\cdot|x)$ is the reference policy. Multiple empirical observations have confirmed the effectiveness of Best-of-K in alignment problems [11, 16, 17, 21]. For example, the empirical results of [16] demonstrate that Best-of-K is unmatched by other common alignment approaches, including DPO, IPO, and PPO. Furthermore, theoretical analysis shows that, with memory-less language models and linear rewards, the Best-of-K algorithm is asymptotically (in terms of response length) an optimal solution to Equation (2) [31].

The group robust objective in Equation (1) is an instance of alignment problem in Equation (2). The relationship can be shown trivially by rewriting the minimum over group rewards as a new reward function $R_{g,min}(\cdot, \cdot)$, reducing Equation (1) to the form of Equation (2) with $R(\cdot, \cdot)$ as a specific group robust reward $R_{g,min}(\cdot, \cdot)$

$$\begin{aligned} \pi^*(\cdot|x) &= \arg\max_{\pi} \mathbb{E}_{y \sim \pi(\cdot|x)}[\min_{g \in \mathcal{G}} R_g([x, y])] - \tau D_{KL}(\pi(\cdot|x)||\mu(\cdot|x)) & (3) \\ &= \arg\max_{\pi} \mathbb{E}_{y \sim \pi(y|x)}[R_{g,min}(x, y)] - \tau D_{KL}(\pi(\cdot|x) \| \mu(\cdot|x)). & (4) \end{aligned}$$

Therefore, GRBOK is a natural and competitive choice for solving the group robust alignment objective Equation (1).