Group Preference Optimization: Few-Shot Alignment of Large Language Models

Siyan Zhao, John Dang, Aditya Grover
Department of Computer Science, University of California, Los Angeles
{siyanz,john.dang,adityag}@cs.ucla.edu

Abstract

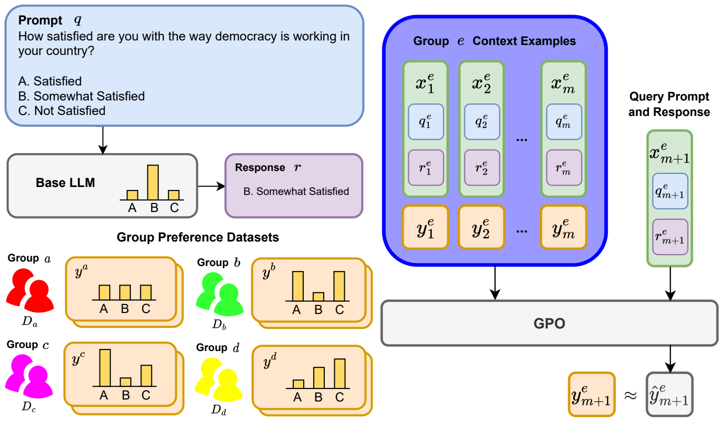
Applications of large language models (LLMs) often demand nuanced judgments that vary among different groups. Existing alignment algorithms can be costly, requiring extensive group-specific data and computation. We present Group Preference Optimization (GPO), a framework that efficiently aligns LLMs to group preferences using a few-shot approach. In GPO, we augment the base LLM with an independent transformer module to predict the preferences of a group for the LLM generations. For few-shot learning, this module acts as an in-context autoregressive transformer and is trained via meta-learning on several groups. Through empirical validation on opinion adaptation tasks involving US demographic groups, global countries, and individuals, GPO demonstrates superior alignment performance, requiring fewer group-specific preferences and reduced training and computational resources, surpassing existing strategies like in-context steering and fine-tuning.

1 Introduction

Large Language Models (LLMs) are increasingly being employed for a wide variety of domains, with use-cases including creative writing and chatbots [36, 34, 21, 2, 3, 7, 5]. Many of these applications are inherently subjective and require generations that cater to different demographics, cultural and societal norms, or simply individual preferences [14, 38, 31, 6, 10]. LLMs can represent diverse opinions due to broad training data [12, 11, 29], but they require steering to express opinions catering to user needs. This brings forth the key question studied in this work: *How do we efficiently adapt LLMs to align closely with the opinions of specific interest groups?* Prior research explores two main approaches: prompt engineering, which involves crafting suitable prompts and in-context examples but may be labor-intensive and less effective for complex behaviors [7, 34, 8, 39, 27, 23, 18, 29], and gradient-based alignment methods, which augment or fine-tune LLMs with additional reward models but require substantial supervision and face challenges when dealing with numerous target groups with limited supervision [21, 2, 12, 4, 1, 32, 3, 35, 37, 30, 26].

We introduce *Group Preference Optimization* (GPO), a few-shot framework for aligning Large Language Models to opinions and preferences of desired interest group(s). At its core, GPO treats LLM alignment as a few-shot adaptation in the LLM's embedded space, augmenting an arbitrary base LLM with an independent few-shot preference module. This module is parameterized via an independent transformer and trained to explicitly perform in-context supervised learning to predict preferences (targets) given joint embeddings (inputs) of prompts and corresponding LLM responses. The use of embeddings guarantees that the preference module can effectively process in-context examples where each example is itself a potentially long sequence of prompt and generated response. In-context learning further provides the ability to efficiently adapt to new, unseen groups at test-time with only a handful of examples. See Figure 1 for an illustration. Finally, we incorporate various architectural design choices to guarantee permutation-specific inductive biases, building on recent

37th R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023(NeurIPS 2023).



(a) (b)

Figure 1: Overview of GPO. **(a):** Group alignment aims to steer pretrained LLMs to preferences catering to a wide range of groups. **(b):** Once trained, GPO provides a few-shot framework for aligning any base LLM to a test group given a small amount of in-context preference data.

work in in-context learning [20]. Once learned, the learned module can serve as a drop-in replacement for a reward or preference function for policy optimization and re-ranking algorithms.

We validate GPO's effectiveness in aligning language models with the opinions of 22 diverse US demographic groups in the OpinionQA dataset [29] and 14 global countries in the GlobalOpinionQA dataset [11], using two different base language models: Alpaca 7B [34] and Llama2 13B chat [36]. GPO consistently outperforms various baselines, showing a 7.1% improvement when adapting to US demographic groups and a 8.4% improvement when aligning with global countries. Moreover, GPO excels in adapting to individual preferences compared to other methods.

2 Group Preference Optimization

2.1 Problem Setup and Related Works

An LLM expresses a probability distribution over natural language, denoted as π. To accomplish any task, users prompt the LLM with a query q to generate a response r from the conditional distribution $\pi(\cdot \mid q)$. Rather than decoding responses from a single distribution $\pi(\cdot \mid q)$, our goal in this work is to align the language model to the preferences of a desired target group $g^* \in G$. We adopt a fairly general definition of a *group* to refer to any collection of agents (e.g., demographic groups, individuals), and represent all possible groups as G. For training, we assume that we are given access to preference datasets for a finite set of training groups G_{train}. In real-world scenarios, while there can be many groups (e.g., varied demographics), preference data per group is typically limited.

Existing approaches for steering LLMs are challenging to apply for group alignment, especially when the groups are complex and per-group supervision is scarce. Prompt engineering approaches are computationally efficient but rely on heuristics and have limited generalization [16, 15, 9, 39, 18, 23], and it has been shown to yield limited gains in aligning LLMs to complex groups on challenging survey datasets [29, 11]. Gradient-based alignment includes methods like supervised fine-tuning and preference-based fine-tuning, but they require substantial group-specific data and face challenges in optimization [21, 40, 30, 33, 28]. We provide additional discussion of related work in Appendix J.

2.2 Proposed Method

We desire an alignment approach that generalizes to a wide variety of groups, even when constrained by the amount of per-group supervision. Accordingly, we view group alignment as a few-shot learning problem and cast it in the framework of in-context meta-learning. For each training group $g \in G_{\text{train}}$, we represent its preference dataset as $\mathcal{D}_g = \{(x_1^g, y_1^g), \ldots, (x_n^g, y_n^g)\}$ where y_i^g denotes the preference of group g to a pair of input prompt query and LLM response $x_i^g = \pi_{\text{emb}}(q_i^g, r_i^g)$. Here, π_{emb} can be the language model embedding function or an identity function that maintains the input's raw textual format. Note that while the inputs x^g can be shared across different groups (e.g., universal surveys), the preferences are different for each group. At test-time, our goal will be steer the default LLM distribution to a new distribution, say π_{g^*}, given a preference dataset $\mathcal{D}_{g^*}$ for the target query

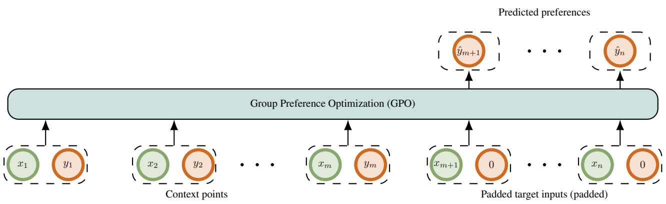


Figure 2: Illustration of the GPO architecture for a sequence of n points, with m context points and $n - m$ target points. The context $(x_{1:m}, y_{1:m})$ serves as few-shot conditioning for GPO. GPO processes the full sequence using a transformer and predicts the preference scores $\hat{y}_{m+1:n}$.

group g^*. For brevity of presentation, we consider the preference to be a real-valued scalars. Our framework extends to other kinds of responses and preferences, such as short-answer questions (e.g., MCQs) and relative pairwise responses, as discussed in Appendix I. Given the setup, GPO aligns to groups with a learned few-shot preference model. This model, once trained, allows LLM updates using standard preference optimization or reweighting algorithms (e.g., Best-of-N, PPO [30]).

We parameterize GPO via a transformer and train it to perform in-context learning on the training preference datasets, as shown in algo box **??**. Given a training group $g \in G_{\text{train}}$, we randomly split its preference dataset $\mathcal{D}_g$ into a set of m context points and $n - m$ target points, where $n = |\mathcal{D}_g|$ is the size of the preference dataset for group g. Thereafter, GPO is trained to predict the target preferences $y^g_{m+1:n}$ given the context points $(x^g_{1:m}, y^g_{1:m})$ and target inputs $x^g_{m+1:n}$. Mathematically, we can express the objective as:

$$L(\theta) = \mathbb{E}_{g,m}\left[\sum_{i=m+1}^{n} \log p_\theta(y_i^g \mid x^g_{1:n}, y^g_{1:m})\right] \tag{1}$$

where the training group $g \sim G_{\text{train}}$ and context size $m \sim \text{Uniform}(1, n-1)$ are sampled uniformly. Figure 2 shows an illustration. For decoding, we assume target preferences are conditionally independent given context and target inputs. In our preliminary experiments, we also investigated alternatives which model the dependencies. We did not find any noticeable improvements and hence use Eq. 1. Following [20], we adjust GPO's transformer to consider permutation invariance over in-context examples by discarding standard positional encodings. We address the loss of pairwise relations between (x_i, y_i) by concatenating them into a single token. Target inputs are padded with a dummy token, e.g., 0. We use a masking strategy, allowing context pairs to self-attend while padded targets attend only to context points, maintaining the conditional independence in Eq. 1. While GPO employs in-context learning, it differs from in-context prompting in base LLMs, which doesn't alter the LLM's parameters. Instead, GPO learns a few-shot model to enhance the base LLM using user preferences. These methods are complementary, allowing interchanged prompts (as in the ablation study in Appendix **??**), including in-context examples, in input x. Additionally, the effective sequence length for the transformer can grow significantly if we use raw representations of prompts and responses within each input x. To scale to long dataset contexts, we propose using embedded representations of text within x by concatenating the prompt and response and computing their joint embedding using the base LLM, with details provided in ablation study in Appendix E.

3 Experiments

Datasets, Evaluation Metrics and Baselines While assessing varied opinions in open-ended questions is intricate and demands extensive human labeling, closed-ended responses, such as multiple-choice questions, provide a consistent way to gather diverse views, minimizing evaluation ambiguity. Surveys like *OpinionQA* [29], spanning 22 US demographic groups, and *GlobalOpinionQA* [11], featuring questions from 14 countries, have highlighted LLMs' limitations in accommodating diverse groups. These datasets are pivotal for benchmarking group alignment advancements. As survey questions overlap among groups, we use x_i for simplicity. Further dataset specifics are in Appendix C. Next, we construct group g preference dataset $\mathcal{D}_g$ from the survey data. Given a set of survey questions Q answered by groups G, each question $q \in Q$ presents T unique answer options, considered as responses $\{r_i\}_{i=1}^T$, yielding a set of T viewpoints $\{x_i\}_{i=1}^T = \{q, r_i\}_{i=1}^T$. The preference score y_i^g for each viewpoint x_i aggregates the responses from group g and is normalized to produce the distribution vector $P_g(q)$ such that $\sum_{i=1}^T y_i^g = 1$. Repeating this process for all n questions in Q yields $\mathcal{D}_g$. During training and testing, all viewpoints from the same question belong to either the context or target set. We apply a softmax layer to predictions for each question, yielding normalized preference scores for the survey questions. We measure the alignment between opinion distributions, P_1 and P_2, using the *Alignment Score* over a question set Q. Depending on the dataset structure, we choose an appropriate similarity function. For OpinionQA, we use the Wasserstein Distance. For GlobalOpinionQA, we utilize the Jensen-Shannon Distance as suggested by its originating study. Further specifics are elaborated in Appendix C. We evaluate our method against several baselines using two base LMs: Alpaca-7B [34] and Llama2-13B chat version [36]. Baselines include: **Uniform**, assuming equal preference scores for all options; ***LM Base***, extracting the LM's default opinion distribution as in Santurkar et al. [29], Durmus et al. [11]; ***LM Steered***, employing diverse prompting strategies; ***Few-shot Prompt***, appending group preference examples to the prompt; ***SFT per group***, fine-tuning the LM per group; ***Reward Model***, training a per-group reward model; and ***In-Context Finetune***, exploring whether the LM's few-shot in-context alignment ability can be fine-tuned. For a comprehensive baseline description, see Appendix G.

3.1 Results and Discussion

On both OpinionQA and GlobalOpinionQA datasets, GPO consistently outperformed other methods in aligning with unseen groups, showcasing superior adaptability to diverse US demographics and global countries. As in Figure 3, GPO achieved alignment score improvements of 7.1% and 8.4% over the closest competitor, the *In-context Finetune* method, for the two datasets respectively. Figure 4 qualitatively illustrates GPO's adept alignment to specific group preferences in OpinionQA using just 15 context samples, highlighting its ability to align beyond learning general dataset tendencies. Further exploring scalability, we evaluated various methods using different in-context example sizes, as shown in Appendix Figure 5. While methods like *Few-shot Prompt* and *In-context Finetune* show incremental or plateauing performance, GPO stands out by efficiently adapting to groups with fewer than 10 preference context samples in the GlobalOpinionQA dataset. Additionally, variations in individual opinions can manifest even within the same demographic groups [15]. Motivated by this, we conducted experiments across 15 topics from the OpinionQA dataset. GPO consistently outperformed other methods in adapting to individual preferences as shown in Appendix Figure 6. For more details and comprehensive discussion on the results, please refer to Appendix section B.

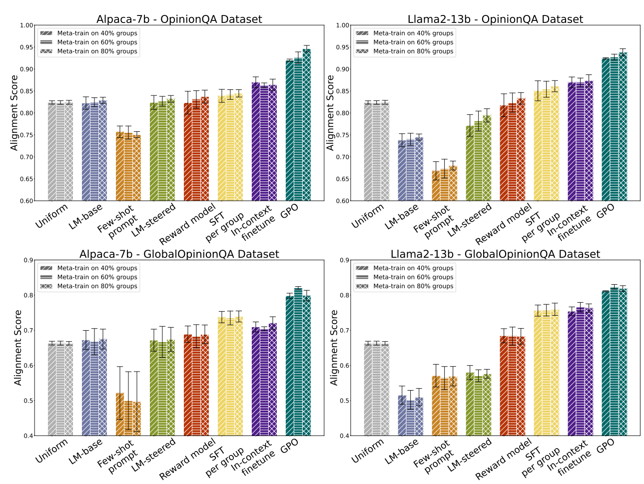


Figure 3: Alignment score comparisons on the OpinionQA dataset and GlobalOpinionQA dataset with Alpaca-7b and Llama2-13b-chat as base models. Results have been averaged across three meta-group split setups and three random seeds, with standard deviations provided.

4 Conclusion

We introduced GPO, a novel method for few-shot alignment of LLM outputs with limited preference data. Trained on a meta-train dataset with group preferences, GPO efficiently adapts to new test groups, offering superior performance over existing methods in alignment scores without needing gradient updates. It excels in alignment scores over other methods and showcases superior sample efficiency across two open-source LLMs, with limitations and future work discussed in Appendix L.

References

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv e-prints*, pages arXiv–2112, 2021.

[2] 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.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Hritik Bansal, John Dang, and Aditya Grover. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*, 2023.

[5] Kush Bhatia, Avanika Narayan, Christopher De Sa, and Christopher Ré. Tart: A plug-and-play transformer module for task-agnostic reasoning. *arXiv preprint arXiv:2306.07536*, 2023.

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[9] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

[10] RIM Dunbar, Anna Marriott, and NDC Duncan. Human conversational behavior. *Human Nature*, 8(3):231–246, 1997.

[11] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv e-prints*, pages arXiv–2306, 2023.

[12] Amelia Glaese, Nat McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[13] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Milena Lagos, Pippa Norris, Eduard Ponarin, and Bianca Puranen. *World values survey: Round seven – country-pooled datafile version 5.0.0*. 2022.

[14] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, 2022.

[15] EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning language models to user opinions. *arXiv e-prints*, pages arXiv–2305, 2023.

[16] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*, 2023.

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[20] Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *International Conference on Machine Learning*, pages 16569–16594. PMLR, 2022.

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[22] PewResearch. Writing survey questions. URL `https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/`.

[23] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, 2021.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

[26] 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.

[27] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[28] Michael Santacroce, Yadong Lu, Han Yu, Yuanzhi Li, and Yelong Shen. Efficient rlhf: Reducing the memory usage of ppo. *arXiv preprint arXiv:2309.00754*, 2023.

[29] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[30] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[31] Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021.

[32] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.

[33] Simeng Sun, Dhawal Gupta, and Mohit Iyyer. Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of rlhf. *arXiv preprint arXiv:2309.09055*, 2023.

[34] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[35] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[37] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[38] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*, 2023.

[39] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

[40] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Ethics Statement

GPO can be used to align models to preferences of diverse interest groups which can provide a more positive, useful, and inclusive experience for end users of LLM applications. We acknowledge that aligning LLMs to the preferences of demographic groups can have malicious applications. For example, making LLMs more capable of producing responses that are more tailored to specific users may be misused to convince or show members of a group how to perform unethical actions. Additionally, GPO's methodology can be used to align a model to a group's preferences even if those preferences are harmful. Biased, offensive, and harmful preferences present in the meta-train or meta-test datasets may be reflected in the outputs of GPO. Future work should investigate methods for aligning LLM outputs to group preferences without amplifying harmful outputs.

B Detailed Results and Discussion

Adapting to US demographics in OpinionQA. We conducted experiments with three distinct meta-train and meta-test splits, allocating 40%, 60%, and 80% of the 22 US demographics groups respectively as the meta-train groups G_{train}. This group split was consistent with the *In-context Finetune* baseline and GPO. For other baselines that operate on a per-group basis, we calculated the alignment score for the meta-test groups and present results averaged over three random seeds.

Our results are presented in Figure 3. Alpaca-7b base model exhibit alignment scores that are similar to the alignment score of a uniform distribution. This does not necessarily imply an absence of biases, as averaging across groups could obscure biases towards certain demographics. Prior work has found LMs may disproportionately over-represent some groups and under-represent others [29]. However, Llama2-13b-chat base model exhibits a lower alignment score as compared to the uniform distribution. This might be attributed to its fine-tuning for safety, causing the model to lean towards the least harmful option, which can be seen from the qualitative examples in Appendix K. When we incorporate group information into the LMs, we deploy various prompting strategies—QA, BIO, and PORTRAY—to convey this information (see Appendix M for examples). We report results for the strategy that yields the best alignment as *LM-steered*. Given explicit group information, *Alpaca-7b-steered* displays slightly lower relative gains as compared to *Llama2-13b-steered*. Next, when provided with few-shot group preference context samples, which serve as an implicit method of conveying group information, the LM's alignment performance significantly declines compared to the base language model's performance. We hypothesize this decline might be due to the prompting format being outside the distribution of the language model's training corpus.

For methods involving gradient updates, we maintain a consistent number of context samples across all baselines, which is also the same number of context examples used in *Few-shot prompt*. Specifically, we use 15 samples for Alpaca-7b and 20 for Llama2-13b experiments. With gradient updates, *SFT per-group* brings improvement as compared to other gradient-free steering methods. However, training a *Reward Model* to predict alignment scores from context samples, and subsequently use it to predict preference scores for query examples underperforms SFT methods. This outcome may suggest a risk of overfitting when working with a limited sample size.

GPO achieves notably higher alignment scores on this dataset compared to the baselines for both the Alpaca and Llama2 base models. GPO uses the same number of context samples for adaptation and the test groups are unseen during training. We observed performance increases when a larger number of meta-training groups were used. GPO's closest baseline, the *In-context Finetune* method ranks second, where the LMs are trained to infer from few-shot context samples. On average over the two base models and the three group split settings, GPO achieves a 7.1% increase over the *In-context Finetune*.

Figure 4 qualitatively illustrates the predicted alignment scores from different methods in response to an OpinionQA example concerning climate change concerns across six demographic groups. The first row depicts the ground truth group opinion distribution. Given just 15 context samples, GPO successfully adapts to match the opinion distributions of different groups. For instance, it increases preference for option A when adapted to the group *Hindus*, while the steered LMs do not exhibit correct distribution changes. For example, *Llama2-13b-steered* appears to be biased towards a specific option, overrepresenting it rather than accurately reflecting the distribution of the targeted group. On the contrary, in demographics with a more balanced distribution like *College graduate/some postgrad*, GPO maintains this balance more consistently. This demonstrates that GPO does not merely adapt to the overall dataset group preferences, but can align to specific groups using limited context.

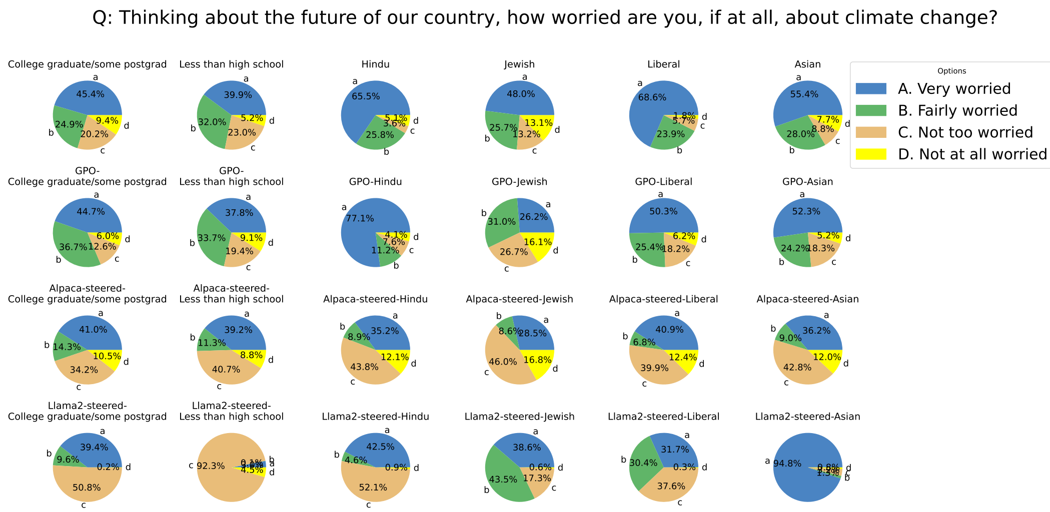


Figure 4: Qualitative comparison of GPO alignment with steered LMs, where each pie chart denotes the preference distribution of the group. Here, GPO uses Alpaca-7b's embedding.

Adapting to cross-nation groups in GlobalOpinionQA. The diverse and highly contrasting opinions across nations in GlobalOpinionQA presents a more complex landscape than the OpinionQA dataset. Upon analyzing performance, trends in the GlobalOpinionQA dataset closely followed those observed in the OpinionQA, as depicted in Figure 3. Notably, the alignment score of the Alpaca-7b base model surpasses that of the uniform distribution while Llama2-13b base model shows lower alignment. For Alpaca-7b *LM-base*, this could suggest that the base models might exhibit stronger alignment to certain specific countries and this hypothesis is supported by the increased standard deviation of the Alpaca-7b *LM-base* alignment scores, hinting at varied alignment across different countries, a phenomenon also reported in the dataset [11]. Alternatively, this could imply that the base models tend to align more with the dataset's general respondents, which naturally would exceed a uniform distribution. With gradient updates, the *SFT per-group* method here surpasses the alignment performance of steering methods, while the *Reward Model* underperforms SFT methods. The *In-context Finetune* method emerges as the third-best and second-best in terms of alignment for Alpaca-7b and Llama-13b respectively, which showcases enhanced in-context few-shot adaptation post meta-training. However, its training demands are substantially higher; it requires approximately 4.7 times more training time as compared with GPO on an NVIDIA RTX A6000 to achieve the depicted performance. Averaged across both base models and the three group split scenarios, GPO posts a 8.4% improvement over the second-best baseline.

Scalability with Increasing Context Samples. We evaluate the scalability of different methods with respect to the size of the in-context examples. Figure 5 demonstrates that for Nigeria in the GlobalOpinionQA dataset, GPO enhances alignment scores with fewer than 10 preference context samples. The performance of *Few-shot Prompt* improves with more examples but plateaus with greater variance. In comparison, *In-context Finetune* exhibits superior adaptability post meta-training than *Few-shot Prompt*, yet its alignment is still suboptimal and the number of group context samples is limited by the context window size of the LM. Both *SFT per-group* and *Reward Model* show incremental improvements with added context samples; however, their sample efficiency is modest. In contrast, GPO adeptly adapts to groups in a sample-efficient manner.

Adapting to Individual Preferences. Variations in individual opinions can manifest even within the same demographic groups [15]. Motivated by this, we assess methods to align with individual-level preferences. From the OpinionQA dataset, encompassing 15 surveys across 15 unique topics, we randomly select 100 participants from each survey, along with their responses to 30 topic-related

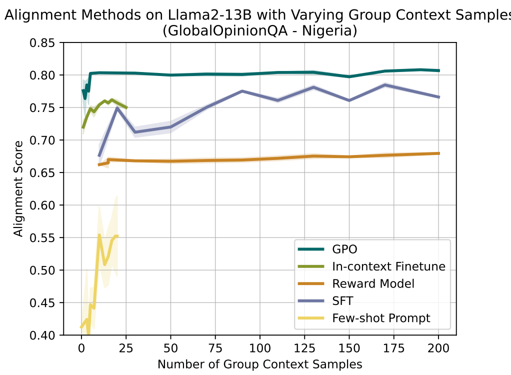


Figure 5: Alignment score of various methods based on Llama2-13B with varying group context sample size. Evaluation conducted on survey questions for Nigeria from the GlobalOpinionQA dataset. The shaded region represents the standard deviation across three different seed results.

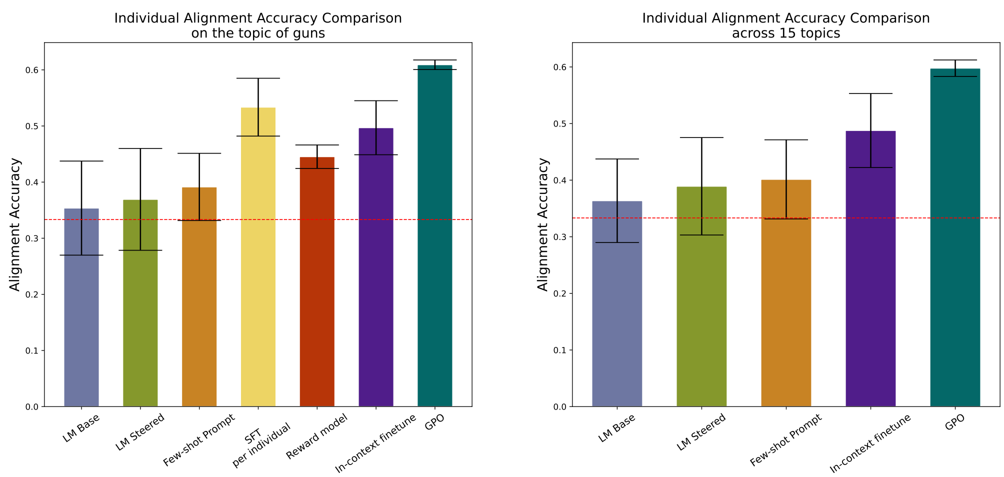


Figure 6: Individual alignment accuracy comparisons from the OpinionQA dataset. **Left:** Individual alignment on the gun topic survey. **Right:** Comprehensive comparison across all 15 topics, showcasing the performance of various methods on diverse subjects. Experiments use Alpaca-7b as the base LM. Both GPO and *In-context finetune* are meta-trained on 40% of individuals and evaluated on the remaining 60%. The horizontal red line represents the average accuracy of a random model.

questions. For each individual, 40% questions serve as context samples and 60% for queries. We use Alpaca-7b here as the base model. To steer the LM with individual information, we create individual context from combined demographic variables, such as income, religion, and age, as demonstrated in Appendix Figure 9. Since each individual only selects one option, we calculate alignment accuracy instead by treating the option with the highest predicted preference score as the predicted option. Due to computational constraints, we confined our evaluations of the SFT per-individual and reward model methods to one survey. Since both of them operate on a per-individual basis, the training needed for about a thousand individuals made broader comparisons of the two baselines impractical. In contrast, other baselines, including in-context finetune and GPO, were assessed across all 15 survey topics. Across the full breadth of the 15 topics, GPO consistently exhibited superior performance in adapting to individual preferences relative to other baselines, as depicted in Figure 6.

C Dataset Details

C.1 OpinionQA dataset

This dataset is sourced from Pew American Trends Panel [22]. This dataset's unique structural characteristics: the answer choices in the survey questions are principally ordinal [29]. For instance, options often extend across a spectrum, ranging from categories such as "A great deal," "Fair amount," "Not much," to "Not at all." Traditional divergence metrics, such as the Kullback-Leibler (KL) divergence, are ill-suited for this task, as they fail to encapsulate the ordinal relationships inherent in the answer choices. In this dataset, the ordinal answer choices are mapped to a metric space using corresponding positive integers. For example, a typical mapping in our dataset might look like $\{A : 1, B : 2, \ldots, D : 4\}$. Therefore, 1-D Wasserstein Distance metric is used. The alignment score for two opinion distributions P_1 and P_2 is consequently expressed as:

$$\mathcal{A}(P_1, P_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} \left[1 - \frac{\mathcal{WD}(P_1(q), P_2(q))}{N - 1} \right] \tag{2}$$

Here, N denotes the total number of selectable answer options, excluding the option to refuse. The term $N - 1$ functions as a normalization factor, representing the maximal possible Wasserstein distance in the given metric space. The score is bounded within the interval $[0, 1]$, with a score of 1 indicating perfect alignment between the two distributions.

We employ the dataset as encompassing 22 demographic groups within the US, as outlined in Table 1. Our analysis focuses on 500 contentious questions, characterized by frequent disagreements among the considered subgroups. These questions are the same ones used in the steerability analysis presented in the OpinionQA dataset [29].

Attribute	**Demographic Group**
CREGION	Northeast, South
EDUCATION	College graduate/some postgrad, Less than high school
GENDER	Male, Female
POLIDEOLOGY	Liberal, Conservative, Moderate
INCOME	More than $100K+, Less than $30,000
POLPARTY	Democrat, Republican
RACE	Black, White, Asian, Hispanic
RELIG	Protestant, Jewish, Hindu, Atheist, Muslim

Table 1: Demographic groups considered in our analysis from the OpinionQA dataset.

C.2 GlobalOpinionQA dataset

The survey questions in this dataset is sourced from the Pew Research Center's Global Attitudes surveys [22] and the World Values Survey [13]. These questions do not generally contain ordinal structures in the options and the ordinal scores are not presented in the datasets. Therefore, we choose to use a different metric for evaluating the alignment in this dataset.

$$\mathcal{A}(P_1, P_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} [1 - \mathcal{JD}(P_1(q), P_2(q))] \tag{3}$$

In this alternate scenario, $\mathcal{JD}$ signifies the Jensen-Shannon Distance following the paper's choice [11].

Out of the 138 countries in the original GlobalOpinionQA dataset, we selected a subsample of 14 countries for our study due to computational constraints. We extract all the survey questions that have the target countries' answers. The countries chosen (in Table 2) span several continents to ensure a broad representation in our evaluation. For instance, Nigeria and Egypt cover Africa, while India and China represent Asia. European nations are represented by countries such as Germany, France, and Spain, and the Americas include the United States, Canada, Brazil, and Argentina. Lastly, the Oceania region is represented by Australia and New Zealand.

Country
Nigeria
Egypt
India (Current national sample)
China
Japan
Germany
France
Spain
United States
Canada
Brazil
Argentina
Australia
New Zealand

Table 2: List of countries considered in our study, from GlobalOpinionQA dataset.

D Ablation on the GPO's transformer architecture

We compare GPO with a standard autoregressive transformer that employs a causal mask, akin to the transformers used in GPT-x series [24, 25, 7]. This basic architecture includes autoregressive generation with the causal mask and uses positional encoding, which we previously omitted to ensure context invariance. Using an autoregressive generation approach violates the target equivalence property since the prediction of each query point relies on previously generated ones. As depicted in Table 3, GPO's inherent inductive biases yield superior alignment performance compared to a traditional transformer. It's noteworthy that in this comparison, we still concatenate the (x, y) pairs into single tokens for the standard transformer, thus preserving the relationship between the viewpoint x and the group preference score.

	Meta train on 40% groups	Meta train on 60% groups	Meta train on 80% groups
GPO	**0.798** $\pm$ **0.007**	**0.820** $\pm$ **0.004**	**0.799** $\pm$ **0.015**
Transformer	0.780 ± 0.009	0.782 ± 0.004	0.772 ± 0.006

Table 3: Comparison of the alignment scores of GPO and a standard autoregressive transformer on alignment tasks on GlobalOpinionQA datasets with three group splits and runs are averaged over three seeds. Experiments are conducted on OpinionQA with Alpaca-7b as the base model.

E Ablation on getting embeddings from the LLM.

Given that the base LLMs we considered in our experiments were not explicitly trained for text summarizing, we examined three methods to generate the embedding x of the sentence: 1) Using the embedding of the last token as the sentence embedding. 2) Averaging over the embeddings of all tokens in the sentence. 3) Concatenating the embeddings obtained from the previous two methods. As depicted in the table 4, averaging over the token embeddings of the sentence yielded the most effective results, whereas relying solely on the last token embedding proved less adept at capturing sentence-level information.

Embedding Method	Alignment Score
Alpaca-7b last token	0.903 ± 0.014
Alpaca-7b average tokens	$\mathbf{0.946 \pm 0.007}$
Alpaca-7b last token + average	0.942 ± 0.009

Table 4: Comparison of different embedding methods using Alpaca-7b as the base model on the OpinionQA dataset, with a meta train split of 80%. Results are averaged across three seeds.

F Ablation on adding group meta-context for GPO

In the primary experiments, viewpoints x are embedded using an LLM. Notably, in our previous experiments, each x_i does not contain group meta-data about the group's identity or attributes. This ablation study explores the potential performance enhancement that could be achieved by integrating meta-data into GPO. Specifically, the context information c_g is embedded into a vector z^g_{ctx}, which is of the same dimension as x as embedded by the same LLM. We examined adding c_g from the three kinds of contextual prompts we study in M. This embedding is then concatenated with each of the (x, y, z_{ctx}) pairs, serving as the one input token for GPO. As illustrated in Table 5, incorporating context embeddings into the structure doesn't bolster GPO's performance across the three group split scenarios, instead it performs worse. We hypothesize this outcome arises because GPO, unlike LLMs, lacks comprehensive world knowledge of diverse group attributes, making it challenging to adapt to the meta-data embeddings of unfamiliar groups. Instead, GPO excels in deducing preference distributions based on the available (x, y) context sample pairs.

	Meta train on 40% groups	Meta train on 60% groups	Meta train on 80% groups
GPO	**0.920 $\pm$ 0.003**	**0.926 $\pm$ 0.013**	**0.946 $\pm$ 0.007**
GPO w/ meta-data	0.900 ± 0.003	0.916 ± 0.017	0.926 ± 0.006

Table 5: Comparison of the alignment scores of GPO with and without meta-data embeddings with three group splits and runs are averaged over three seeds. Experiments are conducted on OpinionQA with Alpaca-7b as the base model.

G Baselines Details

We compare our method against extensive baseline approaches for aligning an LLM's predicted opinion distributions with human groups:

- **Uniform Distribution:** This baseline assumes that all answer options are chosen with equal probability, indicating no preference or bias towards any specific option. For a given question $q \in Q$ with N answer choices, the distribution $P_U(q)$ is represented as: $P_U(q) = \left[\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right]$.
- ***LM Base*:** The opinion distribution, denoted by P_π, is derived from a pre-trained LM without any group-specific steering or fine-tuning. For a given question $q \in Q$, the distribution $P_\pi(q)$ generated by the model is extracted from the output probability distribution across the N available answer choices. We first extract the prediction scores for the next token from the LM, focusing on the top-K tokens. We then normalize the values to obtain $P_\pi(q)$. For a token that is missing from the top-K set, we allocate the smallest prediction score in the top-K set. We use $K = 200$ in our experiments.
- ***LM Steered*:** This baseline gauges the model's adaptability to align with a specific group $g \in G$ when informed of the group information explicitly through the prompt. We use diverse prompting strategies—QA, BIO, and PORTRAY—to convey group information, with examples in Appendix M. The opinion distribution obtained for group g under this steering is expressed as $P_\pi(q; c_g)$, where c_g denotes the context for group g.
- ***Few-shot Prompt*:** Rather than giving the model explicit group information, we input a few examples showing a group's preferences for m context questions, constrained by the LM's context window size. Here the c_g includes the context samples $\{q, r_i, y_i\}_{i=1}^m$. Using this context, the model is prompted to generate a response for a new, unseen question that aligns with the group's opinions. See Figure 8 in the Appendix for examples.
- ***SFT per group*:** The LM is fine-tuned separately for each group g using a supervised loss. Let $Q_{\text{train}} \subset Q$ denote the subset of m context questions used for training. We create training examples (q, r) by sampling q from Q_{train} and then sampling responses r with respect to the preference distribution $P_g(q)$. The loss is defined as:

$$L_{\text{SFT}} = -\mathbb{E}_{q \sim Q_{\text{train}}, r \sim P_g(q)} \log p_\psi(r|q) \tag{4}$$

  where ψ represents the LM parameters and $p_\psi(r|q)$ denotes the probability of producing the response r given the question q. This procedure fine-tunes the LM to maximize the likelihood of the sampled responses that align with the preference distribution of the specific group.

- ***Reward Model*:** We start with the architecture of the base LM and add a linear MLP head. The augmented model is trained on m context samples to predict the preference scores for the $\{x_i\}_{i=1}^m$ using a mean squared error loss. Then, the model is employed to predict the preference scores for the query questions and softmax is applied to ensure that $\sum_{i=1}^{T} \hat{y}_{g,q,i} = 1$ for each query q.
- ***In-Context Finetune*:** We investigate whether the LM can be fine-tuned, akin to GPO, to adapt to a distribution of groups using few-shot learning. This would ideally enable improved few-shot in-context adaptation for unseen groups. To this end, we partition the group set G into a meta-train set G_{train} and a meta-test set G_{test}. During training, each group in G_{train} serves as a training instance. The training questions for each group are split into context samples and query questions. For a given query question q, we supplement it with a few-shot context c_g, consisting of m questions paired with the respective ground truth preference scores. This context mirrors the *Few-shot Prompt* strategy with example shown in Appendix 8. For supervision, for each query, we sample responses r, aligned with the human preference distribution $P_g(q)$. The LM undergoes fine-tuning using a dataset formed from these context-enhanced samples. The associated loss function is:

$$L_{ICT} = -\mathbb{E}_{g\sim G_{train}, q\sim Q, r\sim P_g(q)} \log p_\psi(r|q, c_g) \tag{5}$$

H Training Settings

For all baseline fine-tuning methods, including SFT per group, reward modeling, and in-context fine-tuning that necessitate training the base LM, we employ 8-bit integer quantization and utilize a single Nvidia RTX A6000 GPU with 48GB VRAM. Our parameter search for the learning rate encompassed values {3e-4, 2e-5, 1e-4}. We settled on 1e-4 for the Alpaca baselines and 2e-5 for the Llama2-13B-chat baselines. For both SFT and in-context fine-tuning tasks, our effective batch size was 8, comprised of a batch size of 1 and 8 gradient accumulation steps. In contrast, reward model training had a batch size of 4 with the same gradient accumulation steps. All baseline methodologies were trained with LoRA (with r=12, alpha=32, and a dropout rate of 0.05) with a weight decay of 0.01, utilizing bf16 precision and the AdamW optimizer [19]. For all methods, We use the validation alignment score for early stopping.

For GPO, the transformer's feedforward dimension was set to 128, with an embedding depth of 4, 4 heads, and 6 layers. We sampled m uniformly from the range [10, 100] as context samples for every training task. We also used a learning rate of 3e-4, coupled with the Adam Optimizer [17]. More training details can be found in our codebase.

I Extending GPO Beyond Multiple-Choice Questions

The GPO framework presented in the main paper experiments can be extended beyond the multiple choice setting. GPO works for any LLM generation setting where there is some scalar which represents feedback over an LLM response. We present GPO formulations for producing group aligned LLM responses in the long-form generation setting with two common forms of sparse feedback: (1) relative (e.g. is response 1 or response 2 better) and (2) absolute (e.g. rate the response on a scale of 1-7).

Relative feedback: each context example includes 2 responses and GPO is trained with a binary classification objective for each example. During inference, the GPO module can be used to perform inference through a modified version of best-of-n sampling where n sample responses are sampled from the base LLM and each of the $\binom{n}{2}$ pairs of responses is inputted to GPO as queries. GPO's output can used to calculate a win rate for each of the n responses and the response with the highest win-rate is chosen as the aligned output response.

Absolute feedback: each context example includes 1 prompt and GPO is trained to regress the absolute feedback score. During inference, the GPO module can be used as a reward model in best-of-n sampling to produce a group aligned response.

Since GPO predicts group preference scalars, GPO can be used as a reward model to fine-tune the base LLM with PPO in settings where performing inference with an additional model is not desirable.

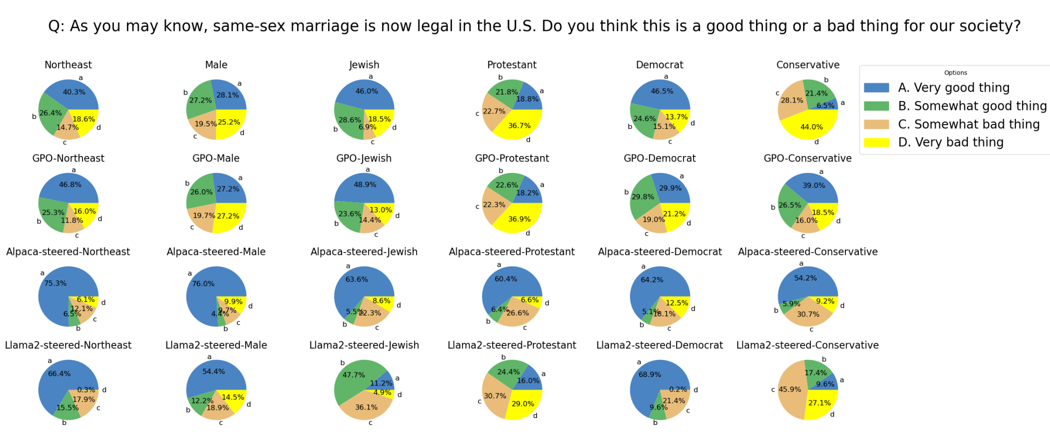
J Additional Related Work

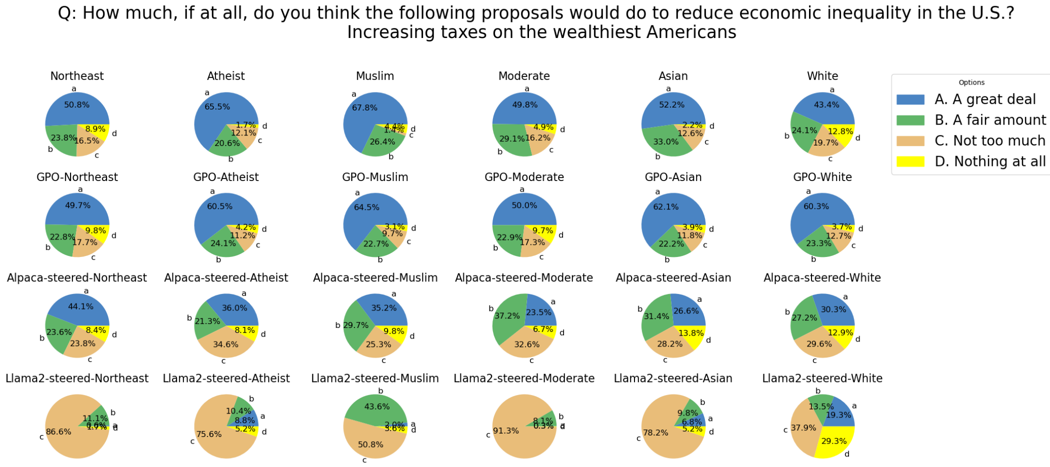
Alignment via Prompting. The conditional nature of LLMs enables them to be conditioned on specific task information or context data and alter their output distribution to respect the conditional information. Various studies have investigated this capability in adapting to groups and personas. Deshpande et al. [9] observed that prompts like *"Speak like xxx"* could elevate LLM's toxicity levels contingent upon the persona's characteristics. Jiang et al. [16] used prompts to guide GPT-3.5 to adopt certain personality traits. Beyond explicit persona or group traits, an LLM's behavior can also be influenced by presenting it with few-shot examples from its in-context learning ability [7]. For example, Hwang et al. [15] uses the previous opinions of an individual to adapt the LLM to align with the user. The advantages of this strategy include its computational efficiency, eliminating the necessity for gradient updates. However, the few-shot examples are restricted by the model's context size, and designing prompts for effective completion of tasks often requires careful prompt engineering [18, 23, 39, 27]. Additionally, when steering the LLM to be more representative of a demographics group on nuanced societal questions from survey datasets, especially on nuanced societal matters, research by Santurkar et al. [29], Durmus et al. [11] shows that this steerability can be constrained, resulting in limited or no enhancements in model alignment.

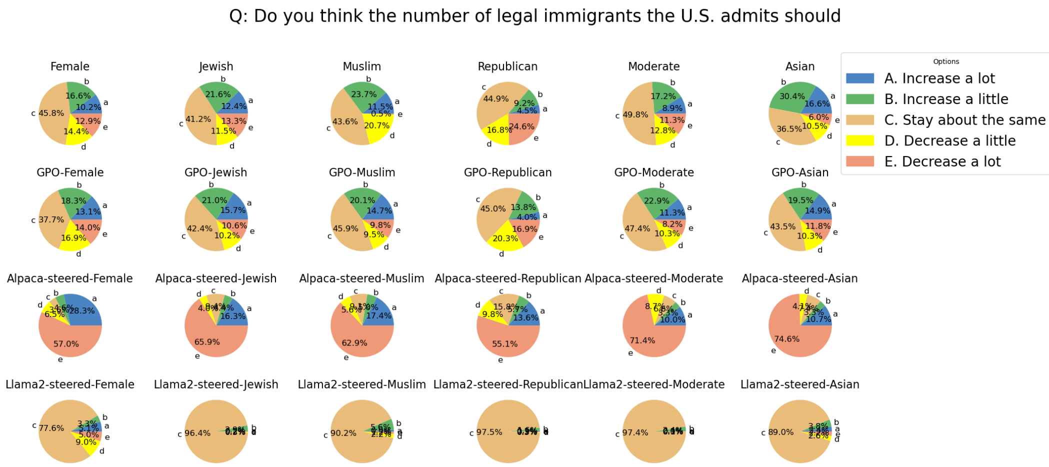
Gradient-based Alignment. Another line of alignment work involves adjusting the LM's parameters using a preference dataset through fine-tuning. Methods have been proposed to align LLMs with specific human values, such as helpfulness, harmlessness, and non-toxicity, using RLHF [12, 21, 26, 2]. This necessitates the modeling of a reward model from human-labeled preference datasets and the use of RL policies, such as PPO [30], to maximize accumulated rewards. This PPO phase poses challenges in terms of computational and memory demands, necessitating the training and storing of large-scale policy, value, reward, and reference models, as well as optimizer states and gradients in GPU memory. Moreover, this process could require complex hyperparameter tuning [33, 28]. To address this, methods such as Direct Preference Optimization [26] and Preference Ranking Optimization [32] have been proposed to directly learn from pairwise or ranking-based preference datasets without reward modeling. However, these methods typically result in specialized models for every alignment task. Adapting to multifaceted, sometimes conflicting group preferences requires fine-tuning distinct models for each subgroup.

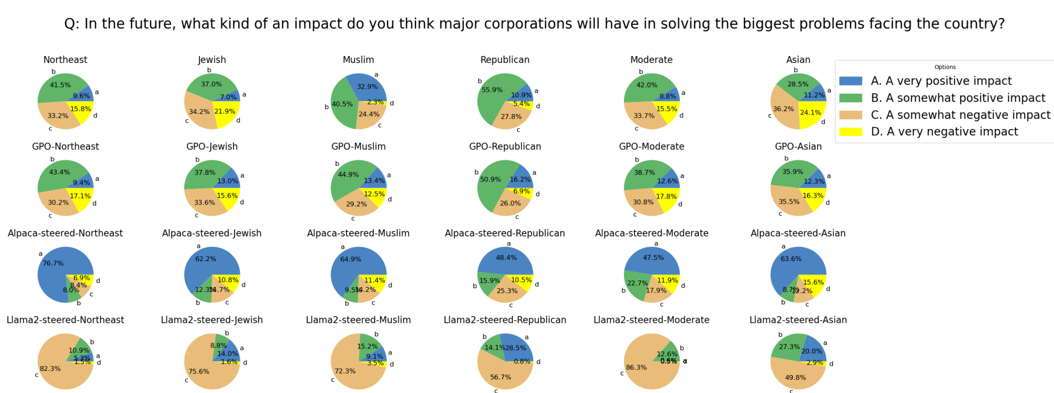
K Qualitative examples of GPO.

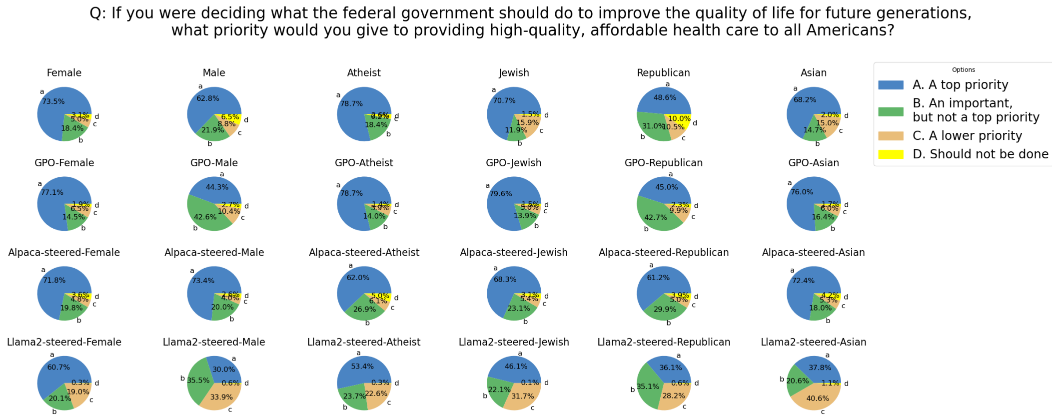
Warning: This section contains qualitative examples that may be viewed as offensive or harmful. Here we demonstrate multiple qualitative examples of GPO's predicted group preference versus the language model's steered performance. Here we used only 15 context examples for GPO and the steered LM uses group's meta-data as context.

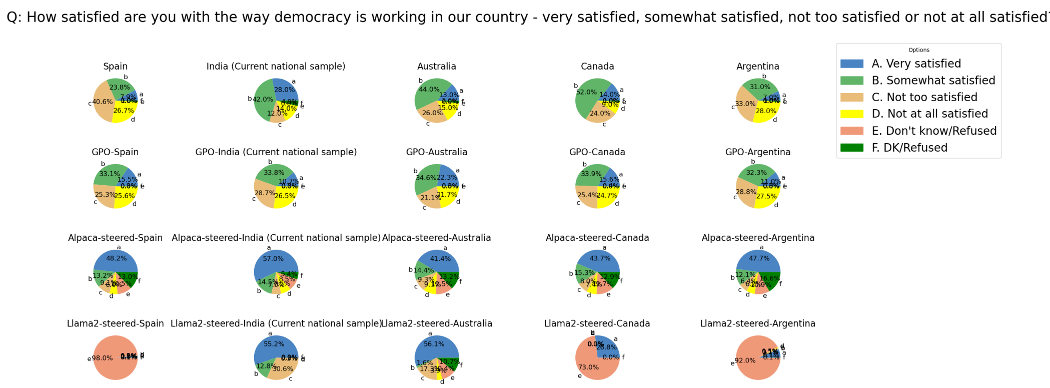



Q: How much, if at all, do you think the following proposals would do to reduce economic inequality in the U.S.?
Increasing taxes on the wealthiest Americans
Northeast
Atheist
Muslim
Moderate
Asian
White
GPO-Northeast
GPO-Atheist
GPO-Muslim
GPO-Moderate
GPO-Asian
GPO-White
Alpaca-steered-Northeast
Alpaca-steered-Atheist
Alpaca-steered-Muslim
Alpaca-steered-Moderate
Alpaca-steered-Asian
Alpaca-steered-White
Llama2-steered-Northeast
Llama2-steered-Atheist
Llama2-steered-Muslim
Llama2-steered-Moderate
Llama2-steered-Asian
Llama2-steered-White
Options
A. A great deal
B. A fair amount
C. Not too much
D. Nothing at all


Q: Do you think the number of legal immigrants the U.S. admits should
Female
Jewish
Muslim
Republican
Moderate
Asian
GPO-Female
GPO-Jewish
GPO-Muslim
GPO-Republican
GPO-Moderate
GPO-Asian
Alpaca-steered-Female
Alpaca-steered-Jewish
Alpaca-steered-Muslim
Alpaca-steered-Republican
Alpaca-steered-Moderate
Alpaca-steered-Asian
Llama2-steered-Female
Llama2-steered-Jewish
Llama2-steered-Muslim
Llama2-steered-Republican
Llama2-steered-Moderate
Llama2-steered-Asian
Options
A. Increase a lot
B. Increase a little
C. Stay about the same
D. Decrease a little
E. Decrease a lot


Q: In the future, what kind of an impact do you think major corporations will have in solving the biggest problems facing the country?
Northeast
Jewish
Muslim
Republican
Moderate
Asian
GPO-Northeast
GPO-Jewish
GPO-Muslim
GPO-Republican
GPO-Moderate
GPO-Asian
Alpaca-steered-Northeast
Alpaca-steered-Jewish
Alpaca-steered-Muslim
Alpaca-steered-Republican
Alpaca-steered-Moderate
Alpaca-steered-Asian
Llama2-steered-Northeast
Llama2-steered-Jewish
Llama2-steered-Muslim
Llama2-steered-Republican
Llama2-steered-Moderate
Llama2-steered-Asian
Options
A. A very positive impact
B. A somewhat positive impact
C. A somewhat negative impact
D. A very negative impact





L Limitations and Future Works

Opinion Datasets: We use datasets containing opinions of various demographic groups to validate GPO. Survey data is imperfect and may not be fully representative of an entire group's population. Additionally, all the datasets that we use in this work are in English. When aligning to groups, the language that is used to collect preference data and during alignment may have a significant effect on alignment metrics, especially if the inputs and outputs are in a different language than the native language of members of a group. Future work should also investigate more challenging few-shot alignment settings, such as adapting to individual creative preferences where there may be much higher variance between group preferences.

Multiple-choice Format: Like many previous works, we focus on a multiple-choice format due to the availability of existing datasets and ease of quantitative evaluations. LLMs are capable of producing much more complicated long-form responses, and it is important that alignment methods can be extended to the general long-form response setting. While the GPO framework extends more broadly to different formats of LLM generations, future work should validate the effectiveness of GPO for longer form responses and additional considerations such as group preference feedback representation and evaluation metrics needed to extend to the long-form setting.

Alignment Objectives: When aligning LLMs, multiple factors beyond group preference alignment are also very important. Aligning to group preferences may result in worse alignment for other factors including as harmlessness and helpfulness especially if the group preference data includes examples that contradicts these values. Moreover, aligning to group preferences may amplify undesirable behaviors from LLMs including biased or harmful outputs. Future work should study the impact of group alignment on other important alignment factors and methods to reduce regressions for these factors when aligning to group preferences.

M Contextual prompt examples

In this paper, we examine three types of contextual prompts, as delineated in [29]. Below, we present examples of the question-answer, biographical, and portrait-based contextual prompts designed for individuals residing in the Northeastern United States.

Question-Answer Prompt:
```
Which part of the United States do you currently live in?
Response:  Northeast
```

Biographical Prompt:
```
Below, please provide a brief description of the region in which you
currently reside within the United States, followed by answers to
several questions.
Description:  I currently reside in the Northeast.
```

Portrait-Based Prompt:
```
Answer the following question as if you currently reside in the
Northeast.
```

Figure 7: Three types of contextual prompts to provide group information.

```
Below is an instruction that describes a task, paired with an input that
provides further context.  Write a response that appropriately completes
the request.

### Instruction:
Given the answer distributions from a specific demographic group for
certain questions in a public opinion survey, answer the subsequent new
question by selecting ONE of the options, as if you are a member of this
identified demographic group:

### Input:

Question:  Question_1
A. Option_1
B. Option_2
C. Option_3
Answer Distribution:
A: 25%, B: 35%, C: 40%

...

Question:  Question_m
A. Option_1
B. Option_2
C. Option_3
Answer Distribution:
A: 35%, B: 25%, C: 40%

Based on the above list of answered questions from a demographic group,
answer the new question by selecting ONE of the options, as if you are a
member of this demographic group:

Question:  Question_m+1
A. Option_1
B. Option_2
C. Option_3

### Response:
```

Figure 8: Few-shot in-context prompt with n context questions in Alpaca prompt format.

```
Below is an instruction that describes a task, paired with an input that
provides further context.  Write a response that appropriately completes
the request.

### Instruction:

Given that you have the following demographics context:
Marital Status:  Married,
Religious attendance:  Roman Catholic,
Region:  Northeast,
Age:  65+,
Sex:  Male,
Education:  Some college or no degree,
Income:  $30,000-$50,000,
Political ideology:  Conservative,
Race:  White,
Answer the following question by picking ONE of the given options

### Input:

Would you say Germany has done a good or bad job dealing with the
coronavirus outbreak?

Options:
A. Very good
B. Somewhat good
C. Somewhat bad
D. Very bad

### Response:
```

Figure 9: A randomly selected individual contextual prompt examples in Alpaca prompt format.