A Dataset Details

A.1 OpinionQA dataset

This dataset is sourced from ew American Trends Panel (PewResearch). This dataset's unique structural characteristics: the answer choices in the survey questions are principally ordinal (Santurkar et al., 2023). 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:

$$\mathcal{A}(P_1, P_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} \left[1 - \frac{\mathcal{WD}(P_1(q), P_2(q))}{N-1} \right] \tag{4}$$

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 (Santurkar et al., 2023).

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.

A.2 GlobalOpinionQA dataset

The survey questions in this dataset is sourced from the Pew Research Center's Global Attitudes surveys (PewResearch) and the World Values Survey (Haerpfer et al., 2022). 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.

$$\mathcal{A}(P_1, P_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} [1 - \mathcal{JD}(P_1(q), P_2(q))] \tag{5}$$

In this alternate scenario, $\mathcal{JD}$ signifies the Jensen-Shannon Distance following the paper's choice (Durmus et al., 2023).

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.

B 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 (Radford et al., 2018; 2019; Brown et al., 2020). 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, stemming from the two properties, 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** ± **0.007**	**0.820** ± **0.004**	**0.799** ± **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.

C 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 of the sentence x: 1) Using the embedding of the last token as the sentence embedding. 2) Averaging over the embeddings of all tokens in the sentence x. 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.

D Ablation on adding group meta-context for GPO

In the primary experiments, viewpoints x are embedded using an LLM. Notably, 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 H. 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 ± 0.003**	**0.926 ± 0.013**	**0.946 ± 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.

E 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_m, is derived from a pre-trained LM without any group-specific steering or fine-tuning. For a given question $q \in Q$, the distribution $P_m(q)$ 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 use softmax to normalize the values to obtain $P_m(q)$. For a token that is missing from the top-K set, we allocate the smallest prediction score in the top-K set.
- ***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 H. The opinion distribution obtained for group g under this steering is expressed as $P_m(q; c_g)$, 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 $\{x_i, y_i\}_{i=1}^m$. 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:

$$L_{\text{SFT}} = -\mathbb{E}_{q \sim Q_{\text{train}}, r \sim P_g(q)} \log p_\psi(r|q) \tag{6}$$

  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:

$$L_{ICT} = -\mathbb{E}_{g \sim G_{train}, q \sim Q, r \sim P_g(q)} \log p_\psi(r|q, c_g) \tag{7}$$

F 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 (Loshchilov & Hutter, 2018).

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 (Kingma & Ba, 2015).

G 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.

H CONTEXTUAL PROMPT EXAMPLES

In this paper, we examine three types of contextual prompts, as delineated in Santurkar et al. (2023). 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.