FSPO: Few-Shot Preference Optimization of Synthetic Preference Data Elicits LLM Personalization to Real Users

Anonymous Authors[1]

Abstract

Effective personalization of LLMs is critical for a broad range of user-interfacing applications such as virtual assistants and content curation. Inspired by the strong in-context learning capabilities of LLMs, we propose few-shot preference optimization (FSPO), an algorithm for LLM personalization that reframes reward modeling as a meta-learning problem. Under FSPO, an LLM learns to quickly infer a personalized reward function for a user via a few labeled preferences. Since real-world preference data is scarce and challenging to collect at scale, we propose careful design choices to construct synthetic preference datasets for personalization, generating over 1M synthetic personalized preferences using publicly available LLMs. In particular, to successfully transfer from synthetic data to real users, we find it crucial for the data to exhibit both high diversity and coherent, self-consistent structure. We evaluate FSPO on personalized open-ended generation for up to 1,500 synthetic users across across three domains: movie reviews, pedagogical adaptation based on educational background, and general question answering. We also run a controlled human study. Overall, FSPO achieves an 87% Alpaca Eval winrate in generating responses that are personalized to synthetic users and a 72% winrate with real human users in open-ended question answering.

1. Introduction

As large language models (LLMs) increasingly interact with a diverse user base, it becomes important for models to generate responses that align with individual user preferences. People exhibit a wide range of preferences and beliefs shaped by their cultural background, personal experience, and individual values. These diverse preferences are human-annotated preference datasets; however, current preferences optimization techniques like reinforcement learning from human feedback (RLHF) largely focus on optimizing a *single* model based on preferences aggregated over the entire population. This approach may neglect minority viewpoints, embed systematic biases into the model, and ultimately lead to worse performance compared to personalized models. Can we create language models that can adaptively align with the personal preferences of each user instead of the aggregated preferences of all users?

Addressing this challenge requires a shift from modeling a singular aggregate reward function to modeling a distribution of reward functions that captures the diversity of human preferences (Sorensen et al., 2024; Jang et al., 2023). By doing so, we can enable personalization in language models, allowing them to generate a wide range of responses tailored to individual subpopulations. This approach not only enhances user satisfaction but also promotes inclusivity by acknowledging and respecting the varied perspectives that exist within any user base. Despite this problem's importance, to our knowledge LLM personalization has yet to be achieved for open-ended question answering with real users.

In this paper, we introduce few-shot preference optimization (FSPO), a novel framework designed to model diverse subpopulations in preference datasets to elicit personalization in language models for open-ended question answering. At a high level, FSPO leverages in-context learning to adapt to new subpopulations. This adaptability is crucial for practical applications, where user preferences can be dynamic and multifaceted. Inspired by past work on black-box meta-learning for language modeling (Chen et al., 2022; Min et al., 2022; Yu et al., 2024), we fine-tune the model in a meta-learning setup using preference-learning objectives such as IPO (Gheshlaghi Azar et al., 2023). To further improve personalized generation, we additionally propose user description chain-of-thought (COT), which allows the model to leverage additional inference-time compute for better reward modeling and instruction following.

Learning a model that effectively personalizes to real people requires training on a realistic, user-stratified preference dataset. One natural approach to consider is to curate such

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

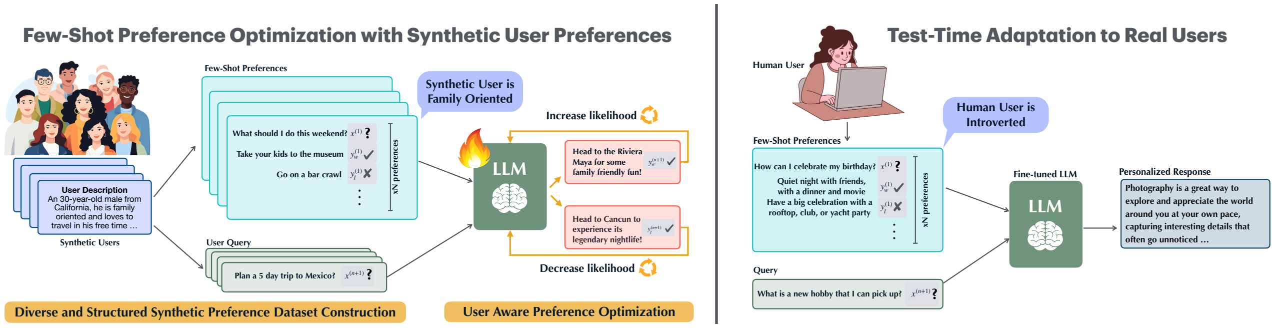

Figure 1: **Overview of FSPO.** N previously collected preferences are fed into the LLM along with the current query, allowing the LLM to personalize its response to the query using the past preferences.

data from humans, but this is difficult and time-consuming. Instead, we propose instantiating this dataset synthetically, and present careful design decisions inspired from the meta-learning literature (Hsu et al., 2019; Yin et al., 2019) to generate a dataset that is both diverse and structured.

To evaluate the efficacy of our approach, we construct a set of three semi-realistic domains to study personalization: (1) **Reviews**, studying the generation ability of models for reviews of movies, TV shows, and books that are consistent with a user's writing style, (2) **Explain Like I'm X (ELIX)**: studying the generation ability of models for responses that are consistent with a user's education level, and (3) **Roleplay**: studying the generation ability of models for responses that are consistent with a user's description, with effective transferability to a real human-study. Here we find that FSPO outperforms an unpersonalized model on average by 87%. We additionally perform a controlled human study showcasing a winrate of 72% of FSPO over unpersonalized models.

By addressing limitations of existing reward modeling techniques, our work paves the way for more inclusive and personalized LLMs. We believe that FSPO represents a significant step toward models that better serve the needs of all users, respecting the rich diversity of human preferences.

2. Related Work

Personalized learning of preferences. Prior research has explored personalization through various methods. One approach is distributional alignment, which focuses on matching model outputs to broad target distributions rather than tailoring them to individual user preferences. For example, some prior work have concentrated on aligning model-generated distributions with desired statistical properties (Siththaranjan et al., 2024; Meister et al., 2024; Melnyk et al., 2024), yet they do not explicitly optimize for individual preference adaptation. Another strategy involves explicitly modeling a distribution of rewards (Lee et al., 2024; Poddar et al., 2024). However, these methods suffer from sample inefficiency during both training and inference (Rafailov et al., 2023; Gheshlaghi Azar et al., 2023). Additionally, these approaches have limited evaluations: Lee et al. (2024) focuses solely on reward modeling, while Poddar et al. (2024) tests with a very limited number of artificial users (e.g helpfulness user and honest user). Other works have investigated personalization in multiple-choice questions, such as GPO (Zhao et al., 2024). Although effective in structured survey settings, these methods have not been validated for open-ended personalization tasks. Similarly, Shaikh et al. (2024) explores personalization via explicit human corrections, but relying on such corrections is expensive and often impractical to scale. Finally, several datasets exist for personalization, such as Prism (Kirk et al., 2024) and Persona Bench (Castricato et al., 2024). Neither of these datasets demonstrate that policies trained on these benchmarks lead to effective personalization. Unlike these prior works which study personalization based off of human values and controversial questions, we instead study more general questions that a user may ask.

Algorithms for preference learning. LLMs are typically fine-tuned via supervised next-token prediction on high-quality responses and later refined with human preference data (Casper et al., 2023; Ouyang et al., 2022). This process can use on-policy reinforcement learning methods like REINFORCE (Sutton et al., 1999) or PPO (Schulman et al., 2017), which optimize a reward model with a KL constraint. Alternatively, supervised fine-tuning may be applied to a curated subset of preferred responses (Dubois et al., 2024b) or iteratively to preferred completions as in ReST (Gulcehre et al., 2023). Other methods, such as DPO (Rafailov et al., 2023), IPO (Gheshlaghi Azar et al., 2023), and KTO (ContextualAI, 2024), learn directly from human preferences without an explicit reward model, with recent work exploring iterative preference modeling applications (Yuan et al., 2024).

Black-box meta-learning. FSPO is an instance of black-box meta-learning, which has been studied in a wide range of domains spanning image classification (Santoro et al., 2016; Mishra et al., 2018), language modeling (Chen et al., 2022; Min et al., 2022; Yu et al., 2024), and reinforcement learning (Duan et al., 2016; Wang et al., 2016). Black-box meta-learning is characterized by the processing of task contexts and queries using generic sequence operations like recurrence or self-attention, instead of specifically designed adaptation mechanisms.

3. Preliminaries and Notation

Preference fine-tuning algorithms, such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), typically involve two main stages (Ouyang et al., 2022; Ouyang et al., 2022): supervised fine-tuning (SFT) and preference optimization (DPO/RLHF). First, a pre-trained model is fine-tuned on high-quality data from the target task using SFT. This process produces a reference model, denoted as π_{ref}. The purpose of this stage is to bring the responses from a particular domain in distribution with supervised learning. To further refine π_{ref} according to human preferences, a preference dataset $\mathcal{D}_{\text{pref}} = \{(\mathbf{x}^{(i)}, \mathbf{y}_w^{(i)}, \mathbf{y}_l^{(i)})\}$ is collected. In this dataset, $\mathbf{x}^{(i)}$ represents a prompt or input context, $\mathbf{y}_w^{(i)}$ is the preferred response, and $\mathbf{y}_l^{(i)}$ is the less preferred response. These responses are typically sampled from the output distribution of π_{ref} and are labeled based on human feedback.

Most fine-tuning pipelines assume the existence of an underlying reward function $r^*(\mathbf{x}, \cdot)$ that quantifies the quality of responses. A common approach to modeling human preferences is the Bradley-Terry (BT) model (Bradley and Terry, 1952), which expresses the probability of preferring response $\mathbf{y}_1$ over $\mathbf{y}_2$, given a prompt $\mathbf{x}$, as:

$$p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \frac{e^{r^*(\mathbf{x}, \mathbf{y}_1)}}{e^{r^*(\mathbf{x}, \mathbf{y}_1)} + e^{r^*(\mathbf{x}, \mathbf{y}_2)}} \tag{1}$$

Here, $p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})$ denotes the probability that $\mathbf{y}_1$ is preferred over $\mathbf{y}_2$ given $\mathbf{x}$.

The objective of preference fine-tuning is to optimize the policy π_θ to maximize the expected reward r^*. However, directly optimizing r^* is often impractical due to model limitations or noise in reward estimation. Therefore, a reward model r_ϕ is trained to approximate r^*. To prevent the fine-tuned policy π_θ from deviating excessively from the reference model π_{ref}, a Kullback-Leibler (KL) divergence constraint is imposed. This leads to the following fine-tuning objective:

$$\max_{\pi} \; \mathbb{E}[r^*(x, y)] - \beta \, D_{\text{KL}}(\pi \parallel \pi_{\text{ref}}) \tag{2}$$

In this equation, the regularization term weighted by β controls how much π_θ diverges from π_{ref}, based on the reverse KL divergence constraint. This constraint ensures that the updated policy remains close to the reference model while improving according to the reward function.

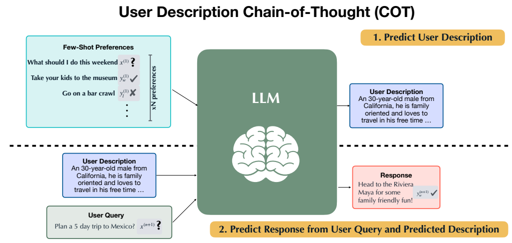

Figure 2: **User description chain-of-thought (COT).** Prediction is a two-stage process: first predicting a (synthetic) user description from the few-shot preferences and next predicting the response.

Reward model training. To fine-tune the large language model (LLM) policy $\pi_\theta(\mathbf{y} \mid \mathbf{x})$, the Bradley-Terry framework allows for either explicitly learning a reward model $r_\phi(\mathbf{x}, \mathbf{y})$ or directly optimizing preferences. Explicit reward models are trained using the following classification objective:

$$\max_{\phi} \; \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\log \sigma \left(r_\phi(\mathbf{x}, \mathbf{y}_w) - r_\phi(\mathbf{x}, \mathbf{y}_l)\right)\right] \tag{3}$$

where σ is the logistic function, used to map the difference in rewards to a probability. Alternatively, contrastive learning objectives such as Direct Preference Optimization (Rafailov et al., 2023) and Implicit Preference Optimization (Gheshlaghi Azar et al., 2023) utilize the policy's log-likelihood $\log \pi_\theta(\mathbf{y} \mid \mathbf{x})$ as an implicit reward:

$$r_\theta(\mathbf{x}, \mathbf{y}) = \beta \log \left(\pi_\theta(\mathbf{y} \mid \mathbf{x}) / \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\right) \tag{4}$$

This approach leverages the policy's log probabilities to represent rewards, thereby simplifying the reward learning process.

4. The Few-Shot Preference Optimization (FSPO) Framework

Personalization as a meta-learning problem. Generally, for fine-tuning a model with RLHF a preference dataset of the form: $\mathcal{D}_{\text{pref}} = \{(\mathbf{x}^{(i)}, \mathbf{y}_w^{(i)}, \mathbf{y}_l^{(i)})\}$ is collected, where x is a prompt, y_w is a preferred response, and y_l is a dispreferred response. Here, preferences from different users are aggregated to learn the preferences over a population. However, through this aggregation, individual user preferences are marginalized, leading to the model losing personalized values or beliefs due to population-based preference learning and RLHF algorithms such as DPO as seen in prior work (Siththaranjan et al., 2024).

How can we incorporate user information when learning from preference datasets? In this work, we have a weak requirement to collect scorer-ids $\mathbf{S}^{(i)}$ of each user for differentiating users that have labeled preferences in our dataset:

$\mathcal{D}_{\text{pref}} = \{(\mathbf{x}^{(i)}, \mathbf{y}_w^{(i)}, \mathbf{y}_l^{(i)}, \mathbf{S}^{(i)})\}$. Now consider each user as a task instance, where the objective is to learn an effective reward function for that user using the user's set of preferences. This can be naturally instantiated as a black-box meta-learning objective, where meta-learning is done over users (also referred to as a task in meta-learning). Meta-learning should enable rapid personalization, i.e. adaptability to new users with just a few preferences.

More formally, consider that each unique user $\mathcal{S}^{(i)}$'s reward function is characterized by a set of preferences with prompt and responses (x, y_1, y_2), and preference label c (indicating if $y_1 \succ y_2$ or $y_1 \prec y_2$). Given a distribution over users $\mathcal{S} = P(\mathcal{S}^{(i)})$, a meta-learning objective can be derived to minimize its expected loss with respect to θ as:

$$\min_\theta \mathbb{E}_{\mathcal{S}^{(i)} \sim \mathcal{S}} \Big[\mathbb{E}_{(x, y_1, y_2, c) \sim \mathcal{D}_i, \{(x, y_1, y_2, c)\}_1^N \sim D_i} \left[\mathcal{L}^\theta_{\textit{pref}} \left(x, y_1, y_2, c | \{(x, y_1, y_2, c)\}_1^N \right) \right] \Big] \tag{5}$$

where D_i is a distribution over preference tuples (x, y_1, y_2, c) for each user $S^{(i)}$, and $\mathcal{L}^\theta_{\textit{pref}}$ is a preference learning objective such as DPO (Rafailov et al., 2023) or IPO (Gheshlaghi Azar et al., 2023):

$$\begin{aligned} \mathcal{L}^\theta_{\textit{pref}} &= || h^{y_w, y_l}_{\pi_\theta} - (2\beta)^{-1} ||_2^2, \\ h^{y_w, y_l}_{\pi_\theta} &= \log \frac{\pi_\theta(y_w|x)}{\pi_{\textit{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\textit{ref}}(y_l|x)} \end{aligned} \tag{6}$$

where y_w and y_l are the preferred and dispreferred responses (respectively) according to the responses y_1, y_2 and class label c in the preference dataset.

Following black-box meta-learning approaches, FSPO receives as input a sequence of preferences $D_i^{fewshot} \sim D_i$ from a User $S^{(i)}$. This is followed by an unlabeled, held-out preference $(x, y_1, y_2) \sim \mathcal{D}_i \backslash \mathcal{D}_i^{\textit{fewshot}}$ for which it outputs its prediction c. To make preferences compatible with a pretrained language model, a few-shot prompt is constructed, comprising of preferences from a user and the held-out query as seen in Figure 1. This construction has an added benefit of leveraging a pretrained language model's capabilities for few-shot conditioning (Brown et al., 2020), which can enable some amount of steerage/personalization. This prediction c is implicitly learned by a preference optimization algorithm such as DPO (Rafailov et al., 2023), which parameterizes the reward model as $\beta \frac{\log \pi_\theta(y|x)}{\log \pi_{\textit{ref}}(y|x)}$. This parameterization enables us to leverage the advantages of preference optimization algorithms such as eliminating policy learning instabilities and computational burden of on-policy sampling, learning an effective model with a simple classification objective.

User description chain-of-thought (COT). If provided with a description of the user (potentially synthetically generated), FSPO can be converted to a two-step prediction problem as seen in Figure 2. In the first step, conditioned on user few-shot preferences, the user description is generated, then conditioned on the prompt, few-shot preferences, and generated user description, a response can then be generated. This prediction of the user description is an interpretable summarization of the fewshot preferences and a better representation to condition on for response generation. Similar to the rationale generated in Zhang et al. (2024) for verifiers, the COT prediction can be viewed as using additional inference-compute for better reward modeling. Additionally, this formulation leverages the instruction following ability of LLMs (Ouyang et al., 2022) for response generation.

User representation through preference labels. From an information-theoretic perspective, the few-shot binary preferences can be seen as a N-bit representation of the user, representing up to 2^N different personas or reward functions. There are several ways to represent users: surveys, chat histories, or other forms of interaction that reveal hidden preferences. We restrict our study to such a N-bit user representation, as such a constrained representation can improve the performance when transferring reward models learned on synthetic personalities to real users. We defer the study of less constrained user representations to future work.

We summarize FSPO in Algorithm 1. Next, we will discuss domains to study FSPO.

Algorithm 1 Overview of few-Shot preference optimization (FSPO).

```
Input: For each unique user S^(i), a dataset of preferences D := (x, y_1, y_2, c)_i, and optionally user description y_S^(i) for COT, ∀i
Output: Learned policy π_θ
while not done do
    Sample training user S^(i) (or minibatch)
    Sample a subset of preferences from the user D_i^fewshot ~ D_i
    Sample held-out preference examples D_i^heldout ~ D_i \ D_i^fewshot
    if COT then
        Use Equation (5) and Equation (6) to predict the loss on the user description y_S^(i)
    end if
    Conditioning on D_i^fewshot (optionally y_S^(i)), use Equation (5) and Equation (6) to predict the loss on the held-out preference example D_i^heldout
    Update learner parameters θ, using gradient of loss on D_i^heldout
end while
Return π_θ =0
```

5. Domains to Study Personalization

To study personalization with FSPO we construct a benchmark across 3 domains ranging from generating personalized movie reviews (**Reviews**), generating personalized responses based off a user's education background (**ELIX**), and personalizing for general question answering (**Roleplay**). We open-source preference datasets and evaluation protocols from each of these tasks for future work looking to study personalization (sample in supplementary).

Reviews. The Reviews task is inspired by the IMDB dataset (Maas et al., 2011), containing reviews for movies. We curate a list of popular media such as movies, TV shows, anime, and books for a language model to review. We consider two independent axes of variation for users: sentiment (positive and negative) and conciseness (concise and verbose). Here being able to pick up the user is crucial as the users from the same axes (e.g positive and negative) would have opposite preferences, making this *difficult* to learn with any population based RLHF method. We also study the steerability of the model considering the axes of verbosity and sentiment in tandem (e.g positive + verbose).

ELIX. The Explain Like I'm X (ELIX) task is inspired by the subreddit "Explain Like I'm 5" where users answer questions at a very basic level appropriate for a 5 year old. Here we study the ability of the model to personalize a pedagogical explanation to a user's education background. We construct two variants of the task. The first variant is **ELIX-easy** where users are one of 5 education levels (elementary school, middle school, high school, college, expert) and the goal of the task is to explain a question such as "How are beaches formed?" to a user of that education background. The second, more realistic variant is **ELIX-hard**, which consists of question answering at a high school to university level. Here, users may have different levels of expertise in different domains. For example, a PhD student in computer science may have a very different educational background from an undergraduate studying studying biology, allowing for preferences from diverse users (550 users).

Roleplay. The Roleplay task tackles general question answering across a wide set of users, following PRISM (Kirk et al., 2024) and PERSONA Bench (Castricato et al., 2024) to study personalization representative of the broad human population. We start by identifying three demographic traits (age, geographic location, and gender) that humans differ in that can lead to personalization. For each trait combination, we generate 30 personas, leading to 1,500 total personas. To more accurately model the distribution of questions, we split our questions into two categories: global and specific. Global questions are general where anyone may ask it, but specific questions revolve around a trait, for example an elderly person asking about retirement or a female asking about breast cancer screening.

One crucial detail for each task is the construction of a preference dataset that spans multiple users. But how should one construct such a dataset that is realistic and effective?

6. Sim2Real: Synthetic Preference Data Transfers to Real Users

Collecting personalized data at scale presents significant challenges, primarily due to the high cost and inherent unreliability of human annotation. Curating a diverse set of users to capture the full spectrum of real-world variability further complicates the process, often limiting the scope and representativeness of the data. Synthetically generating data using a language model (Li et al., 2024; Bai et al., 2022) is a promising alternative, since it can both reduce costly human data generation and annotation and streamline the data curation process. Can we generate diverse user preference data using language models in a way that transfers to real people?

We draw inspiration from simulation-to-real transfer in non-language domains like robotics (Makoviychuk et al., 2021) and self-driving cars (Yang et al., 2023), where the idea of domain randomization (Tobin et al., 2018) has been particularly useful in enabling transfer to real environments. Domain randomization enables efficient adaptation to novel test scenarios by training models in numerous simulated environments with varied, randomized properties.

But why is this relevant to personalization? As mentioned previously, each user can be viewed as a different "environment" to simulate as each user has a unique reward function that is represented by their preferences. To ensure models trained on synthetic data generalize to real human users, we employ domain randomization to simulate a diverse set of synthetic preferences. However, diversity alone isn't sufficient to learn a personalized LM. As studied in prior work (Hsu et al., 2019; Yin et al., 2019), it is crucial that the task distribution in meta-learning exhibits sufficient structure to rule out learning shortcuts that do not generalize. But how can we elicit both **diversity** and **structure** in our preference datasets?

Encouraging diversity. Diversity of data is crucial to learning a reward function that generalizes across prompts. Each domain has a slightly different generation setup as described in Section 5, but there are some general design decisions that are shared across all tasks to ensure diversity.

One source of diversity is in the questions used in the preferences. We use a variety of strategies to procure questions for the three tasks. For question selection for ELIX, we first sourced questions from human writers and then synthetically augmented the set of questions by prompting GPT-4o (OpenAI et al., 2024) with subsets of these human-generated questions. This allows us to scalably augment the human

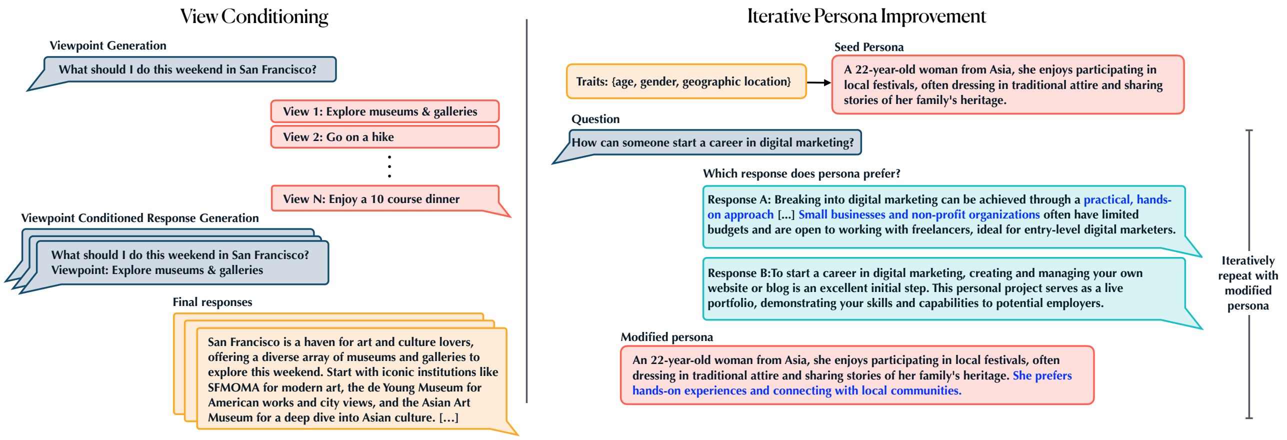

Figure 3: **Overview of domain randomization techniques.** View-conditioning (left) decomposes a given question into multiple viewpoints, allowing for diverse response generation. Iterative persona generation (right) allows for better structure by removing underspecification of the persona by iteratively refining a persona if it is insufficient to make a preference prediction.

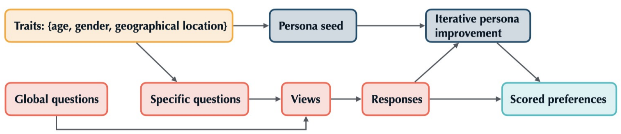

Figure 4: **Flowchart of Roleplay dataset generation:** Starting from a set of traits, a seed persona is constructed and a set of specific questions about that trait. Then responses are constructed with View-Conditioning. The seed personas are then iteratively refined to not be underspecified. Finally, the refined persona is used to score consistent preferences.

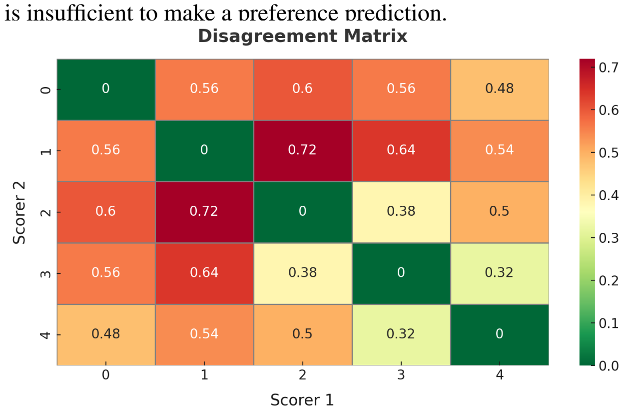

Figure 5: **Disagreement matrix across 5 users in Roleplay.** Here we plot the disagreement of preferences for 5 users. There is a mix of users with high and low disagreement.

question dataset, while preserving the stylistic choices and beliefs of human writers. For the reviews dataset, we compiled a list of popular media from sites such as Goodreads, IMDb, and MyAnimeList. For the Roleplay dataset, we prompted GPT-4o to generate questions all users would ask (global) or questions only people with a specific trait would ask (specific). This allows us to have questions that are more consistent with the distribution of questions people may ask.

Additionally, having a diversity of responses is crucial for not only training the model on many viewpoints but also reward labeling, allowing for greater support over the set of possible responses for a question. To achieve diverse responses, we employ two strategies: Persona Steering (Cheng et al., 2023) and view conditioning. For ELIX and Reviews, we use persona steering by prompting the model with a question and asking it to generate an answer for a randomly selected persona. For Roleplay, the user description was often underspecified so responses generated with persona steering were similar. Therefore, we considered a multi-turn approach to generating a response. First, we asked the model to generate different viewpoints that may be possible for a question. Then, conditioned on each viewpoint independently, we prompted the model with the question and the viewpoint and asked it to answer the question adhering to the viewpoint presented. For example, if you consider the question, "How can I learn to cook a delicious meal?", one viewpoint here could be "watching a youtube video", better suited for a younger, more tech savvy individual, whereas viewpoints such as "using a recipe book" or "taking a cooking class" may be better for an older population or those who would have the time or money to spend on a cooking class. This allowed for more diversity in the responses and resulting preferences.

Finally, we sampled responses from an ensemble of models with a high temperature, including those larger than the base model we fine-tuned such as Llama 3.3 70b (Grattafiori et al., 2024) and Gemma 2 27b (Team et al., 2024), allowing for better instruction following abilities of the fine-tuned model, than the Llama 3.2 3B we fine-tune.

Encouraging task structure. Meta-learning leverages a shared latent structure across tasks to adapt to a new task quickly. The structure can be considered as similar feature representations, function families, or transition dynamics that the meta-learning algorithm can discover and leverage. For a preference dataset, this structure can be represented as the distribution of preferences across different users and is controlled by the scoring function and the distribution of responses.

One thing we controlled to enable better structure is the scoring function used to generate synthetic preferences. Firstly, we wanted to ensure consistent preference labeling. We use AI Feedback (Bai et al., 2022) to construct this, using relative pairwise feedback for preference labels, akin to AlpacaEval (Dubois et al., 2024b), as an alternative to absolute rubric based scoring, which we found to be noisy and inaccurate. The preference label along with being conditioned on the prompt, response, and general guidance on scoring, is now also conditioned on the scoring user description and additional scoring guidelines for user-aware preference labeling. Additionally, due to context length constraints, many responses for our preference dataset are shorter than the instruct model that we fine-tune from. Therefore, we prompt the model to ignore this bias. Furthermore, we provide each preference example to the model twice, flipping the order of the responses, and keeping filtering out responses that are not robust to order bias for both training and evaluation (win rates).

Additionally, as mentioned above, in some cases, such as with the Roleplay dataset, the user description is underspecified, leading to challenges in labeling consistent preferences. For example, if a user description does not have information about dietary preferences, inconsistency may arise for labeling preferences about that topic. For instance, in one preference pair, vegan cake recipes may be preferred but in another, steakhouses are preferred for date night. To fix this, we take an iterative process to constructing user descriptions. Firstly, we start with a seed set of user descriptions generated from the trait attributes. After generating questions and responses based on these seed descriptions, we take a set of question and response pairs. For each pair, we iteratively refine the user description by prompting a model like GPT-4o to either label the preference pair or if the user description is insufficient, to randomly choose a preference and append information to the description so a future scorer would make the same decision. Finally, we utilize the updated user description to relabel preferences for the set of questions and responses allocated to that user with the labeling scheme above. This fix for underspecification also helps the COT prediction as predicting an underspecified user persona, can lead to ambiguous generated descriptions.

Finally, we desire structured relationships between users. To ensure this, we analyzed the disagreement (average difference of preference labels) of user's preferences across prompts to understand where users agreed and disagreed, and regenerated data if this disagreement was too high across users. By having users with some overlap, meta-learning algorithms can learn how to transfer knowledge effectively from one user to another. A sample disagreement plot for a subset of users in the Roleplay task can be found in Figure 5. We outline our full dataset generation process in Figure 4 in the Roleplay Task, starting from just a simple set of demographic traits.

Strategy	Mean Similarity (↓)	Median Similarity (↓)
Temp. = 0.3	0.96	0.97
Temp. = 1.0	0.94	0.95
Persona Steering (ours)	0.81	0.82
View Steering (ours)	0.78	0.78
Ensemble + View (ours)	**0.71**	**0.73**

Table 1: Comparison of diversity-inducing strategies as evaluated under ALOE (Wu et al., 2024).

Evaluating diversity and structure. We evaluate our design decisions with the following vignettes. For diversity, we measure semantic similarity using the dense score from the BGE-M3 model, following ALOE (Wu et al., 2024), on 100 randomly sampled prompts and 10 responses per prompt in the Roleplay task. As seen in Table 1, our proposed steering and ensembling mechanisms result in the base Llama 3.2 3B Instruct model exhibiting significantly reduced mean similarity. For structure, we estimate binary Shannon entropy of the preference label before and after iterative refinement. We condition on the persona and an unlabeled preference tuple (prompt and responses) and sample a preference label with a fixed temperature of 1.0 on 100 randomly sampled prompts from the Roleplay task with 100 pairs of personas and 10 samples per prompt. We use GPT-4o as the scoring model. Iterative persona refinement causes the entropy to drop from 0.64 nats to 0.13 nats, validating the efficacy of this approach in inducing better persona-prompt-response consistency.

7. Experimental Evaluation

Baselines. We compare FSPO against five baselines: (1) a base model generating user-agnostic responses, (2) few-shot prompting with a base model, following Meister et al. (2024), (3) few-shot supervised fine-tuning (Pref-FT) based off the maximum likelihood objective from GPO (Zhao et al., 2024), (4) prompting with an oracle user description following Persona Steering (Cheng et al., 2023), and (5) Rewards-in-Context (Yang et al., 2024b). Specifically, for (1) we use a standard instruct model that is prompted solely with the query, resulting in unconditioned responses. For (2) and (3), the base instruct model is provided with the same few-shot personalization examples as in FSPO, but

Method	Trained	Interpolated
Llama 3.2 3B Instruct	50.0	50.0
4-shot Prompted	66.6	61.9
4-shot Pref-FT	66.5	66.1
4-shot FSPO (Ours)	**78.4**	**71.3**
8-shot Prompted	69.1	59.1
8-shot Pref-FT	65.6	70.7
8-shot FSPO (Ours)	**80.4**	**73.6**
8-shot FSPO + COT (Ours)	**92.3**	**84.6**

Table 2: Review Winrates

Method	ELIX-easy	ELIX-hard
Llama 3.2 3B Instruct	50.0	50.0
Few-shot Prompted	92.4	81.4
Few-shot Pref-FT	91.2	82.9
FSPO (Ours)	**97.8**	**91.8**

Table 4: GPT-4o Winrates on ELIX

Method	Winrate (%)
Llama 3.2 3B Instruct	50.0
IPO	72.4
Few-shot Prompting	63.2
Few-shot Pref-FT (GPO (Zhao et al., 2024))	62.8
RIC (Yang et al., 2024b)	53.3
FSPO (Ours, DPO)	**81.3**
FSPO (Ours, IPO)	**82.6**
FSPO + COT (Ours)	**90.3**
Oracle (prompt w/ g.t. persona)	90.9

Table 3: GPT-4o Winrates on Roleplay (1500 users)

Baseline Method	Winrate (%)
FSPO vs Base	**71.2**
FSPO vs SFT	**72.3**

Table 5: Roleplay: Human Eval Winrates

(2) zero-shot predicts the preferred response and (3) is optimized with SFT to increase the likelihood on the preferred response. In (4), the base model is prompted with the oracle, ground truth user description, representing an upper bound on FSPO's performance.

Synthetic winrates. We first generate automated win rates using the modified AlpacaEval procedure from Section 6. In the ELIX task in Table 4, we study two levels of difficulty (easy, hard), where we find a consistent improvement of FSPO over baselines. Next, in Table 2 for the Review task, on both Trained and Interpolated Users, FSPO allows for better performance on held-out questions. Finally, in Table 3, we study Roleplay, scaling to 1500 real users, seeing a win rate of 82.6% on both held-out users and questions. Also, COT closes the gap to the oracle response, effectively recovering the ground-truth user description. In Appendix A.2, sample generations from FSPO show effective personalization to the oracle user description. Given this result, can we personalize to real people?

Preliminary human study. We evaluate our model trained on the Roleplay task by personalizing responses for *real human participants*. We build a data collection app (Figure 7), interacting with a user in two stages. First, we ask participants to label preference pairs, used as the few-shot examples in FSPO. Then, for held out questions, we show a user a set of two responses: (1) a response from FSPO personalized based on their preferences and (2) a baseline response. Prolific is used to recruit a diverse set of study participants, evenly split across genders and continents, corresponding to the traits used to construct user descriptions. Question and response order is randomized to remove confounding factors. We evaluate with 25 users and 11 questions. As seen in Figure 5, we find that FSPO has a 71% win rate over the Base model and a 72% win rate over an SFT model trained on diverse viewpoints from the preference dataset.

8. Discussion and Conclusion

We introduce FSPO, a novel framework for eliciting personalization in language models for open-ended question answering that models a distribution of reward functions to capture diverse human preferences. Our approach leverages meta-learning for rapid adaptation to each user, addressing limitations of conventional reward modeling techniques that learn from aggregated preferences. Through rigorous evaluation in 3 domains, we demonstrate that FSPO's generations are consistent with user context and preferred by real human users. Our findings also underscore the importance of diversity and structure in synthetic personalized preference datasets to bridge the Sim2Real gap. Overall, FSPO is a step towards developing more inclusive, user-centric language models.

8.1. Limitations, Potential Risks, and Societal Impact

Key limitations include ethical concerns like reinforcing user biases, requiring future work on safeguards. The preliminary human study, though a first for personalized open-ended QA, had controlled elements and needs further ablation. Computational constraints restricted us to smaller models with limited context; investigating larger, more capable models is an open question, especially for Chain-of-Thought processes.

References

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Biyik, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification.

L. Castricato, N. Lile, R. Rafailov, J.-P. Fränken, and C. Finn. Persona: A reproducible testbed for pluralistic alignment, 2024. URL https://arxiv.org/abs/2407.17387.

Y. Chen, R. Zhong, S. Zha, G. Karypis, and H. He. Meta-learning via language model in-context tuning, 2022. URL https://arxiv.org/abs/2110.07814.

M. Cheng, E. Durmus, and D. Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.84. URL https://aclanthology.org/2023.acl-long.84/.

ContextualAI. Human-centered loss functions (halos), 2024. URL https://github.com/ContextualAI/HALOs.

Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2024a. URL https://arxiv.org/abs/2404.04475.

Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2024b.

M. Gheshlaghi Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, and R. Munos. A General Theoretical Paradigm to Understand Learning from Human Preferences. *arXiv e-prints*, art. arXiv:2310.12036, Oct. 2023. doi: 10.48550/arXiv.2310.12036.

Goodreads. Goodreads: Book reviews, recommendations, and discussion, 2025. URL https://www.goodreads.com/. Accessed: 2025-02-15.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca,

J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024.

C. Gulcehre, T. L. Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, C. Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

K. Hsu, S. Levine, and C. Finn. Unsupervised learning via meta-learning. In *International Conference on Learning Representations*, 2019.

IMDb. Imdb: Ratings, reviews, and where to watch the best movies & tv shows, 2025. URL https://www.imdb.com/. Accessed: 2025-02-15.

J. Jang, S. Kim, B. Y. Lin, Y. Wang, J. Hessel, L. Zettlemoyer, H. Hajishirzi, Y. Choi, and P. Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023. URL https://arxiv.org/abs/2310.11564.

H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, B. Vidgen, and S. A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. URL https://arxiv.org/abs/2404.16019.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Y. Lee, J. Williams, H. Marklund, A. Sharma, E. Mitchell, A. Singh, and C. Finn. Test-time alignment via hypothesis reweighting, 2024. URL https://arxiv.org/abs/2412.08812.

H. Li, Q. Dong, Z. Tang, C. Wang, X. Zhang, H. Huang, S. Huang, X. Huang, Z. Huang, D. Zhang, Y. Gu, X. Cheng, X. Wang, S.-Q. Chen, L. Dong, W. Lu, Z. Sui, B. Wang, W. Lam, and F. Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models, 2024. URL https://arxiv.org/abs/2402.13064.

S. Li, F. Xue, C. Baranwal, Y. Li, and Y. You. Sequence parallelism: Long sequence training from system perspective, 2022. URL https://arxiv.org/abs/2105.13120.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. URL https://arxiv.org/abs/2108.10470.

N. Meister, C. Guestrin, and T. Hashimoto. Benchmarking distributional alignment of large language models, 2024. URL https://arxiv.org/abs/2411.05403.

I. Melnyk, Y. Mroueh, B. Belgodere, M. Rigotti, A. Nitsure, M. Yurochkin, K. Greenewald, J. Navratil, and J. Ross. Distributional preference alignment of llms via optimal transport, 2024. URL https://arxiv.org/abs/2406.05882.

S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi. Metaicl: Learning to learn in context, 2022. URL https://arxiv.org/abs/2110.15943.

N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner, 2018. URL https://arxiv.org/abs/1707.03141.

MyAnimeList. Myanimelist: Track, discover, and discuss anime & manga, 2025. URL https://myanimelist.net/. Accessed: 2025-02-15.

OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva,

J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *arXiv e-prints*, art. arXiv:2203.02155, Mar. 2022. doi: 10.48550/arXiv.2203.02155.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

S. Poddar, Y. Wan, H. Ivison, A. Gupta, and N. Jaques. Personalizing reinforcement learning from human feedback with variational preference learning, 2024. URL https://arxiv.org/abs/2408.10075.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. One-shot learning with memory-augmented neural networks, 2016. URL https://arxiv.org/abs/1605.06065.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. *arXiv e-prints*, art. arXiv:1707.06347, July 2017. doi: 10.48550/arXiv.1707.06347.

O. Shaikh, M. Lam, J. Hejna, Y. Shao, M. Bernstein, and D. Yang. Show, don't tell: Aligning language models with demonstrated feedback, 2024. URL https://arxiv.org/abs/2406.00888.

A. Siththaranjan, C. Laidlaw, and D. Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf, 2024. URL https://arxiv.org/abs/2312.08358.

T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

J. Tobin, L. Biewald, R. Duan, M. Andrychowicz, A. Handa, V. Kumar, B. McGrew, J. Schneider, P. Welinder, W. Zaremba, and P. Abbeel. Domain randomization and generative models for robotic grasping, 2018. URL https://arxiv.org/abs/1710.06425.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

S. Wu, M. Fung, C. Qian, J. Kim, D. Hakkani-Tur, and H. Ji. Aligning llms with individual preferences via interaction. *arXiv preprint arXiv:2410.03642*, 2024.

A. Yang, J. Yang, A. Ibrahim, X. Xie, B. Tang, G. Sizov, J. Reizenstein, J. Park, and J. Huang. Context parallelism for scalable million-token inference, 2024a. URL https://arxiv.org/abs/2411.01783.

R. Yang, X. Pan, F. Luo, S. Qiu, H. Zhong, D. Yu, and J. Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *International Conference on Machine Learning*, 2024b.

Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun. Unisim: A neural closed-loop sensor simulator, 2023. URL https://arxiv.org/abs/2308.01898.

M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024. URL https://arxiv.org/abs/2309.12284.

W. Yuan, R. Yuanzhe Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston. Self-Rewarding Language Models. *arXiv e-prints*, art. arXiv:2401.10020, Jan. 2024. doi: 10.48550/arXiv.2401.10020.

L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2024. URL https://arxiv.org/abs/2408.15240.

S. Zhao, J. Dang, and A. Grover. Group preference optimization: Few-shot alignment of large language models, 2024. URL https://arxiv.org/abs/2310.11523.

L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

A. Appendix

A.1. Limitations, Potential Risks, and Societal Impact

There are several limitations and potential risks. One limitation pertains to the ethical and fairness considerations of personalization. While FSPO improves inclusivity by modeling diverse preferences, the risk of reinforcing user biases (echo chambers) or inadvertently amplifying harmful viewpoints requires careful scrutiny. Future work should explore mechanisms to balance personalization with ethical safeguards, ensuring that models remain aligned with fairness principles while respecting user individuality. Additionally, our human study was preliminary with control over the questions that a user may ask, format normalization where formatting details such as markdown are removed, and view normalization comparing the same number of viewpoints for both FSPO and the baselines. However, to the best of our knowledge, we are the first approach to perform such a human study for personalization to open-ended question answering. Future work should do further ablations with human evaluation for personalization. Additionally, due to compute constraints, we work with models in the parameter range of 3B (specifically Llama 3.2 Instruct 3B) with a limited context window of 128K, and without context optimization such as sequence parallelism (Li et al., 2022; Yang et al., 2024a), further limiting the effective context window. It is an open question on how fine-tuning base models with better long-context and reasoning capabilities would help with FSPO for personalization, such as the 2M context window of Gemini Flash Thinking models, especially in the case of COT.

A.1.1. Additional Ablations

Percentage of Preference Data	Winrate (%)
10%	70.1
25%	69.5
50%	78.3
100%	82.6 (reported)

Table 6: Ablation study showing the effect of varying the percentage of preference data per user for FSPO without COT on roleplay task winrates with held-out synthetic users.

Number of Few-Shot Examples	Winrate (%)
1	65.7
2	69.3
4	72.1
8	82.6 (reported)

Table 7: Ablation study showing the impact of varying the number of few-shot examples per user on roleplay task performance.

A.2. Sample Personalized Responses

We provide sample responses from FSPO in Figure 6 across the 3 tasks that were studied (ELIX, Reviews, and Roleplay). We additionally include the oracle scoring description for each response, demonstrating FSPO's ability to personalize. Additionally, for the Review task, we show responses for interpolated users which were not found in the preference dataset, showing the ability of FSPO to generalize and be controllable.

A.2.1. Roleplay CoT

```
QUESTION:
What are some good volunteer opportunities for me?

TRUE PERSONA:
A 53-year-old woman living in North America, an avid traveler and writer who
    shares her adventures through a popular travel blog. She loves spontaneous
    adventures and values long-term financial security. She is at a point in her
    career where opportunities for advancement might be time-sensitive. She
```

```
    enjoys movies that challenge perspectives and spark discussions. She
    prioritizes well-being over being fully informed. She is interested in the
    van life movement. She appreciates creativity and variety in meal options.
    She enjoys engaging in creative and tactile activities and prefers
    traditional learning methods. She values compassionate and responsible
    choices. She enjoys exploring new and dynamic experiences. She is concerned
    about fast food. She values mental well-being and mindfulness. She enjoys
    attending events and conferences to meet new people. She appreciates
    practicality and health. She enjoys shows with dark humor and suspenseful
    storytelling.

COT GENERATED PERSONA:
Based on the examples provided, the user appears to value a more holistic and
    experiential approach to learning and personal growth. They seem to
    prioritize human connection, nature, and personal development over technology
     and material wealth.

This user is likely someone who is interested in personal stories, emotions, and
    experiences, as evident from their appreciation for Sylvia Plath's poetry and
     the suggestion to weave storytelling into public speaking. They may also be
    someone who values authenticity, simplicity, and community, as seen in their
    suggestion of house swapping and volunteering as alternative vacation options
    .

In terms of lifestyle, this user likely prefers a balanced and flexible approach
    to life, allowing for a mix of work, personal growth, and leisure activities.
     They may also be someone who values environmental sustainability and social
    responsibility, as hinted at by their consideration of eco-friendly home
    improvements and sustainable financial planning. Overall, this user seems to
    be a thoughtful, empathetic, and nature-loving individual who prioritizes
    human connections, experiences, and personal growth over material possessions
     and technological advancements.

FINAL RESPONSE:
Based on your affinity for human connections, personal growth, and environmental
    sustainability, I'd like to suggest the following volunteer opportunities
    that might resonate with you:

**Community Garden or Urban Farm**: Many organizations in your area offer
    community gardens or urban farms where you can help cultivate and maintain a
    shared space, promoting sustainability and food security. This allows you to
    connect with nature and contribute to your community's well-being.

**Nature Reserves or Conservation Efforts**: Consider volunteering with local
    wildlife conservation organizations, such as bird watching tours, tree
    planting initiatives, or habitat restoration projects. This will enable you
    to work outdoors, connect with others who share your passion for nature, and
    contribute to preserving the environment.

**Community Centers or Environmental Education Programs**: Many community centers
     offer various programs, such as gardening workshops, environmental education
     classes, or sustainable living workshops. By volunteering at these centers,
    you can share your knowledge, learn from others, and make a positive impact
    on your community.
```

Name	Value
Learning Rate (SFT/Pref-FT)	$1e^{-5}, 1e^{-6}, \mathbf{1e^{-7}}$
Learning Rate (IPO)	$1e^{-5}, \mathbf{1e^{-6}}, 1e^{-7}$
Beta (IPO)	$0.1, 0.05, 0.01, \mathbf{0.005}, 0.001$
Number of Shots	$4, \mathbf{8}$
Model Name	Llama 3.2 3B Instruct (Grattafiori et al., 2024)

Table 8: Sweep over hyperparameters for FSPO, recommended hyperparameters in bold.

```
**Animal Shelters or Wildlife Rehabilitation Centers**: If you're an animal lover
    , consider volunteering at a local animal shelter or wildlife rehabilitation
    center. You'll have the opportunity to care for animals, interact with like-
    minded individuals, and contribute to animal welfare and conservation efforts
    .

**Neighborhood Events or Community Festivals**: Look for opportunities to
    volunteer at local events, such as festivals, farmers' markets, or street
    fairs. This allows you to engage with your community, promote sustainability,
     and build connections with neighbors through shared passion for
    environmental causes.

These opportunities will not only nurture your spirit of volunteering but also
    help you foster meaningful connections with nature, others, and your
    community while nurturing a sense of personal growth and fulfillment.
```

A.3. Additional Details on Dataset Construction and Evaluation

Dataset Details We include a representative subset of the roleplay task and a jupyter notebook demonstrating shot construction for training and evaluation in the supplementary material (due to size restrictions). We will release the full datasets for each task in the final release of the paper.

Evaluation Prompt for Synthetic Preferences We used GPT-4o as a Judge using a modified variant of the Alpaca Eval (Dubois et al., 2024b;a) Prompt to be aware of a user description when scoring preference examples.

Here is the system prompt:

```
You are a highly efficient assistant, who evaluates and selects the best large
    language model (LLMs) based on the quality of their responses to a given
    instruction. This process will be used to create a leaderboard reflecting the
     most accurate and human-preferred answers.
```

Here is the user prompt:

```
You are tasked with evaluating the outputs of multiple large language models to
    determine which model produces the best response from a human perspective.

## Instructions

You will receive:
1. A **User Instruction**: This is the query or task provided to the models.
2. **Model Outputs**: Unordered responses from different models, each identified
    by a unique model identifier.
```

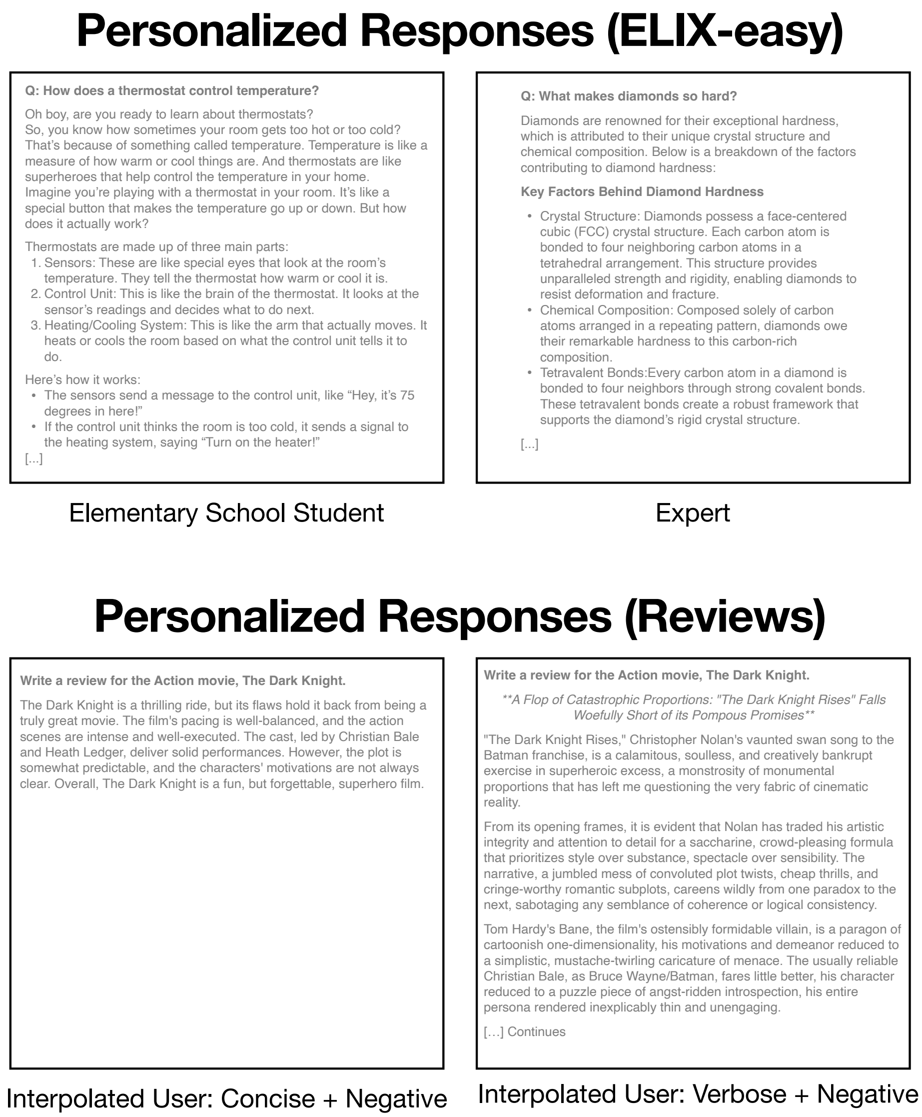

Figure 6: **Sample Personalized Response for ELIX (top) and Reviews (bottom).**

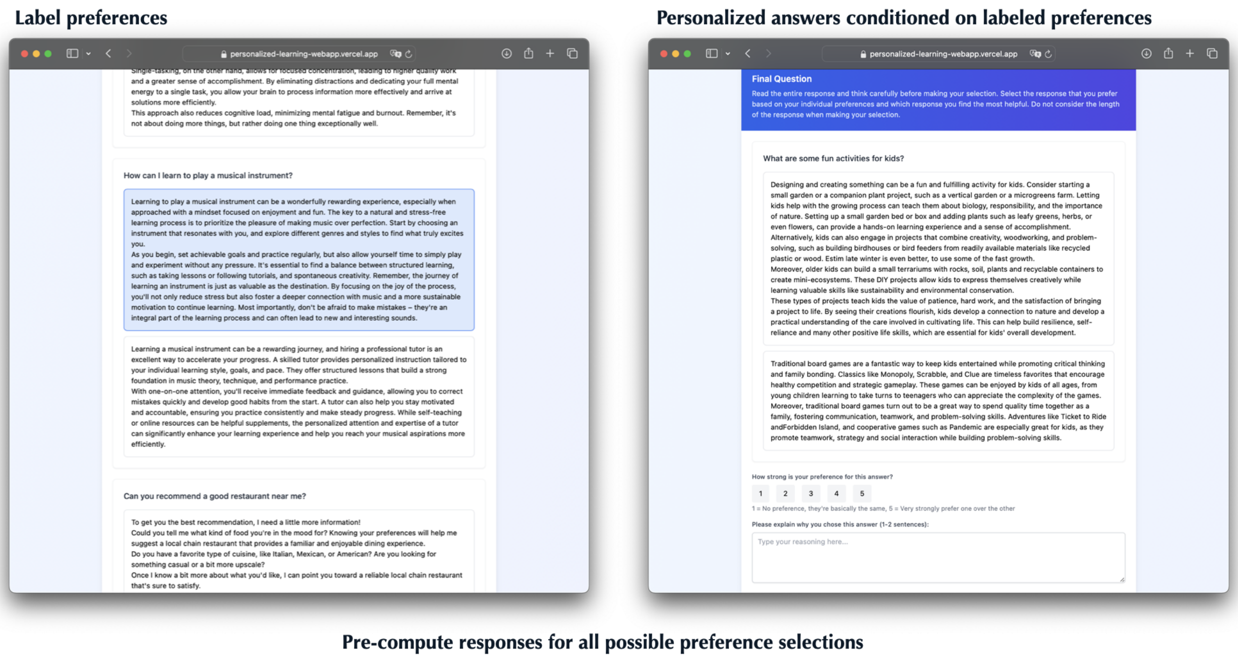

Figure 7: **An overview of the Human Study Interface.** First, users label a set of preferences. Then, a set of personalized answers are provided, conditioned on label preferences.

```
3. A **User Description**: This describes the user's preferences or additional
   context to guide your evaluation.

Your task is to:
1. Evaluate the outputs based on quality and relevance to the user's instruction
   and description.
2. Select the best output that meets the user's needs.

## Input Format

### User Instruction
{QUESTION}

### Model Outputs
- Model "m": {RESPONSE_A}
- Model "M": {RESPONSE_B}

### User Description
{USER_DESCRIPTION}

## Task

From the provided outputs, determine which model produces the best response.
   Output only the model identifier of the best response (either ‘m‘ or ‘M‘)
   with no additional text, quotes, spaces, or new lines.

## Best Model Identifier
```

Additional Human Study Details As shown in Alpaca Eval 2.0 (Dubois et al., 2024a), several biases can affect the evaluation of language models such as length, format, and more. For this reason, we took action to normalize both FSPO and baselines in 3 different categories. First, length is an evaluation bias. For this reason, we computed the average length of responses from FSPO and prompted the base model during evaluation to keep its responses around the average length in words (≈ 250 words). For the SFT baseline, we found that this was consistent with FSPO since it was fine-tuned on the same preference dataset. Additionally, due to context length restrictions and the instruction following abilities of smaller open-source LLMs, we decided to have formatting be consistent as paragraphs rather than markdown for the Roleplay task. Thus, we similarly prompted the Base model with this behavior. Finally, a differing number of views can also skew the evaluation, as a large proportion of users seem to prefer direct answers. Additionally, if more views are presented, a user may prefer just one of the many views provided, skewing evaluation. Thus, we ensure that when two responses are compared, they have the same number of views. In future, work, it would be interesting to consider how to relax some of the design decisions needed for the human study. We additionally provide screenshots of the human study interface in Figure 7.

Below is the full text of instructions given to the participants:
"This is a study about personalization. You will be asked to read a set of 20 questions (9 on the first page, 11 on the second page). For each question, there are two responses. Please select the response that you prefer. Make this selection based on your individual preferences and which response you find the most helpful. Read the entire response and think carefully before making your selection."

We utilize the demographic information that Prolific provides for each user such as their age group, continent and gender to chose questions but do not store that information about the user. We collect no identifying information about the user and will not make any of the individual preferences from a user public. We pay each user a fair wage subject to the current region that we reside in. We received consent from the people whose data we are using and curating as the very first question in our survey. The demographic and geographic characteristics of the annotator population is exactly the same as Prolific. We do no filtering of this at all.

A.4. Training Details and Hyperparameters for FSPO and baselines

Similar to DPO (Rafailov et al., 2023) and IPO (Gheshlaghi Azar et al., 2023), we trained FSPO in a two stage manner. The first stage is Fewshot Pref-FT, increasing the likelihood of the preferred response. The second stage is Fewshot IPO, initialized from the checkpoint of Fewshot Pref-FT. One epoch of the dataset was performed for each stage. For the IPO baseline, we followed a similar procedure. Additional hyperparameters can be found in Table 8.

A.5. Additional Details of Setup for Reproducability

We used both code, models, and data as scientific artifacts. In particular, for code, we built off of the codebase from Rafailov et al. (2023), with an Apache 2.0 license. We additionally adapted our evaluation script from Alpaca EVAL, including the prompt, and other criterion for evaluation and normalization. We have reported the implementation details for synthetic evaluation in Section 6 and human study evaluation in Section A.3.

For models, we used a combination of open-source and closed-source models. The models that we used for sampling data are the Llama family of models (Grattafiori et al., 2024) (Llama 3.2 3b, Llama 3.1 8b, Llama 3.3 70b) with the llama license (3.1, 3.2, 3.3), the Qwen family of models (Qwen et al., 2025) (Qwen 2.5 3b, Qwen 2.5 32b, Qwen 2.5 72b) with the qwen license, the Gemma 2 family of models (Team et al., 2024) (Gemma 2 2b, Gemma 2 9b, and Gemma 2 27b) with the gemma license, and the OpenAI (OpenAI et al., 2024) family of models (GPT4o, GPT4o-mini) with the OpenAI API License (based off of the MIT License). We used SGLang (Zheng et al., 2024) and VLLM (Kwon et al., 2023) for model inference. For training, we used 1 node of A100 GPUs (8 GPUs) for 8 hours for each experiment with FSDP. Cumulatively, we used approximately 4000 hours of GPU hours for ablations over dataset, architecture design and other details.

With respect to the dataset, for questions for the review dataset, we sourced media names from IMDb (IMDb, 2025), Goodreads (Goodreads, 2025), and MyAnimeList (MyAnimeList, 2025). We define the domains in more detail in section 5. Seed questions for ELIX were human generated, sourced from Prolific. The dataset is entirely in English, with some artifacts of Chinese from the Qwen model family, which will be filtered out for the final release of the dataset. None of this data has identifying information about individual people or offensive content as the dataset was sourced from instruction and safety-tuned models, with each step of the dataset having a manual check of the inputs and outputs. In terms of statistics of the dataset, the review dataset has 130K train/dev examples and 32.4K test examples, the ELIX-easy dataset has 235K train/dev examples and 26.1K test examples, the ELIX-hard dataset has 267K train/dev examples and 267K test examples, and the roleplay dataset has 362K train/dev examples and 58.2K test examples, with a total of 1.378 million examples. For our statistics, we reported the average winrate % for each method on both synthetic and human evals, following prior work in alignment like AlpacaFarm (Dubois et al., 2024b).

Each of the artifacts above was consistent with its intended use and the code, models, and datasets should be usable outside of research contexts.