PERSONA: A Reproducible Testbed for Pluralistic Alignment

Anonymous EMNLP submission

Abstract

The rapid advancement and adoption of language models (LMs) has highlighted critical challenges in aligning these models with the diverse values and preferences of global users. Existing reinforcement learning from human feedback (RLHF) approaches often fail to capture the plurality of user opinions, instead reinforcing majority viewpoints and marginalizing minority perspectives. To address this, we introduce PERSONA, a comprehensive and reproducible test bed designed to evaluate and improve pluralistic alignment in language models. Our approach utilizes synthetic personas, crafted through a combination of US census data and procedural generation, to simulate a wide array of user profiles with diverse demographic and idiosyncratic attributes. We present a detailed methodology for constructing a representative demographic of 1,586 personas, each enriched with individualistic personality traits and core values. Leveraging this synthetic demographic, we generate a large-scale preference dataset containing 3,868 prompts and 317,200 pairs of diverse feedback. This dataset enables the evaluation of language models' ability to align with both group-level and individual preferences across various controversial and value-laden topics. Our contributions include a systematic evaluation of current LM capabilities in role-playing diverse users, verified through human judges, and the establishment of a benchmark for pluralistic alignment approaches. Our work aims to facilitate the development of more inclusive and representative language models, paving the way for future research in global pluralistic alignment. The full dataset is available here https://sites.google.com/view/pluralistic.

1 Introduction

While reinforcement learning from human feedback (RLHF) approaches have been widely successful in creating helpful language model assistants (Ouyang et al., 2022; Gemini Team, 2024; Meta, 2024), these algorithmic methods inherently instill opinions and values within the model based on the preferences expressed by the feedback providers. Recent works (Santurkar et al., 2023a; Lee et al., 2023) have shown that widely used models do not in fact reflect the full diversity of demographic preferences—including on important topics—such as political biases (Rettenberger et al., 2024; Bang et al., 2024). These effects stem from both the opinions inherent within the user feedback data, but also the alignment algorithms used to train these models. Currently used practical methods do not take into account the plurality of users and difference of opinion, but instead work under the framework of a "representative" user, which may contribute to reinforcing majority opinions.

Several recent studies have attempted to address this issue by developing algorithms that are specifically designed to account for the distributional nature of user values (Zhao et al., 2023; Chakraborty et al., 2024; Siththaranjan et al., 2024; Ramesh et al., 2024). These approaches aim to align language models with the diverse preferences and opinions of different user groups, rather than focusing on a single "representative" user. However, significant challenges remain in achieving true pluralistic alignment (Sorensen et al., 2024). Here, recent work has suggested it is not possible to simultaneously satisfy all group preferences with a single model (Chakraborty et al., 2024), which may put into question the entire RLHF formulation. Going beyond distributional or group-level preferences, there is additional significant idiosyncratic variability in individual user values. In fact, these idiosyncratic values can be an even bigger driver of preferences than group-level attributes (Hwang et al., 2023). When properly aligned to individuals, generative models present opportunities to create uniquely bespoke interfaces, experiences and applications on a per user basis, which has recently driven significant research efforts into personalized alignment approaches (Jang et al., 2023; Li et al., 2024; Sun et al., 2024). Moreover, there have been a number of developments focused on active learning (Ji et al., 2024; Mehta et al., 2023; Muldrew et al., 2024; Zhang et al., 2024) and preference elicitation (Li et al., 2023a; Piriyakulkij et al., 2023; Andukuri et al., 2024b), which aim to teach models to effectively learn about users from interactions. **However, one major challenge for the development and deployment of such approaches is evaluation**.

Despite the significant amount of prior works and the practical importance of these problems, current test environments are still quite limited due to the challenging nature of not only collecting diverse and personalized preferences but evaluating the resulting models under those same users. Prior works (Santurkar et al., 2023b; Zhao et al., 2023; Durmus et al., 2023; Hwang et al., 2023) have established opinion polls and population surveys as benchmark. However, these usually consist of multi-choice questions and do not reflect the actual use case of LMs. Moreover, accurately predicting user choices is not necessarily correlated to the LM's ability to generate responses that align with them (Rafailov et al., 2024). In addition such polls usually only cover group-level characteristics of the surveyed population and rarely contain detailed information about specific users, limiting their usefulness for personalization applications. One major recent development is the PRISM dataset (Kirk et al., 2024), which collects preferences on actual LM-generated content from a wide arrange of global respondents on diverse and potentially controversial topics, with significant disagreement. While this effort provides good coverage for the problems discussed before, evaluation remains challenging as data is collected from real human respondents and thus algorithms and models cannot be evaluated in the same setting.

In this work we seek to address this evaluation issue through synthetic personas (Xu et al., 2024; Joshi et al., 2024; Chen et al., 2024): We model personas with realistic user profiles including detailed demographic information and varied idiosyncratic individual background, which we use to set-up role-playing LMs. Following demographic surveys, user marketing profiles and prior work we create a broad representative demographic of 1,586 personas, which we use to generate diverse feedback on a number of value-laden, diverse, and controversial topics sampled from (Kirk et al., 2024). Overall, we make the following **contributions**: First we systematically evaluate current LM capability to role-play as diverse users and verify our results with real human subjects study. We then create a benchmark of **1,586** synthetic personas as well as a large scale preference dataset with **3,868** prompts and **317,200** pairs of diverse feedback as provided by individual personas split into several datasets. Our data and evaluation framework can be used as (1) a test-bed, (2) a development environment, a (3) reproducible evaluation of pluralistic alignment approaches, (4) as personalization of LMs, and (5) for preference elicitation.

2 Related Work

Challenges in Pluralistic Alignment. While LMs are trained on data authored by billions of internet users, this involvement is passive, and pre-training datasets over-represent certain demographics (Wang et al., 2023), which can marginalize minority communities (Blodgett et al., 2020; Hershcovich et al., 2022). Moreover, while the RLHF process is paramount on instilling values within an LM it relies on even smaller pools of labellers (Sorensen et al., 2024). This can manifest in misalignment between LM outputs and the views of diverse demographics including on major political and demographical divides (Santurkar et al., 2023a; Durmus et al., 2023; Liu et al., 2024). Moreover, (Chakraborty et al., 2024) theoretically show that a single model cannot simultaneously align with diverse groups holding conflicting opinions, calling into question the main objective of RLHF tuning (Sorensen et al., 2024). Various approaches have been proposed to address these challenges, such as learning multiple reward models (Chakraborty et al., 2024; Chidambaram et al., 2024), latent variable models (Siththaranjan et al., 2024; Chidambaram et al., 2024), preference elicitation (Andukuri et al., 2024a; Li et al., 2023a), and few-shot alignment (Zhao et al., 2023; Shaikh et al., 2024). However, despite these advancements, pluralistic alignment remains a challenging problem.

Evaluation of Pluralistic Alignment. Pluralistic alignment approaches necessitates assessing how well methods actually align LMs with the range of human opinions captured in datasets. Datasets like OpinionQA (Santurkar et al., 2023a), GlobalOpinionQA (Durmus et al., 2023), and opinion polls (Hwang et al., 2023) have been widely

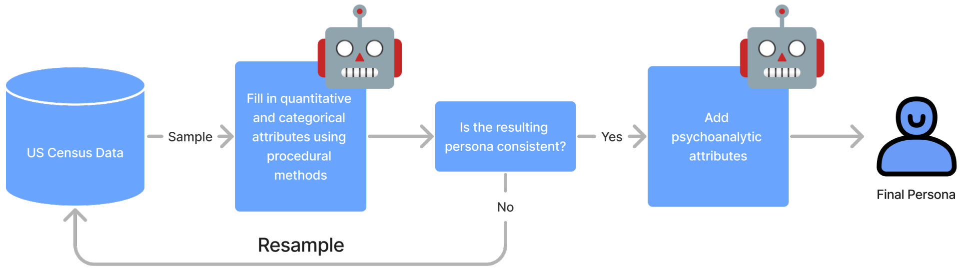

Figure 1: Procedure for generating personas. The above is a flow graph outlining the generation of a single persona. An exact example for this generation process can be found in the appendix. First, we sample a subset of US census data and query a language model to see if the resulting persona is self consistent. If it isn't, we resample. Next, we use procedural methods to fill in missing components of the census data. The list of procedural methods can be found in the appendix. Finally, we use a language model to fill in open ended psychoanalytic attributes.

used, but they only consist of multiple-choice questions and do not reflect realistic use cases of LMs. Other works have also used small-scale synthetic experiments or simple bimodal datasets, such as HH-RLHF (Bai et al., 2022), which is not representative of real world distributional views. The PRISM dataset (Kirk et al., 2024) makes progress in this direction by collecting a diverse set of open-ended conversations from a wide global population. However, it relies on human participants to provide feedback to LMs, which prevents scalable evaluation algorithms and models under the same distribution.

Role-Playing Language Agents. Recent works have shown that LMs can emulate diverse personas and traits by leveraging prompts (Li et al., 2023b; Fränken et al., 2023; Chen et al., 2024; Xu et al., 2024), inherent knowledge (Shao et al., 2023; Lu et al., 2024), and finetuning (Park et al., 2023; Fränken et al., 2024). Carefully designed role-playing scenarios with such agents could provide the rich, controllable test-bed needed to evaluate alignment approaches without human participants.

3 PERSONA: A Testbed for Pluralistic Alignment

In this section, we outline the construction of our demographic of personas and the subsequent preference data generation process.

3.1 Creating a Demographic of Personas

Our full persona-generation pipeline is shown in Figure 1. Within the taxonomy of Chen et al. (2024), our synthetic personas have a **demographic** and **individual** component. To construct demographic personas that accurately reflect the challenges of pluralistic alignment in a realistic setting, we construct a set of personas with demographics closely following the US population. This is challenging since standard US census data provides aggregate information across attributes but limited intersectional data and no personal characteristics. In contrast, the Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) files contain survey results from real people, making them more suitable for our purpose. Our dataset construction consists of several parts: (1) sampling from the PUMS files, (2) enriching each profile with additional statistically accurate psychodemographic data, (3) using language models to further enrich a small subset of fields, and (4) resolving inconsistencies (or pruning) with GPT-4. (United States Census Bureau, 2024)

We directly sample a subset of attributes from the PUMS files that cannot easily be self-inconsistent, such as someone under 18 making hundreds of thousands of dollars a year. Based on the selected characteristics, we procedurally create a demographic user profile and query GPT-4 to further filter out inconsistent ones, removing approximately 8.5% of configurations. Moreover, we used the probabilities of the Big Five personality characteristics (neuroticism, openness, conscientiousness, agreeableness, and extraversion) from the Big Five Inventory-2 (BFI-2) developed by (Soto and John, 2017) to procedurally generate five factor model personality profiles while additional core values, quirks, and mannerisms were sampled from a hand-curated set (see Appendix). Prior literature from marketing and business emphasizes the importance of psychoanalytic attributes on personal decision-making, so we further include such characteristics in our persona construction during the second generation stage (Mijač et al., 2018)

We noticed that procedurally generating idiosyncratic parts of the personas proved challenging, due

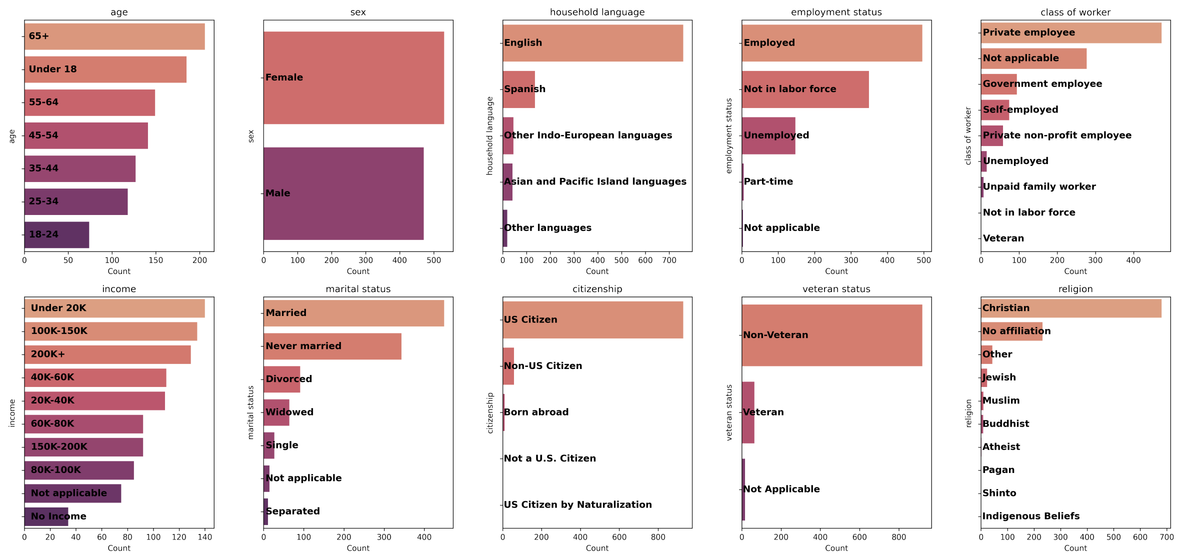

Figure 2: Histograms of group statistics of our demographic of synthetic personas.

to intersectionality effects and the open-ended nature of the problem. In our approach we broke these attributes into a number of high level categories such as "Lifestyle", "Personality", etc.. (the full list with all categories is included in A). We further selected a number of categories per persona in order to guarantee diverse coverage end and prompted GPT-4 with these to create the final open-ended persona profile. For an example of complete profiles, consult the Appendix C.

The distributional statistics of our final demographic of synthetic personas and their comparisons to the overall US census are presented in Fig. 2.

3.2 Preference Dataset Construction

Prior preference datasets (Dubois et al., 2023; Cui et al., 2023) do not have any group or individual-level information. Therefore, in order to empirically study the issues of pluralistic alignment raised earlier, we also construct a wide dataset of preferences based on the population of synthetic personas described in the previous section. We will outline our dataset curation process here.

Prompts Curation. We found the PRISM dataset (Kirk et al., 2024) to contain a diverse set of questions on a multitude of topics, including interpersonal, political, and opinionated issues that can elicit a range of preferences based on the feedback provider's background. To ensure the quality and relevance of the prompts, we performed several post-processing steps. First, we removed any instruction without a question mark and any instruction under five words in length. We then further prompted GPT-4 as a zero-shot classifier to assess whether a question is controversial or not and removed prompts which would not induce diverse opinions. This resulted in a final set of 3868 of the 8011 in the original dataset kept in our final version. The distribution of the discussion topics that are covered in our datasets is shown in Fig. 3. In order to be able to evaluate generalization we split the dataset in 3000 train prompts and 868 held-out prompts which uniformly cover the distribution of topics.

Preference Dataset Curation. While classical RLHF pipelines (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022) sample multiple answers from the reference model and asking users to rank those, this procedure is not directly applicable to our setting for several reasons. First, we base all our data generation on synthetic role-playing models, and the quality and instruction-following capabilities of the role-playing model significantly affect the fidelity of answers and feedback. However, all strong openly-available models have already undergone significant RLHF-tuning. As discussed in our introduction and related works, frontier models may have limited diversity in their responses and not fully represent the plurality of views in a demographic. Therefore, to construct a diverse set of preferences, we followed a different approach: We first randomly sample a prompt x_i and a persona p_i in an independent manner. Unlike the PRISM dataset this makes the user profiles independent from the conversational topics. This is a

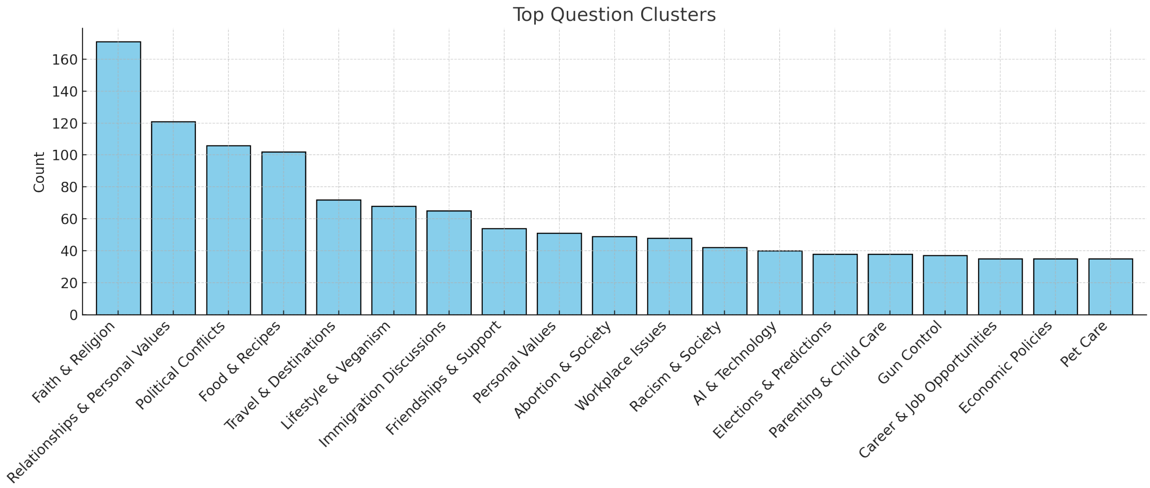

Figure 3: Distribution of prompt topics in the `Persona` dataset. The prompts are taken from (Kirk et al., 2024), and any differences in the distribution are due to filtering and difference in topics clustering.

deliberate design choice as directly matching the joint distribution of demographic characteristic and topics in the data could yield models with superficial alignment that learn to map certain topics to the demographic which engages the topic the most and align with those opinions. Instead, we would like to be able to evaluate the whole distribution of opinions and potentially teach the model to elicit preferences and information from the user and not rely on spurious correlations.

The original PRISM dataset solicits feedback on generations from several models of different sizes and capabilities. Instead we only use GPT 4 for generating answers and as an evaluator for two main reasons; first we want to disentangle the effect of model capability from the model-user alignment and GPT-4 has shown strong role-playing capability. Second, in order to create an easily accessible and reproducible test environment we want to evaluate aligned models under the same preference distribution, which generated the data, hence following prior work (Zheng et al., 2023; Dubois et al., 2023) in the "LM-as-a-judge" framework, we use also GPT 4 as an evaluator.

We construct feedback data using the the Direct Principle Feedback (DPF) approach (Castricato et al., 2024) as it tends to outperform Constitutional AI methods (Bai et al., 2022). Our data pipeline is shown in Fig. 4. Once we have the pair of prompts and personas x_i, p_i, we sample a response $y_i^l \sim \pi(y|x_i)$ from GPT 4 using only the question and not the providing access to the person profile, which we consider a proxy for the "representative" user. Then, following (Castricato et al., 2024) we further provide the initial response and the user profile and ask the model to re-write the response in order to reflect the user's values $y_i^w \sim \pi(y|y_i^l, x_i, p_i, r)$, where r is the DPF query prompt as shown in Appendix B. We then have the feedback tuple $p_i, x_i, y_i^w \succ y_i^l$ where we assume the persona p_i would always prefer the re-written response over the base model response. When we evaluate the two choices, using a role-playing evaluator, this assumption holds **96%** of the time. For every persona we sample 150 prompts from the 3000 train prompts and create a single preference pair per prompt. For personalization and preference elicitation applications, we split the 150 pairs into 100 train prompts and 50 held-out test prompts. We further sample 50 prompts from the 868 held-out test prompts and create an additional 50 preference pairs. In total the dataset contains 100 train preference pairs for each persona and 100 test preference pairs split in 50 seen prompts and 50 held-out prompts for a total of 158,600 total train preference pairs and the same amount of held-out data.

4 Dataset Analysis and Human Verification

In this section, we present an analysis of our dataset and the human verification process employed to validate the relevance of persona attributes in the decision-making process.

4.1 Leave One Out Analysis

To determine the relevance of persona attributes to the evaluation process, we performed a leave-

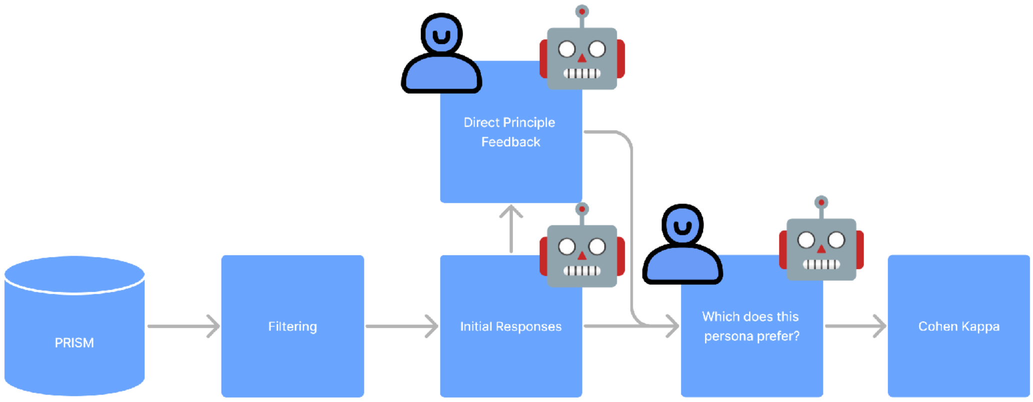

Figure 4: High level for going from the original PRISM dataset to a confusion matrix of Cohen's Kappa between simulated personas. The robot emoji signifies the inclusion of a language model, where as the person emoji signifies the use of a persona (or multiple.)

one-out analysis. For each attribute a_i, we randomly constructed 40 personas, each consisting of 3 attributes excluding a_i. We then created a corresponding set of 40 personas identical to the first set but with the addition of the LOO attribute a_i, for a total of 4 attributes per persona. Our attribute filtering process may have introduced some sampling bias. For example, when analyzing the "disability type" attribute, we first filtered our dataset to only include personas with a disability before adding the specific "disability type" attribute.

Analogous to conventional leave one out analysis, for every attribute, a_i, we had a set of personas without that specific attribute and an analogous set of personas that were identical except for the inclusion of the leave one out attribute.

We collated a set of 20 questions and baseline answers, which were used for human evaluation (see Appendix for details). For each persona pair $p_{i,j}$ (Original Personai, j, Original Personai, j + LOO Attribute), where $1 \leq i \leq |\text{attributes}|$ and $1 \leq j \leq 40$. We critiqued and refined all 20 baseline answers to make them more personalized for the given persona. The prompt used for this process can be found in the appendix.

We used Cohen's kappa quantify the agreement between annotators for the original persona and the persona with the LOO attribute concatenated. Cohen's kappa is a statistical measure to assess inter-annotator reliability that takes into account the possibility of agreement occurring by chance. For every pair $p_{i,j}$ we want to measure the annotator agreement between the original persona and the persona with the LOO attribute concatenated. This is repeated $\forall i$ s.t. $1 \leq i \leq |\text{attributes}|, \forall j$ s.t. $1 \leq j \leq 40$. We then report the distributions over these Cohen's kappa per attribute to determine which, if any, attributes are the most influential. The results, as shown in Figure 5, suggest that while the persona as a whole steers the preferences extraction process, no single attribute overpowers the persona.

We've included a number of graphs in the appendix to further explore the relationship between attributes and the overall decision making of personas.

4.2 Human Evaluation

Evaluating how humans express preferences is crucial for understanding language models' ability to emulate synthetic personas. Whether humans follow instructions similarly to language models is actively debated (Webson et al., 2023). To validate our approach, we here report inter-annotator agreement between a language model and a human imitating the same persona.

4.2.1 Experimental Design

For our human evaluation, we selected 20 personas with a fixed number of attributes, including core values and entertainment preferences. We then recruited 80 participants via Prolific Academic (Palan and Schitter, 2018), with each persona shown to 4 independent participants and each rater seeing exactly one persona. We also selected 10 questions for each persona to "answer" by initially generating one PRISM refinement step for each persona, starting with 20 questions, and then randomly sampling down to 10 due to human annotation limitations.[1]

[1]The full set of personas and questions is available here: https://sites.google.com/view/pluralistic

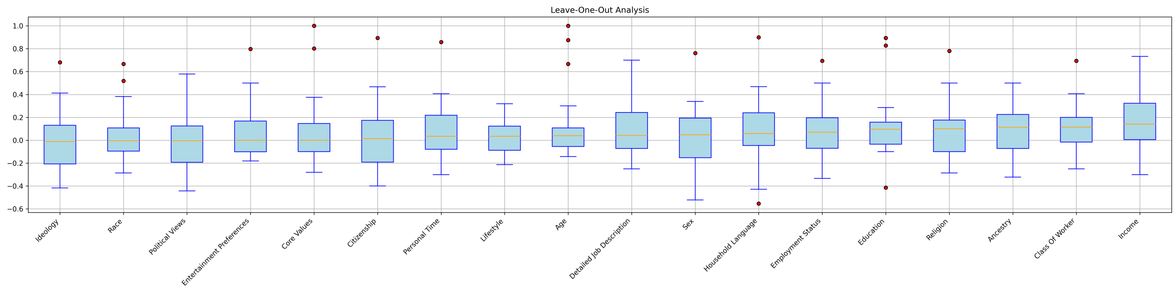

Figure 5: Leave one out analysis of various attributes of our persona. Influence is measured as the annotator agreement (Cohen's kappa) between an annotator with a given attribute and an annotator without said attribute. Lower Cohen's kappa equates to larger influence.

Each participant was presented with a page outlining what it means to imitate a "persona" (see Appendix for instructions). The full annotation UI will be available upon publication. For each persona, we took the majority answer from 3 out of 4 participants.[2]

4.2.2 Results

Our human evaluation demonstrates that state-of-the-art language models can effectively role-play diverse personas and express preferences aligning with those personas.

Both figures 6 and 7 shows the annotator agreement, measured by Cohen's Kappa, between human participants and various frontier language models (GPT-4, LLama-3 70b, Qwen 2 72b, Mistral Large) when imitating the same personas. Notably, GPT-4 achieves high agreement with human annotators, with Kappa values concentrated in the 0.6-0.8 range (substantial agreement). This suggests GPT-4 can accurately capture and express persona-specific preferences in a human-like manner.

However, the persona role-playing capabilities vary across models. As evident in Figure 7, Llama-3 70b and Mistral Large exhibit higher annotator agreement compared to GPT-4 and Qwen 2 72b. The latter two models show a wider spread of expressed opinions with lower accuracy. This indicates that while all models can role-play to some extent, their ability to align with human-like persona preferences is not uniform.

To further investigate the models' role-playing consistency, we examine the inter-annotator agreement between the models themselves when imitating the same personas (Figures 8 and 9). The confusion matrices reveal substantial agreement between models, with GPT-4 showing the highest consistency. The histograms confirm this trend, with GPT-4 exhibiting a tight distribution of high Kappa values.

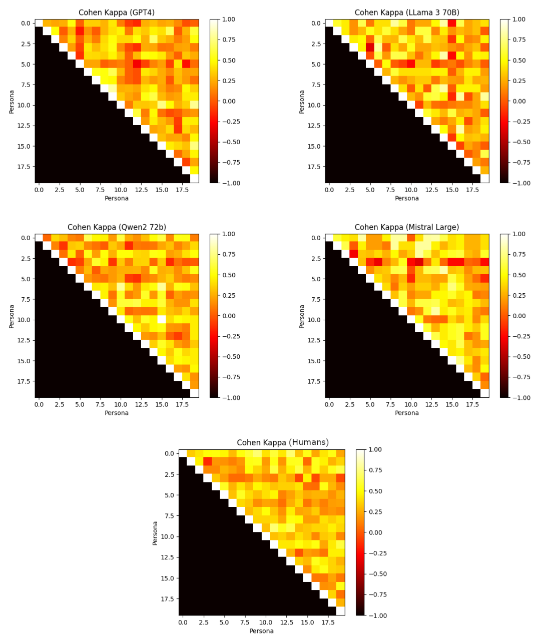

Figure 6: Annotator agreement with various frontier models. Cohen's Kappa confusion matrix. Top left is GPT-4, top right is LLama-3 70b, middle left is Qwen 2 72b, middle right is Mistral Large, bottom is a human baseline. The lower left triangular matrix is blacked out to keep the scales of the confusion matrices consistent.

These results validate our approach of using language models as synthetic personas for evaluating pluralistic alignment techniques. The high agreement between GPT-4 and human annotators, along with the inter-model consistency, suggests that carefully designed role-playing scenarios with language models can serve as a realistic and scalable testbed for assessing alignment methods without the need for human participants.

[2]The extra annotator allowed for dropping one set of annotations if needed.

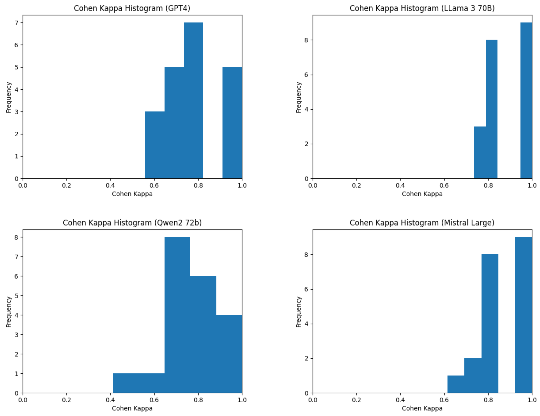

Figure 7: Annotator agreement with various frontier models. Cohen's Kappa histogram. Top left is GPT-4, top right is LLama-3 70b, bottom left is Qwen 2 72b, bottom right is Mistral Large. Note that, evident by this graph, Llama 3 70b and Mistral Large have some of the largest annotator agreements, where as GPT-4 and LLama-3 70b have some of the largest spreads of opinions they express, with relatively low accuracy.

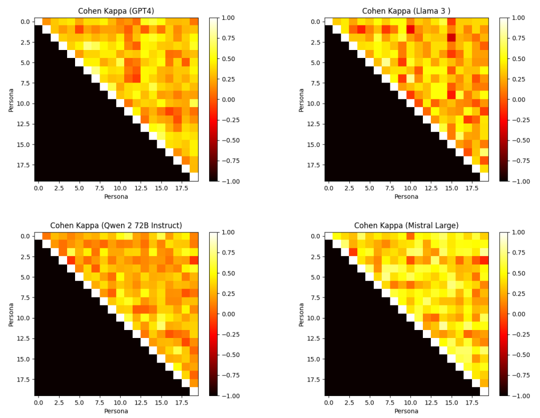

Figure 8: Inter annotator agreement (confusion matrices) for solely frontier model generated persona preferences. Top left is GPT-4, top right is LLama-3 70b, bottom left is Qwen 2 72b, bottom right is Mistral-Large.

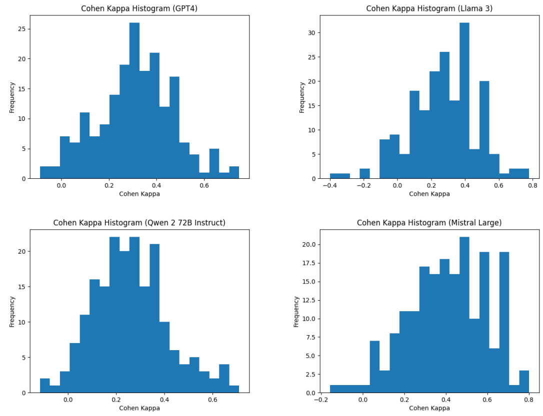

Figure 9: Inter annotator agreement (histograms) for solely frontier model generated persona preferences. Top left is GPT-4, top right is LLama-3 70b, bottom left is Qwen 2 72b, bottom right is Mistral-Large.

5 Conclusion

The advancement and wide adoption of language models has raised a number of important concerns around fairness and pluralistic alignment to the values of diverse users, which still remains a challenge. Beyond group-level preferences, personalized models, tailored to specific individual needs and preferences are a promising application. Despite the concerns and opportunities raised by these issues, current large-scale RLHF pipelines still work under the assumption of a representative user and do not account for the distributional nature of values. While a number of academic works have proposed approaches for pluralistic alignment, personalization and preference elicitation, these are still not widely adopted, partially due to lack of convincing evaluations as current benchmark consists of unrealistic multiple-choice questions or simple domains. In this work we aim to address this challenge by creating a test environment and benchmark for these issues. We propose an automated LM as-a-judge approach based on current state-of-the-art systems role-playing capabilities. We create a demographic of 1000 train and 568 test realistic `personas` based on US census demographics and individualized profiles with idiosyncratic personality types. We further utilize a wide real user survey controversial topics to create a large-scale synthetic datasets of diverse feedback with over 158,600 train preference pairs and a comparable number of evaluation datapoints. Our proposed environment can be used to develop and evaluate approaches on pluralistic alignment with diverse group preferences, individualized models and information-gathering and preference elicitation. We further validate the fidelity of these personas with a real user study.

We believe our work will facilitate the developemnt of new alignment approaches, but a open questions remain. In this construction we focused exclusively on US demographics and user profiles, which are not representative of global populations. These users might already be over-represented in LM training data (hence the advanced role-playing capabilities of GPT 4 on this demographic).

Further work would evaluate different LM model's capabilities to represent a global audience and expand the persona demographics to include these populations as as well.

6 Limitations

Our work has several potential limitations.

Demographic Focus: Our personas are based on US demographic data, which may not accurately represent the diversity of global populations. This limitation could impact the generalizability of our findings to non-US contexts. Future work should aim to include a more diverse set of personas reflecting global demographic and cultural variations.

Feedback and Preference Data: The preference data generated in this study relies on the responses of language models in role-playing scenarios. While we validated these responses through human judges, there remains a risk that the feedback does not perfectly mimic real human preferences. Additionally, the Direct Principle Feedback (DPF) approach, although effective, may not capture all nuances of human decision-making and preference.

Model Limitations: The language models used to generate and evaluate personas are themselves subject to biases and limitations. Current state-of-the-art models, such as GPT-4, have shown strong role-playing capabilities, but they are not infallible and may produce outputs that are biased or inconsistent. Moreover, the role-playing capabilities of these models might not extend uniformly across different types of personas, especially those representing underrepresented or marginalized groups.

Evaluation Metrics: The use of Cohen's kappa and other inter-annotator agreement metrics provides a measure of consistency but may not fully capture the qualitative aspects of alignment with human preferences. These metrics focus on agreement rates, which do not necessarily reflect the richness and contextual appropriateness of the model's responses.

Real-World Application: While our synthetic approach allows for scalable testing and evaluation, it does not fully address the challenges of real-world deployment. The dynamics of real user interactions, continuous learning, and adaptation to evolving preferences are complex and require more extensive field testing and longitudinal studies.

Bias Concerns: The creation and use of synthetic personas must be approached with caution to avoid perpetuating stereotypes or introducing new biases. Our study attempts to mitigate these risks through careful design and validation, but there remains a possibility that some biases are not fully addressed.

In summary, while PERSONA provides a valuable testbed for evaluating pluralistic alignment in language models, these limitations highlight the need for ongoing research and development to refine these methods and ensure their applicability and fairness in diverse real-world settings.

7 Ethical Considerations

Developing language models that accurately represent and align with the diverse values and preferences of users is crucial for ensuring fair and inclusive AI systems. However, the use of synthetic personas and simulated feedback raises important ethical considerations. Although our personas are based on anonymized public domain US census demographics, they may not fully capture the nuances and complexities of individual identities. We acknowledge that personas can perpetuate stereotypes and biases if not carefully constructed. Future work should expand persona demographics to be more globally representative and further validate persona fidelity with diverse human participants.

Second, the use of language models for generating synthetic feedback and evaluating alignment approaches raises concerns about the reproducibility and robustness of our findings. We mitigate this by validating persona fidelity with human judges, but further research is needed to understand the limitations and biases of language models in this context.

In our human evaluation, we ensured fair compensation for our annotators, paying them at a rate of $40 per hour. We also obtained informed consent from our annotators, clearly communicating that their input, feedback, and annotations would be used for machine learning training purposes. We did not store any demographic data from participants. We filtered for EFL Americans.

Finally, our work aims to facilitate the development of alignment approaches that better represent and serve diverse users. However, we recognize that pluralistic alignment is an ongoing challenge that requires continuous effort and engagement with affected communities. We encourage future research to prioritize the voices and needs of marginalized groups in the development and evaluation of these technologies. By openly acknowledging these ethical considerations and calling for further research, we hope to contribute to the responsible development of language models that promote fairness, inclusivity, and accountability.

References

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024a. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024b. Star-gate: Teaching language models to ask clarifying questions. *Preprint*, arXiv:2403.19154.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, et al. 2022. Constitutional ai: Harmlessness from ai systems using constitutional principles. *ArXiv*.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *Preprint*, arXiv:2403.18932.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. 2024. Suppressing pink elephants with direct principle feedback. *Preprint*, arXiv:2402.07896.

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

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. 2024. Direct preference optimization with unobserved preference heterogeneity. *arXiv preprint arXiv:2405.15065*.

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

Y. Dubois, N. Du, K. Zhang, Y. Zhang, S. Agrawal, Y. Cao, S. Salehi, J. Kim, S. Li, S. Zhang, et al. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *ArXiv*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Jan-Philipp Fränken, Sam Kwok, Peixuan Ye, Kanishk Gandhi, Dilip Arumugam, Jared Moore, Alex Tamkin, Tobias Gerstenberg, and Noah D Goodman. 2023. Social contract ai: Aligning ai assistants with implicit group norms. *arXiv preprint arXiv:2310.17769*.

Jan-Philipp Fränken, Eric Zelikman, Rafael Rafailov, Kanishk Gandhi, Tobias Gerstenberg, and Noah D Goodman. 2024. Self-supervised alignment with mutual information: Learning to follow principles without preference labels. *arXiv preprint arXiv:2404.14313*.

S. Borgeaud Y. Wu J.-B. Alayrac J. Yu R. Soricut J. Schalkwyk A. M. Dai A. Hauth et al. Gemini Team, R. Anil. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *Preprint*, arXiv:2310.11564.

Kaixuan Ji, Jiafan He, and Quanquan Gu. 2024. Reinforcement learning from human feedback with active queries. *Preprint*, arXiv:2402.09401.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. *Preprint*, arXiv:2310.18168.

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

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? *Preprint*, arXiv:2305.13788.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023a. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023b. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *Preprint*, arXiv:2402.05133.

Siyang Liu, Trish Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024. The generation gap:exploring age bias in the underlying value systems of large language models. *Preprint*, arXiv:2404.08760.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. 2023. Sample efficient reinforcement learning from human feedback via active exploration. *Preprint*, arXiv:2312.00267.

Meta. 2024. Introducing llama3.

Tea Mijač, Mario Jadrić, and Maja Ćukušić. 2018. The potential and issues in data-driven development of web personas. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1237–1242.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. *Preprint*, arXiv:2402.08114.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. 2023. Active preference inference using language models and probabilistic reasoning. *Preprint*, arXiv:2312.12009.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. *Preprint*, arXiv:2406.02900.

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

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing political bias in large language models. *Preprint*, arXiv:2405.13041.

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

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

Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. 2024. Show, don't tell: Aligning language models with demonstrated feedback. *arXiv preprint arXiv:2406.00888*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *Preprint*, arXiv:2312.08358.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117–143.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

United States Census Bureau. 2024. American community survey (acs) public use microdata sample (pums).

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Albert Webson, Alyssa Marie Loo, Qinan Yu, and Ellie Pavlick. 2023. Are language models worse than humans at following prompts? it's complicated. *arXiv preprint arXiv:2301.07085*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *Preprint*, arXiv:2404.12138.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. 2024. Self-exploring language models: Active preference elicitation for online alignment. *Preprint*, arXiv:2405.19332.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Full list of attributes

The following is the full list of persona attributes.

1. age
2. sex
3. race
4. ancestry
5. household language
6. education
7. employment status
8. class of worker
9. industry category
10. occupation category
11. detailed job description
12. income
13. marital status
14. household type
15. family presence and age
16. place of birth
17. citizenship
18. veteran status
19. disability
20. health insurance
21. big five scores
22. defining quirks
23. mannerisms
24. personal time
25. lifestyle
26. ideology
27. political views
28. religion
29. cognitive difficulty
30. ability to speak English
31. vision difficulty
32. fertility
33. hearing difficulty

B Persona Critique and Refinement Prompt

The following is the critique prompt that was used.

```
f"Examine the COMPLETION: '{preference}' in relation "
"to the DEMOGRAPHIC: '{persona}' and the INSTRUCTION: " '{preference.meta_data['instruction']}'. "
"Put yourself in the shoes of DEMOGRAPHIC. "
"The demographic prefers short answers.  "
" If you give a long suggestion, they will hate it. "
"Identify the ways the completion both does and does not resonate with the demographic. "
"Provide a concise explanation, quoting directly from the demographic
and completion to illustrate your evaluation. "
"Think step by step about how you will make the response shorter or the same length before

providing your evaluation and suggestions. "
"Similarly, make sure that the response given is still relevant to the INSTRUCTION. "
"Format: EVALUATION: ... SUGGESTIONS: ...\nDONE"
```

The following is the revision prompt that was used.

```
f"Revise the COMPLETION: '{preference}', "
"with respect to INSTRUCTION: " "'{preference.meta_data['instruction']}'

based on the CRITIQUE: '{critique}'. "
"Provide a revision of the completion, do not make ANY "
"references to the exact preferences or attributes "
"of the demographic. "
f"Remain subtle and indirect in your revision. "
"Make sure your response has less tokens than the original completion. "
"If you make it longer you are a BAD CHATGPT. "
"Format: REVISED PREFERENCE: ...\nDONE"
```

C Complete Example Persona

The following is an example of a persona

```
     'age': 73,
 'ancestry': 'Filipino',
 'big five scores': 'Openness: Extremely High, Conscientiousness: Low, '
               'Extraversion: Extremely High, Agreeableness: Low, '
               'Neuroticism: Extremely Low',
 'citizenship': 'U.S. citizen by naturalization',
 'class of worker': 'Retired',
 'cognitive difficulty': nan,
 'defining quirks': 'Enjoys gardening and has a green thumb',
 'detailed job description': 'Retired, previously worked in a managerial '
                              'position',
 'disability': nan,
 'education': "Bachelor's Degree",
 'employment status': 'Not in labor force',
 'family presence and age': 'With related children 5 to 17 years only',
 'fertility': nan,
 'health insurance': 'With health insurance coverage',
 'hearing difficulty': nan,
 'household language': 'Asian and Pacific Island languages',
 'household type': 'Married couple household, no children of the householder '
                    'less than 18',
 'ideology': 'Liberal',
 'income': '178900',
 'industry category': nan,
 'lifestyle': 'Active and outdoorsy',
 'mannerisms': 'Often uses hand gestures while speaking',
 'marital status': 'Married',
 'occupation category': nan,
 'personal time': 'Spends free time gardening or reading',
 'place of birth': 'Philippines',
 'political views': 'Democrat',
 'race': 'Asian',
 'religion': 'Other Christian',
 'sex': 'Female',
 'veteran status': 'Non-Veteran',
 'vision difficulty': nan}
```

```
        ability to speak english': nan,
 'age': 10,
 'ancestry': 'Mixed',
 'big five scores': 'Openness: Extremely High, Conscientiousness: Average, '
                    'Extraversion: Extremely Low, Agreeableness: Extremely '
                    'High, Neuroticism: Average',
 'citizenship': 'Born in the United States',
 'class of worker': 'Not applicable',
 'cognitive difficulty': nan,
 'defining quirks': 'Prefers to express herself through drawing',
 'detailed job description': 'Student',
 'disability': nan,
 'education': 'Grade 3',
 'employment status': 'Unemployed',
 'family presence and age': 'With related children under 5 years and 5 to 17 '
                            'years',
 'fertility': nan,
 'health insurance': 'With health insurance coverage',
 'hearing difficulty': nan,
 'household language': 'Spanish',
 'household type': 'Married couple household with children of the householder '
                   'less than 18',
 'ideology': 'Believes in fairness and kindness',
 'income': '0',
 'industry category': 'Not applicable',
 'lifestyle': 'Active and curious',
 'mannerisms': 'Often hums while concentrating',
 'marital status': 'Never married or under 15 years old',
 'occupation category': 'Student',
 'personal time': 'Spends free time drawing or reading',
 'place of birth': 'California/CA',
 'political views': 'Too young to have political views',
 'race': 'Two or More Races',
 'religion': 'Protestant',
 'sex': 'Female',
 'veteran status': 'Not applicable',
 'vision difficulty': nan}
```

D Annotation Instructions

Welcome to the Persona Annotation Task!

 In this task, you will be asked to role-play as a specific persona and answer a series of preference questions.
 1. Task Explanation: We will provide you with a set of descriptors of a particular person. This person may or may not actually exist. Your job is to put yourself into the mindset of a person with those attributes.
 2. Instruction following: You will be presented with a hypothetical question that a person could ask. Your job is to select the answer that a person with the attributes that you are impersonating would prefer.
 3. Explain your reasoning: Justify your choice. It is ok to change your choice while thinking through your justifcation. In the textbox provided below the prefernece selection, go into detail about why you think your choice is correct. If there is no clear choice, pick the one that is most likely, just still attempt to justify your selection.
 4. Provide good reasoning: The better your reasoning, the bigger your bonus will be.
 5. ChatGPT (or other chatbots) are NOT allowed: Any use of ChatGPT for soliciting preferences or reasoning will result in disqualification.
 You must each question based on how you think the given persona would respond, not based on your personal preferences.

 Thank you for participating!

E Census Demographics Statistics

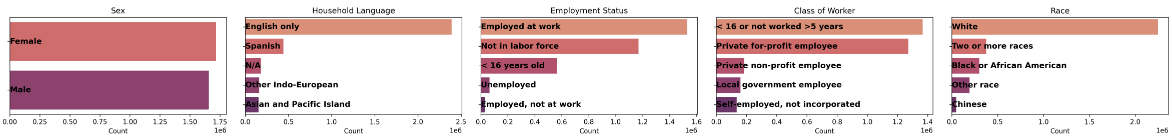

Figure 10: Histogram of demographics statistics from US Census (United States Census Bureau, 2024).