PRIMEX: A Dataset of Worldview, Opinion, and Explanation

Anonymous ACL submission

Abstract

As the adoption of language models advances, so does the need to better represent individual users to the model. Are there aspects of an individual's belief system that a language model can utilize for improved alignment? Following prior research, we investigate this question in the domain of opinion prediction by developing PRIMEX, a dataset of public opinion survey data from 885 US residents with two additional sources of belief information: written explanations from the respondents for why they hold specific opinions, and the Primal World Belief survey for assessing respondent worldview. We provide an extensive initial analysis of our data and show the value of belief explanations and worldview for personalizing language models. Our results demonstrate how the additional belief information in PRIMEX can benefit both the NLP and psychological research communities and open up avenues for further study.

1 Introduction

Psychological research and clinical successes give evidence that a person's beliefs about themselves, their future, and their environment shape their behavior (Beck, 1976; Dweck et al., 1995; Hofmann et al., 2012). Recent work shows that an individual's *worldview* — or beliefs about the overall character of the world — can explain not only persistent behavioral patterns but correlates with personality, well-being, political, religious, and demographic variables (Clifton et al., 2019). As such, worldview can be viewed as a powerful, compact, and predictive model of the individual's belief system.

Simultaneously, advancements in NLP have made it possible to incorporate higher-level user beliefs into predictive models (Sun et al., 2024). A better understanding of individual belief systems can improve personalization of language models, for instance by building better representations of an individual user's *persona* – characteristics, preferences, and behavior. Persona-adapted language models (PA-LMs) have been used to create realistic simulated communities (Park et al., 2022; Zhou et al., 2024; Park et al., 2024), generate arbitrary amounts of diverse, synthetic data (Moon et al., 2024; Ge et al., 2024), and simulate partners in training applications for a variety of professional domains (Markel et al., 2023; Louie et al., 2024; Shaikh et al., 2024). A common evaluation of PA-LMs is predicting user responses to surveys and behavioral tests (Argyle et al., 2023; Santurkar et al., 2023; Hwang et al., 2024; Joshi et al., 2025).

To facilitate both worldview and persona research, we introduce the PRIMEX dataset of opinions, explanations, and beliefs about the world. PRIMEX consists of anonymous survey responses from 885 US residents from various geographic regions, age groups, education levels, and genders. Our respondents complete a subset of questions from each of three American Trends Panel public opinion surveys (Pew Research Center, 2014), allowing for the study of a single individual's opinions across different topics, which is not possible with existing datasets. We also collect two supplemental categories of user information which are, to our knowledge, novel in persona research. First, for a portion of opinion questions, we collect free-form written explanations of the respondent's opinions. These explanations often draw on the respondent's higher-level beliefs about the world. We show that these explanations can help PA-LMs predict an individual's other opinions.

Second, we collect participants' responses to the 18-question version of the Primal World Beliefs survey, an instrument for characterizing an individual's worldview which generally takes less than 10 minutes to answer (Clifton et al., 2019; Clifton and Yaden, 2021). We find significant correlations between worldview and opinions across topics and show that worldview impacts stylistic characteristics of written explanations. Including worldview in user representations for PA-LMs cam improve

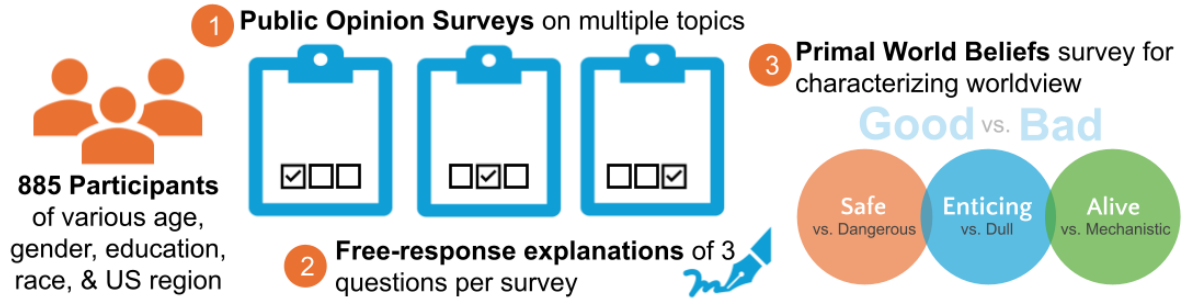

Figure 1: Overview of the PRIMEX data. We collect three types of responses from a diverse pool of participants: Opinions from 3 Pew Research surveys; explanations for 3 opinions per survey; and Primal World Belief survey of participant worldview.

opinion prediction. Additionally, we show how an individual's Primal World Beliefs can be predicted from their opinions and explanations, an interesting new avenue for building general user representations from specific user data.

Our experiments and analysis of PRIMEX data highlight the value of belief explanations and worldview for personalizing language models. Though extensive, they are far from exhaustive — we believe this dataset constitutes a rich source of persona information for continued analysis in both the NLP and psychological research communities.[1]

2 Background

Primal World Beliefs Primal World Beliefs, or *Primals*, aim to capture an individual's beliefs about the general character of the world (Clifton et al., 2019). Examples of Primals include *The world is Safe* and *The world is Interesting*. Research has shown these beliefs to be stable across time and correlated with a number of personality and well-being variables. We hypothesize that LMs have some knowledge of Primals due to how the theory of Primals itself was developed, which involved extensive linguistic analysis of text that is likely to be part of many LM's pretraining data. In particular, researchers scoured hundreds of historical texts (including sacred texts, novels, films, speeches, and philosophical works) and over 80K tweets for statements about how people view the world as a whole, using NLP extraction tools and Latent Dirichlet Allocation for topic clustering. Over a span of 5 years (2014-2019), they coded the statements and consulted with social science experts as well as religious focus groups to identify 26 Primal World Beliefs. Primals are organized under the top-level belief that *The world is Good* and secondary beliefs that it is *Safe* (versus dangerous), *Alive* (intentionally and purposefully interacting with us versus inanimate and mechanical), and *Enticing* (interesting and beautful versus dull and ugly). These beliefs were ultimately validated through multiple psychometric measures. In our dataset, participants filled out the 18-item survey (Clifton and Yaden, 2021) measuring their top-level and secondary Primals as part of their 30 minute session; in general, it took less than 10 minutes for most participants to fill out the survey.

Public Opinion Public opinion surveys are used in PA-LM research due to their easy availability and the rigorous validation of their construction over many decades (Pew Research Center, 2014). The complexity and deeply personal nature of opinion offers a difficult challenge for personalized ML, and only recently have models become powerful enough to take on this task. Prior opinion datasets for PA-LM research borrow data originally intended for demographic and economic analysis, resulting in limited individual information (Santurkar et al., 2023). Our work enriches public opinion data by addressing several shortcomings that hinder generalization: 1) opinions from a single user across multiple topics are not available; and 2) demographic distributions of existing data can bias the output of LLMs, which often under-represent certain viewpoints. In addition, our data allows us to correlate opinion and demographics with new variables of interest: viz., worldview and explanation style.

Explanations Social scientists often conduct free-form interviews, in part because participant explanations of responses can provide deeper insights than structured formats (Stanford Center on

[1] Our data will be made available for further study.

Poverty and Inequality, 2021). Inspired by this, we ask our participants to explain a subset of their survey opinions in a free text format in hopes of deriving a better understanding of their personae. A work similar to ours has demonstrated the value of conducting a free form interview followed by refinement processes, but this method of gathering persona information is both expensive and intensive for users (Ge et al., 2024). Our work introduces a lower-cost persona format and elucidates how explanation interacts with both opinion and worldview. Model-generated explanations of reasoning have proven useful for improving performance on many tasks (Wei et al., 2023), including preference modeling and opinion prediction (Sun et al., 2024; Do et al., 2025; Joshi et al., 2025). We analyze human-written explanations and model-generated explanations for prediction to determine characteristics of helpful or unhelpful explanations.

Personalized LMs The advancement of large language models has enabled new possibilities for personalized machine learning. Adapting an LM to the preferences of individuals can be done via alignment strategies such as RLHF and DPO, but these require expensive, large scale data (Ouyang et al., 2022; Rafailov et al., 2023; Wu et al., 2023). Recent datasets for personalizing LMs address issues of representation (Kirk et al., 2024; Aroyo et al., 2023), but focus on demographics of human feedback data for conversational content. Weaker personalization can be accomplished quickly and cheaply using low-data techniques such as prompting (Hwang et al., 2023) or refinement (Sun et al., 2024). These works make use of persona to adapt language models to an individual user's preferences. PRISMX provides rich user data and can serve as training and testing data for personalization methods.

3 Dataset Construction

The goal of PRISMX is to extend current resources along multiple dimensions. Addressing a shortcoming in existing opinion data, we collect responses on multiple topics from each individual. This enables the development of personae which generalize across topics. For a subset of opinion questions, we collect free-form explanations for why the respondent holds their particular opinion. Explanations give insights into an individual's belief system and can also improve personae development. Lastly, we consider a source of user information which has not yet been brought to bear on PA-LM development: individual worldview. Prior works have made some use of Big 5 personality traits (Goldberg, 1993), but worldview has been shown to explain broader aspects of personality (Clifton et al., 2019). Hence, we collect responses to the 18 question Primal World Beliefs survey to capture an individual's worldview. In total, PRISMX includes responses from 885 individuals, similar in size to recent works in LM personalization (Park et al. (2024), $N = 1052$) and personality psychology (Ludwig et al. (2022), $N = 529$)

Do you favor or oppose the use of animals in scientific research?
User 1: Favor — Most of the vaccines and oncology drugs were discovered and invented due to trials on animals which I think I favor
User 2: Oppose — It is clear by now that animals experience a range of emotions just like we do, so what was once thought acceptable is no longer. Just because we have all the power doesn't mean we should inflict pain.
User 3: ...

Figure 2: Examples of opinions with user explanations.

3.1 Survey Questions

Our data consists of three types of questions: public opinion, free-response, and Primals. We use public opinion questions from the American Trends Panel surveys (Pew Research Center, 2014), which have been carefully developed by experts at Pew Research to mitigate bias, ambiguity, difficulty, and other confounding factors. We choose 10 questions from each of 3 surveys: ATP Wave 34, dealing with biomedical and food issues (Pew Research Center, 2018a); ATP Wave 41, dealing with the condition of America in the year 2050 (Pew Research Center, 2018b); and ATP Wave 54, dealing with economic inequality (Pew Research Center, 2019). From each of these surveys, we manually select questions that meet two characteristics: they ask about personal opinions rather than biological or economic facts; and their response distribution in the larger population has higher entropy, as these are more likely to produce diverse answers from our participants. The full list of questions, response choices, and shorthand names used in this work (e.g. ORGANIC FOODS, GOVT RETIREMENT) are listed in Table 10 in Appendix A.2.

For 3 questions from each ATP survey, we ask participants to explain their answer in a free-

response format. We instruct respondents to "draw on any aspect of your personal history, social life, experiences, thoughts, feelings, beliefs, or values" in their explanation. Examples of elicited explanations are shown in Figure 2 (additional examples in Appendix A.3).

We also include the 18 item Primal World Beliefs Inventory (PI-18) (Clifton and Yaden, 2021). This shorter instrument balances brevity and granularity, measuring top-level (*Good*) and secondary Primals (*Safe*, *Enticing*, and *Alive*). The PI-18 has been shown to have high correlation with the full 99 question inventory. Questions in the PI-18 are multiple choice with responses ranging from "Strongly Disagree" to "Strongly Agree", which are converted to real-values ranging from 0 to 5. We follow the administration and scoring instructions given in Clifton (2021) (questions and scoring functions are repeated in Appendix C).

In addition to the opinion, explanations, and Primal World Belief data, we also ask questions covering basic demographic self-identification: geographic region, age range, gender, English proficiency, number of children, employment status, political affiliation, hobbies, other languages spoken at home, and races.

3.2 Data Collection

We recruit 885 participants through a third-party user study firm, User Research International. Participants were selected to achieve relative balance in terms of male/female ratio, age range, and geographic distribution.[2] Each participant was offered a fair wage for their participation in the survey. The projected time to complete the survey was 30 minutes. Participants gave their informed consent before participation and were made aware of the intended uses of their data. They were offered the chance to stop at any time, and given the option to not answer any question.

The survey questions were presented in the same order for all participants: first the subsampled ATP Wave 34, ATP Wave 54, and ATP Wave 41 surveys with additional explanation questions, followed by the PI-18, and lastly some additional optional additional demographic questions. The 3 opinion questions which are each followed by explanations are given at the beginning of each section, followed by the 7 remaining opinion questions from the same ATP survey. This was done in an effort to reduce the cognitive load required by switching between topics. The order of questions within sections of the survey was also fixed.

[2]We provided a non-binary gender option, but we did not control for representative non-binary participation.

Primal	*N*=	Avg	Std	min	max	US Avg.
Good	809	3.08	0.69	0.53	4.93	2.9
Safe	853	2.50	0.90	0.0	5.0	2.5
Alive	805	2.65	1.09	0.0	5.0	2.8
Enticing	856	3.72	0.76	0.57	5.0	3.7

Table 1: Primal Belief scores of our respondents

From each of our 885 participants we collected 30 opinion question responses, 9 explanations, answers to the Primal World Beliefs survey comprising 18 scalar ranked questions questions, and 11 demographic attributes. A condensed version of the demographic distribution of this data is provided in Table 9 in Appendix A.1. Our respondents reflect a balance of geographic regions and age groups. To maintain this balance, we struggled to recruit male and female respondents at equal rates resulting in a female bias. Compared to the national average, people with college degrees or higher are overrepresented in our data. The reported race of our respondents shows nationally representative numbers of Black and Asian respondents, a slight over-representation of White respondents, and under-representation of Spanish, Hispanic, or Latino respondents. Future data collection efforts should consider additional controls for better representation of this demographic if necessary.

4 Analysis of Primals

The PI-18 measures the top-level *Good* Primal and secondary *Safe*, *Enticing*, and *Alive* Primals. The aggregate statistics for our respondents is shown in Table 1 along with the US average reported in existing research (Clifton, 2018). Our sample averages are similar to the US population, and our standard deviations cover the spread of reported scores. Our respondents can choose not to answer any questions including those needed to compute their score for a particular Primal, resulting in different but still large sample sizes N.

4.1 Primals and Opinion

We compute correlations between Primals and responses to each opinion question to determine the effect of a higher or lower score for each Primal on a person's opinion. We ignore "Prefer not to answer" opinion responses and respondents without a particular high-level Primal score on a per-

	Good			Safe			Alive			Enticing		
	↑	↓	Δ	↑	↓	Δ	↑	↓	Δ	↑	↓	Δ
Length	173	180	3.89%	158	186	17.92%	163	249	52.48%	201	164	-18.59%
1st Person	0.57	0.53	-6.96%	0.64	0.47	-26.77%	0.60	0.46	-24.00%	0.52	0.53	1.74%
All Pronouns	1.33	1.32	-1.15%	1.47	1.24	-15.69%	1.37	1.04	-24.01%	1.29	1.27	-1.51%

Table 2: Lexical characteristics of explanations from users with highest ↑ and lowest ↓ Primal scores.

question/primal basis, resulting in different but still sizable sample size N for each correlation. Opinion questions with 2 response options are treated as binary; the remaining opinion questions are mapped to integers following Santurkar et al. (2023). The full tables of correlations for all Primals and opinion questions are shown in Appendix B.

Effect size Funder and Ozer (2019) recommend reporting effect sizes relative to a benchmark for comparison. One benchmark they suggest, which we use in this work, is the typical effect size found in a large scale literature review. The average effect size of 708 meta-analytically determined correlations in personality and individual difference research determined by Gignac and Szodorai (2016) is $r = 0.19$. As such, we follow their suggested thresholds for relatively small ($r = 0.1$), typical ($r = 0.2$), and relatively large ($r = 0.3$).

Notable Effects We observe relatively large correlations between the opinion that children will have a better standard of living in the future (CHILD STANDARD) and high *Good* and *Safe* scores ($r = 0.341$ and $r = 0.325$ respectively). Not surprisingly, we observe a large correlation between the role of God versus evolution in determining the development of human life (EVOLUTION) and high *Alive* scores ($r = 0.32$). Interestingly, we also observe relatively large correlations between how much gas prices impact the view of the economy (GAS PRICES) and *Alive* scores ($r = 0.308$).

Overall, we observe small or stronger correlations ($r > 0.1$) between all Primals and at least some questions from each topic with the exception of *Enticing* and questions on economic inequality. The strongest correlations are found between Primals and questions from Wave 41 on the likely condition of America in 2050. This may indicate that Primals better encode how a person views the future world compared to the present one, but further analysis is needed.

4.2 Primals and Explanations

Table 2 shows variations in lexical features of explanations based on the respondent's Primal scores. In this table, we group the 50 respondents with the highest and lowest scores for a given Primal (900 explanations per Primal in total). We compute the average explanation length in characters, as well as counts average pronouns per sentence which may indicate belief. The Δ column indicates the change moving from the high to low group for each Primal.

We observe large differences in average explanation length; respondents with lower *Safe*, lower *Alive*, or higher *Enticing* scores give much longer responses than their counterparts. Pronoun usage mostly correlates inversely with the length of the explanations, though there is a larger difference in first person pronoun usage between people with high and low *Safe* scores and a almost no difference between people with high and low *Enticing* scores. A higher *Safe* score has been correlated with lower neuroticism and higher trust, which may enable such people to speak freely about themselves in their explanations.

An interpretable vocabulary analysis of explanations from these groups is complicated by the underlying topicality of the explanations and the length difference between responses from each group, but predictive results of LMs conditioned on explanations in Section 5 shed some light on the differences between text from these groups.

5 Predicting User Responses

In order to highlight the value of PRIMEX for personalizing language models, we now consider the problem of predicting the survey responses of a user in our dataset using a PA-LM. Prior works on opinion prediction represent a user by their demographic attributes (Santurkar et al., 2023) or by including a seed set of opinion questions and the user's answers (Hwang et al., 2023). We study the value of the additional data from PRIMEX— Primals and explanations – in user representations. Our data also enables the analysis of representation generalization through the prediction of opinions

User Representation	GPT-4o	Mistral
All Topics		
DEMOGRAPHICS	42.22	42.30
DEMO & OPINIONS	45.10	44.30
+ Primals	46.17	44.24
+ Explanations	48.09	46.02
+ Generated Explanations	46.15	45.02
PRIMEX PERSONA	48.28	45.84
Cross Topic		
DEMOGRAPHICS	39.13	39.61
DEMO & OPINIONS	39.68	39.90
+ Primals	40.31	40.89
+ Explanations	40.17	40.29
+ Generated Explanations	39.91	40.21
PRIMEX PERSONA	40.57	40.67

Table 3: Predicting user opinions from PRIMEX. Underlined results are significantly different from DEMO & OPINIONS at $p < 0.05$.

of the same user across different topics. Finally, we explore whether a model can predict a user's Primals from their persona, and find that training on PRIMEX data facilitates this prediction.

5.1 Opinion Prediction

Task In the opinion prediction task, a model is prompted with a user representation and instructions to predict the user's response to unseen test questions one at a time. In the *all topics* settings, seed opinions included in the user representation and test questions are drawn from all Waves of the ATP survey. We use the 9 explained opinions as seeds and test on the remaining 21 for each user. In the *cross topic* settings, seed opinions include all 10 opinions from Wave 34 and the test questions come from Waves 41 and 54. This is a harder setting, since less is known about the user's opinion within a given test topic. To enable the study of finetuned models, we split PRIMEX in half for training and testing (442 users). We use both a large (GPT-4o (OpenAI et al., 2024)) and smaller (Mistral 7B Instruct (Jiang et al., 2023)) for predicting survey responses from the PRIMEX test set given different user representations.

User Representations The main baseline for comparison is DEMO & OPINIONS, which represents users with their demographics and seed opinions. To this we add different types of novel data: The *+Primals* setting includes information from the Primal World Beliefs survey. For long context models (GPT-4o), we include Primal scores with contextualizing information from Clifton (2018); for short context (Mistral) we provide the question/response pairs from the user's PI-18. The *+Explanations* setting includes the human written explanations for seed opinions (all 9 seed opinions have explanations in *all topics*; 3 of the 10 in *cross topic* settings). The PRIMEX PERSONA setting uses all information collected from users (demographics, seed opinions, explanations, and Primals).

Primal	Correlated	Uncorrelated
Good	49.68	47.30
Safe	51.38	46.07
Alive	49.25	44.39
Enticing	53.51	46.82

Table 4: Accuracy of model predictions for correlated and uncorrelated questions.

To explore the generalization capability of the explanations in PRIMEX, we study a *+ Generated Explanations* setting. We use a finetuned GPT-4o to explain each seed opinion in the test set independently and include the generated explanations for each user in their representations. The explanation model is finetuned from GPT-4o with a user's demographics and a seed opinion as input and the user's explanation as the target output. Finally, we include a demographics-only setting (DEMOGRAPHICS) which allows the default alignment positions of models to be more prominent in the response distribution.

Results Table 3 shows the average per-user zero-shot opinion prediction performance on the PRIMEX test set. Underlined results are significantly better than the DEMO & OPINIONS baseline for each model using paired t-tests with $p < 0.05$. We see that both the explanations and worldview information provided in PRIMEX enables better prediction of unseen opinions. In the *all topics* setting, GPT-4o can effectively use all information from PRIMEX including model-generated explanations, whereas Mistral requires human explanations to achieve significant results. In the *cross topic* setting, we see that both models struggle to generalize from off topic explanations but can combine these with worldview in the PRIMEX PERSONA setting to make significant improvements in prediction. Notably, the smaller Mistral model benefits more from worldview in the *cross topic* setting, indicating the generality of this form of user data.

Primals and Accuracy Continuing the analysis of Section 4, we compare the model prediction accuracy of opinion questions which are correlated with different Primals against those which are un-

	Good	Safe	Alive	Enticing
				GPT-4o
DEMOGRAPHICS	0.14	0.11	-0.04	0.09
PRIMEX PERSONA	0.06	0.01	-0.04	0.03
				Mistral
DEMOGRAPHICS	0.14	0.10	-0.12	0.12
PRIMEX PERSONA	0.04	0.01	-0.06	0.04

Table 5: Correlation of model accuracy and user Primal score. Underlined values are significant at $p < 0.05$

	Good	Safe	Alive	Enticing
DEMO & OPINIONS	0.56	1.13	1.43	0.61
+ *Explanations*	0.55	1.41	1.06	0.63
TRAINED	0.46	0.70	1.22	0.65

Table 6: Predicting a user's Primals (MSE).

correlated. We use a threshold of $r = 0.1$ to distinguish these sets of questions. Table 4 shows the average accuracy of the DEMO & OPINIONS+ *Explanations* representation for the test opinions in these groups. We see that both in aggregate and across individual Primals the model is better at predicting opinions for correlated questions. These trends hold for other user representations but with smaller gaps between the accuracies. This indicates that the Primal beliefs involved in the correlations identified in Section 4 are partly encoded by other user demographics, seed opinions, or explanations.

In Table 5 we show correlations between a user's Primals and opinion prediction accuracy under different user representations and models. Using the DEMOGRAPHICS representation, models are more accurate for users with higher Good, Safe and Enticing beliefs, and for users with lower Alive beliefs. This aligns with results reported in Santurkar et al. (2023) showing that the default values encoded in LLMs represent particular populations, but characterizes default values of LLMs in terms of worldview rather than demographic attributes. These correlations weaken in the PRIMEX PERSONA setting, where the additional user data allows the LLMs to align to users with diverse Primals.

5.2 Primals Prediction

If a user representation encodes worldview, we should be able to recover a user's Primals from thier representation. Table 6 shows the performance of predicting Primal scores from various inputs, measured in mean squared error across test users. Here, the model is prompted with a persona description and tries to predict the user's responses to the PI-18. Scores for each Primal are computed from these synthesized responses and compared with the user's actual scores. The TRAINED predictor is a version of GPT-4o trained on the PRIMEX training data. It takes as input a user's demographics, seed opinions, and explanations and predicts the answer to each PI-18 item independently.

The results in Table 6 show varying degrees of success at recovering user Primals for the zero-shot DEMO & OPINIONS and + *Explanations* settings depending on the Primal being predicted. However, there is an strong improvement in predicting the top-level *Good* and secondary *Safe* scores using the TRAINED model; this suggests that it is possible for a model to learn to predict some aspects of a user's worldview from opinion and explanation data. It would be worth investigating if Primals can be approximated from other sources of user data.

6 Measuring Explanation Helpfulness

Explanations have been shown to improve the predictive accuracy of PA-LMs; do some explanations help these models generalize better than others? To study this, we develop a measure of explanation *helpfulness*, or its utility to an LM for predicting a variety of opinions.

Let s^u denote a seed question and answer pair for user u, and let e^u_s be the explanation given by the user for their answer. Let $T^u = \{(q_j, a^u_j)\}$ be the user specific test set, where a^u_j is the user's response for test question q_j. The helpfulness e^u_s is defined as the difference in probability assigned to user answers by an LM when it is conditioned on e^u_s versus not, averaged across T^u. Formally:

$$\mathcal{M}(e^u_s) = \sum_{(q_j, a^u_j) \in T^u} \mathcal{P}(a^u_j | q_j, U + e^u_s) - \mathcal{P}(a^u_j | q_j, U)$$

Here, U is a user representation consisting of demographic information plus the single seed opinion s^u explained by e^u_s. We model $\mathcal{P}$ using the Mistral-7B-Instruct model (Jiang et al., 2023) which we prompt with the user representation as well as the test question and answer choices. The answer choices are enumerated with letters; $\mathcal{P}$ is restricted to the letters corresponding to valid answer choices and renormalized. We consider $\mathcal{P}(a^u_j|\cdot)$ to represent the probability assigned by the language model to the user's true choice. $\mathcal{M}(e^u_s)$ then represents the change in probability of the true user answers

	Length	1st Person	All Pronouns
Most Helpful	309.60	0.58	1.26
Least Helpful	145.34	0.59	1.32

Table 7: Lexical characteristics of the most and least helpful explanations

	Seed Question	Test Questions
Most Helpful	0.560	0.183
Least Helpful	0.447	0.140

Table 8: Explanation similarity with seed and test pairs.

under the model when provided with the extra information in the user's explanation, averaged over the test questions. $\mathcal{M}$ can be and often is negative, as some explanations provide information which causes the language model to move probability mass away from the user's answers.

More and Less Helpful Explanations We compute $\mathcal{M}$ for every explained user opinion; altogether a total of 7965. The scores on this dataset range from -0.155 to 0.109. The least helpful explanation is for a "Prefer not to answer" response to CHURCH ECON - "I think that this is not for religious causes but they can help". The most helpful explanation is for a "Yes" answer to GOVT RETIREMENT which begins – "Universal Basic Income or Guaranteed Basic Income should be implemented immediately, as should Universal Healthcare…".

Certain questions in PRIMEX seem to elicit more helpful explanations. Comparing the aggregate helpfulness of explanations, we find that the question eliciting the most helpful explanations is GOVT RETIREMENT; least helpful explanations are in response to ORGANIC FOODS. The helpfulness of the explanations for these questions differs significantly from the aggregate helpfulness of all explanations with a effect size of $d = 0.280$ and $d = 0.297$ respectively. The helpfulness of explanations for GOVT RETIREMENT is significantly different from that of explanations for ORGANIC FOODS with effect size $d = 0.579$. These results illustrate the importance of crafting explanation elicitation materials when collecting such data.

Quantitative Characterization To characterize the textual difference between the most and least explanations, we aggregate the 50 most and least helpful explanations for each question. Table 7 shows the average length as well as average counts of pronouns per sentence for the explanations in each category. The strongest signal here is the difference in length between the most and least helpful explanations, with the most helpful explanations averaging twice as long as the least. Longer explanations may include more overlapping information with test questions, improving their helpfulness.

To test this, we measure the semantic similarity of explanations e_s^u from each category with the seed opinion s^u they explain, as well as their average similarity with the user's test set T^u. Similarity is computed using embeddings from the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). Results in Table 8 show that the most helpful explanations are more similar to both their seed questions and test set. If we trust that respondents' explanations are relevant (see qualitative analysis below), this indicates that off-the-shelf models may not be able to generalize well from relevant but semantically dissimilar explanations.

Finally, we consider the in-group similarity by taking the average similarity of all explanations for the same question within the best and worst groups. Better explanations are more similar to each other than are worse explanations ($d = 0.80$). This indicates that models can only generalize from a small part of the semantic space of possible explanations.

Qualitative Analysis We manually examine 25 samples of the most and least helpful explanations. Our analysis reveals that most explanations are relevant to the opinion question they explain (23 best, 21 worst). Both the best and worst explanation groups contain a substantial number of *ambivalent* explanations, which describe possible reasons for taking different sides on the opinion issue (10 best, 8 worst), and so it seems that ambivalence is not strongly connected to explanation helpfulness. A major difference between the groups arises when considering *vacant* responses which only restate the provided opinion (1 best, 6 worst).

7 Conclusion

We introduce PRIMEX, a novel dataset of opinion question responses, explanations from respondents, and their answers to the Primal World Beliefs survey. We provide new insights into the relationships between personal opinions and worldview, and conduct detailed analysis of the utility of user beliefs in PA-LMs. The analyses described here are only some of what is possible with PRIMEX. We encourage its continued study in the NLP and psychological research communities.

8 Limitations

The participant pool for PRIMEX was restricted to English-speaking US. residents. We faced challenges collecting data from all demographic groups either equally or in proportion to that group's portion of the US. population. As a result, PRIMEX under-represents "Spanish, Hispanic, or Latino" respondents and "Male" respondents. Due to the cost of collecting survey data, the number of participants in PRIMEX is relatively small for the purposes of training NLP systems. The online format of this survey may have posed additional problems for people with less technological familiarity. Particularly, if a respondent did not have access to text-to-speech on their device they would have had to type out their explanation answers, a burden for those with weaker typing skills. This could have resulted in suboptimal collection of their explanations.

This work uses GPT-4o (gpt-4o-2024-11-20) accessed through the OpenAI API. This models is subject to a proprietary license which may change. The specific model may not be available indefinitely which impacts the reproducibility of the results reported in this paper. We also use Mistral 7B Instruct (v0.3), which is subject to the Apache 2.0 license.

9 Ethical Considerations

The intention of PRIMEX is to provide researchers from the psychological and NLP research science communities a rich source of data for analysis of opinion, explanation, and worldview. Our data contains subjective opinions from respondents which ma be offensive to some people. Our data was collected under the guidance of an ethics review board to ensure participant safety.

We study the impact of richer persona information for prompting LMs on the assumption that better user representations will enable more positive user experiences. Language models and especially PA-LMs have been shown to exhibit unfair biases (Gupta et al., 2024). We believe that richer user representations can counteract these biases by encouraging models to consider the individuality of each user rather than resorting to coarse generalizations.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Preprint*, arXiv:2306.11247.

A. T. Beck. 1976. *Cognitive therapy and the emotional disorders*. International Universities Press.

Jeremy Clifton. 2021. Primals inventories administration instructions$_{j}une2022$.

Jeremy D. W. Clifton. 2018. Primal world beliefs. https://www.myprimals.com/. Accessed: 2025-01-27.

Jeremy D. W. Clifton, Joshua D. Baker, Crystal L. Park, David B. Yaden, Alicia B. W. Clifton, Paolo Terni, Jessica L. Miller, Guang Zeng, Salvatore Giorgi, H. Andrew Schwartz, and Martin E. P. Seligman. 2019. Primal world beliefs. *Psychological Assessment*, 31(1):82–99.

Jeremy D. W. Clifton and David Bryce Yaden. 2021. Brief measures of the four highest-order primal world beliefs. *Psychological assessment*.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547, Abu Dhabi, UAE. Association for Computational Linguistics.

Carol S. Dweck, Chi yue Chiu, and Ying yi Hong. 1995. Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, 6:267–285.

David C. Funder and Daniel J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2:156 – 168.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Gilles E. Gignac and Eva T. Szodorai. 2016. Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102:74–78.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *Preprint*, arXiv:2311.04892.

Stefan G. Hofmann, Anu Asnaani, Imke J. J. Vonk, Alice T. Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, 36:427–440.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

EunJeong Hwang, Vered Shwartz, Dan Gutfreund, and Veronika Thost. 2024. A graph per persona: Reasoning about subjective natural language descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1928–1942, Bangkok, Thailand. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving llm personas via rationalization with psychological scaffolds. *Preprint*, arXiv:2504.17993.

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

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Conference on Empirical Methods in Natural Language Processing*.

Vera U. Ludwig, Damien L. Crone, Jeremy D. W. Clifton, Reb W Rebele, Jordyn Schor, and Michael Louis Platt. 2022. Resilience of primal world beliefs to the initial shock of the covid-19 pandemic. *Journal of Personality*.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. *Proceedings of the Tenth ACM Conference on Learning @ Scale*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. *Preprint*, arXiv:2407.06576.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Lan-

ders, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. *Preprint*, arXiv:2208.04024.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Pew Research Center. 2014. The american trends panel. https://www.pewresearch.org/the-american-trends-panel/. Accessed: 2025-01-27.

Pew Research Center. 2018a. American trends panel wave 34. https://www.pewresearch.org/dataset/american-trends-panel-wave-34. Accessed: 2024-09-15.

Pew Research Center. 2018b. American trends panel wave 41. https://www.pewresearch.org/dataset/american-trends-panel-wave-41. Accessed: 2024-09-15.

Pew Research Center. 2019. American trends panel wave 54. https://www.pewresearch.org/dataset/american-trends-panel-wave-54. Accessed: 2024-09-15.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Stanford Center on Poverty and Inequality. 2021. American v oices project. https://inequality.stanford.edu/avp/methodology. Accessed: 2025-02-14.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *ArXiv*, abs/2306.01693.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A Data Details

A.1 Demographics

Demographic distribution is shown in Table 9

A.2 Pew Survey Questions

All questions are taken from Pew Survey website. An option "Prefer not to answer" was included for all multiple choice questions to meet internal review requirements. Slight formatting changes compared to Pew presentation to accommodate our survey software.

	US Region
South	275
West	269
Midwest	189
Northeast	152
	Age
30 to 49	236
50 to 64	226
65 or older	220
18 to 29	203
	Gender
Female	508
Male	356
Non-binary	18
Prefer not to say	3
	Education
High school	323
Undergraduate degree (Bachelor's)	277
Graduate degree (Master's)	138
Associate's degree	134
Other responses	13
	Race
White or Caucasian	574
Black or African American	126
Asian	81
Spanish, Hispanic, or Latino	65
Other responses	39

Table 9: Demographic distribution of PRIMEX

Name	Question Text	Options
		Wave 34
†Medical Costs	Which of these statements comes closer to your point of view, even if neither is exactly right?	a. Medical treatments these days often create as many problems as they solve b. Medical treatments these days are worth the costs because they allow people to live longer and better quality lives
†Animal Research	All in all, do you favor or oppose the use of animals in scientific research?	a. Oppose b. Favor
†Organic Foods	Do you think organic fruits and vegetables are generally...	a. Worse for one's health than conventionally grown foods b. Neither better nor worse for one's health than conventionally grown foods c. Better for one's health than conventionally grown foods
Gene Risks	Thinking about what you have heard or read, how well do you think medical researchers understand the health risks and benefits of changing a baby's genetic characteristics?	a. Not well at all b. Not too well c. Fairly well d. Very well
Gene Disease	Do you think changing a baby's genetic characteristics to treat a serious disease or condition the baby would have at birth is an appropriate use of medical technology?	a. Taking medical technology too far b. An appropriate use of medical technology
Meat Hormone	How much health risk, if any, does eating meat from animals that have been given antibiotics or hormones have for the average person over the course of their lifetime?	a. No health risk at all b. Not too much health risk c. Some health risk d. A great deal of health risk
New Treatments	Thinking about medical treatments these days, how much of a problem, if at all, is the following: New treatments are made available before we fully understand how they affect people's health	a. Not a problem b. A small problem c. A big problem
Science Funding	Which statement comes closer to your view, even if neither is exactly right?	a. Private investment will ensure that enough scientific progress is made, even without government investment b. Government investment in research is ESSENTIAL for scientific progress
Food Additives	Which of these statements comes closer to your view, even if neither is exactly right?	a. The average person is exposed to additives in the food they eat every day but they eat such a small amount that this does not pose a serious health risk b. The average person is exposed to additives in the food they eat every day, which pose a serious risk to their health
Evolution	Thinking about the development of human life on Earth: Which statement comes closest to your view?	a. Humans have evolved over time due to processes that were guided or allowed by God or a higher power b. Humans have existed in their present form since the beginning of time c. Humans have evolved over time due to processes such as natural selection; God or a higher power had no role in this process
		Wave 54
†Govt Retirement	Do you think adequate income in retirement is something the federal government has a responsibility to provide for all Americans?	a. No, not the responsibility of the federal government to provide b. Yes, a responsibility of the federal government to provide for all Americans
†Church Econ	How much responsibility, if any, should churches and other religious organizations have in reducing economic inequality in our country	a. None b. Only a little c. Some d. A lot
†Immigrant Econ	How much, if at all, do you think the growing number of legal immigrants working in the US contributes to economic inequality in this country?	a. Contributes not at all b. Contributes not too much c. Contributes a fair amount d. Contributes a great deal

Gas Prices	How much, if at all, do you think gas prices are contributing to your opinion about how the economy is doing?	a. Not at all b. Not too much c. A fair amount d. A great deal
House Prices	How much, if at all, do you think real estate values are contributing to your opinion about how the economy is doing?	a. Not at all b. Not too much c. A fair amount d. A great deal
Job Confidence	How much, if at all, do you think the availability of jobs in your area are contributing to your opinion about how the economy is doing?	a. Not at all b. Not too much c. A fair amount d. A great deal
Race Econ	How much, if at all, do you think discrimination against racial and ethnic minorities contributes to economic inequality in this country?	a. Contributes not at all b. Contributes not too much c. Contributes a fair amount d. Contributes a great deal
Corporate Econ	How much, if at all, do you think regulation of major corporations contributes to economic inequality in this country?	a. Contributes not at all b. Contributes not too much c. Contributes a fair amount d. Contributes a great deal
Benefits Econ	How much, if at all, do you think the following proposals would do to reduce economic inequality in the U.S.? Expanding government benefits for the poor	a. Nothing at all b. Not too much c. A fair amount d. A great deal
Antitrust Econ	How much, if at all, do you think the following proposals would do to reduce economic inequality in the U.S.? Breaking up large corporations	a. Nothing at all b. Not too much c. A fair amount d. A great deal
		Wave 41
†Population	In 2050, do you think population growth in the US will be a ...	a. Not a problem b. Minor problem c. Major problem
†Energy Crisis	How likely do you think it is that the following will happen in the next 30 years? The world will face a major energy crisis	a. Will definitely not happen b. Will probably not happen c. Will probably happen d. Will definitely happen
†Public Ed.	Thinking ahead 30 years, which do you think is more likely to happen in the U.S.?	a. The public education system will get worse b. The public education system will improve
Child Standard	Thinking ahead 30 years, which do you think is more likely to happen in the U.S.?	a. Children will have a worse standard of living b. Children will have a better standard of living
China vs US	How likely do you think it is that the following will happen in the next 30 years? China will overtake the US as the world's main superpower	a. Will definitely not happen b. Will probably not happen c. Will probably happen d. Will definitely happen
Race Relations	Thinking ahead 30 years, which do you think is more likely to happen in the U.S.?	a. Race relations will improve b. Race relations will get worse
Climate Change	Thinking about the future of our country, how worried are you, if at all, about climate change?	a. Not worried at all b. Not too worried c. Fairly worried d. Very worried
Alzheimer Cure	How likely do you think it is that the following will happen in the next 30 years? There will be a cure for Alzheimer's disease	a. Will definitely not happen b. Will probably not happen c. Will probably happen d. Will definitely happen
Military Cost	If you were deciding what the federal government should do to improve the quality of life for future generations, what priority would you give to reducing military spending?	a. Should not be done b. A lower priority c. An important, but not a top priority d. A top priority
Religion	Thinking ahead 30 years, which do you think is more likely to happen in the U.S.?	a. Religion will be about as important as it is now b. Religion will become less important

Table 10: Public opinion survey questions in PRIMEX. For questions marked with †we elicit explanations of participant responses.

How much responsibility, if any, should churches and other religious organizations have in reducing economic inequality in our country?
User 11: Only a little — In my opinion, church members should address social and economic issues only as expressions of their faith. Other than that, there should be strict separation of church and state.
User 12: None — Many religions teach the importance of charity, but in a country with no state official religion, we should not depend on, or demand, some or all religious organizations be part of a nationwide effort to redistribute wealth. Extremely large organizations, such as megachurches, should become taxable to an extent, but as a society, we should use our framework of government to reduce economic inequality, not attempt to create a system based on vastly different religions working together.
User 13: A lot — Churches are social groups. We should support ourselves as a community and churches are part of the community . . . .

In 2050, do you think population growth in the US will be a ...
User 234: Major Problem — We are growing really fast. I know all over the world and the US, we don't have enough for people. That includes basics and I know growth is just going up still.
User 235: Not a problem — It will be opposite, population will be less than they expect given no one is having babies these days
User 236: Minor Problem — I don't expect population growth to be unmanageable if we do a good job managing it. The US is a huge and vast country with more than enough room and resources to handle population growth, especially if it lets more cities outside of the main urban areas grow. . . .

Do you think organic fruits and vegetables are generally ...
User 58: Better for one's health than conventionally grown foods — Fruits n vegetables are way more better than supplements and medicines
User 59: Neither better nor worse for one's health than conventionally grown foods — I do not ever consume organic products because there are no legal or official standards for organic farming practices, although I do not believe those foods are necessarily worse than non-organic foods. I simply think those foods are marked up unnecesarrily to take advantage of a recent trend, even though those products are often inferior (smaller, less hearty, more prone to disease, etc).
User 60: Better for one's health than conventionally grown foods — I've read a lot of research on the dangers of consuming pesticides. Pesticides are toxic to humans as well as other important life like pollinating insects. . . .

Figure 3: Examples of opinions with user explanations.

A.3 Explanation Examples

Figure 3 provides additional examples of explanations from PRIMEX.

B Full Correlations

This section contains all correlation results between Primals and survey responses.

	Good				
Question	n	r	p of r	ρ	p of ρ
Medical Costs	793	0.243	4.13e-12	–	–
Animal Research	756	0.103	4.52e-03	–	–
Organic Foods	809	0.077	2.75e-02	0.08	2.21e-02
Gene Risks	778	0.099	5.50e-03	0.098	6.40e-03
Gene Disease	757	0.066	6.82e-02	–	–
Meat Hormone	800	0.018	6.13e-01	0.022	5.30e-01
New Treatments	805	-0.025	4.86e-01	-0.011	7.58e-01
Science Funding	788	0.061	8.58e-02	–	–
Food Additives	801	-0.044	2.14e-01	–	–
Evolution	809	0.043	2.22e-01	–	–
Govt Retirement	809	-0.087	1.33e-02	–	–
Church Econ	791	0.107	2.59e-03	0.113	1.51e-03
Immigrant Econ	783	-0.085	1.78e-02	-0.094	8.35e-03
Gas Prices	804	0.046	1.96e-01	0.055	1.21e-01
House Prices	806	-0.018	6.08e-01	-0.004	9.10e-01
Job Confidence	806	-0.012	7.27e-01	0.002	9.44e-01
Race Econ	798	-0.024	4.97e-01	-0.034	3.38e-01
Corporate Econ	792	-0.101	4.62e-03	-0.102	3.90e-03
Benefits Econ	803	-0.036	3.13e-01	-0.048	1.72e-01
Antitrust Econ	800	-0.14	6.71e-05	-0.151	1.75e-05
Population	795	-0.156	1.02e-05	-0.175	6.65e-07
Energy Crisis	784	-0.065	6.98e-02	-0.063	7.81e-02
Public Ed.	765	0.293	1.34e-16	–	–
Child Standard	762	0.341	2.91e-22	–	–
China vs US	789	-0.184	1.84e-07	-0.178	4.68e-07
Race Relations	773	-0.272	1.27e-14	–	–
Climate Change	807	-0.019	5.98e-01	-0.007	8.33e-01
Alzheimer Cure	802	0.187	9.24e-08	0.186	1.16e-07
Military Cost	805	-0.041	2.43e-01	-0.041	2.46e-01
Religion	792	-0.13	2.54e-04	–	–

Table 11: Correlations of Pew Opinion responses with Good primal. For questions with only 2 answer options, Spearman rank correlation is unavailable.

	Safe				
Question	n	r	p of r	ρ	p of ρ
Medical Costs	831	0.25	2.42e-13	–	–
Animal Research	794	0.144	4.37e-05	–	–
Organic Foods	852	0.028	4.12e-01	0.028	4.07e-01
Gene Risks	813	0.074	3.56e-02	0.061	8.00e-02
Gene Disease	790	0.034	3.37e-01	–	–
Meat Hormone	841	-0.133	1.13e-04	-0.13	1.62e-04
New Treatments	847	-0.124	3.03e-04	-0.115	7.58e-04
Science Funding	827	0.054	1.18e-01	–	–
Food Additives	842	-0.149	1.41e-05	–	–
Evolution	853	0.02	5.60e-01	–	–
Govt Retirement	853	-0.117	5.96e-04	–	–
Church Econ	827	0.053	1.28e-01	0.053	1.28e-01
Immigrant Econ	824	-0.078	2.50e-02	-0.066	5.70e-02
Gas Prices	848	-0.03	3.84e-01	-0.028	4.20e-01
House Prices	845	-0.1	3.72e-03	-0.096	5.05e-03
Job Confidence	849	-0.058	9.16e-02	-0.065	5.91e-02
Race Econ	840	-0.071	3.95e-02	-0.079	2.21e-02
Corporate Econ	830	-0.168	1.14e-06	-0.174	4.45e-07
Benefits Econ	842	-0.084	1.44e-02	-0.108	1.64e-03
Antitrust Econ	838	-0.171	6.33e-07	-0.183	8.90e-08
Population	833	-0.183	1.12e-07	-0.195	1.37e-08
Energy Crisis	822	-0.136	8.74e-05	-0.149	1.78e-05
Public Ed.	800	0.29	6.40e-17	–	–
Child Standard	788	0.325	7.86e-21	–	–
China vs US	820	-0.182	1.58e-07	-0.177	3.48e-07
Race Relations	808	-0.285	1.33e-16	–	–
Climate Change	850	-0.041	2.27e-01	-0.04	2.45e-01
Alzheimer Cure	840	0.07	4.26e-02	0.071	3.85e-02
Military Cost	846	-0.033	3.43e-01	-0.031	3.75e-01
Religion	829	-0.088	1.11e-02	–	–

Table 12: Correlations of Pew Opinion responses with Safe primal. For questions with only 2 answer options, Spearman rank correlation is unavailable.

	Enticing				
Question	n	r	p of r	ρ	p of ρ
Medical Costs	835	0.198	7.94e-09	–	–
Animal Research	796	0.074	3.78e-02	–	–
Organic Foods	854	0.094	5.99e-03	0.094	5.85e-03
Gene Risks	820	0.054	1.21e-01	0.054	1.25e-01
Gene Disease	794	0.093	8.58e-03	–	–
Meat Hormone	843	0.071	3.81e-02	0.069	4.59e-02
New Treatments	851	0.032	3.49e-01	0.044	2.02e-01
Science Funding	830	0.077	2.60e-02	–	–
Food Additives	845	0.047	1.73e-01	–	–
Evolution	856	-0.021	5.44e-01	–	–
Govt Retirement	856	-0.074	2.94e-02	–	–
Church Econ	834	0.084	1.50e-02	0.081	1.97e-02
Immigrant Econ	825	-0.098	4.87e-03	-0.111	1.42e-03
Gas Prices	851	0.036	2.89e-01	0.029	3.91e-01
House Prices	851	0.037	2.85e-01	0.045	1.94e-01
Job Confidence	853	0.049	1.53e-01	0.062	7.07e-02
Race Econ	844	0.019	5.90e-01	0.013	6.97e-01
Corporate Econ	835	-0.053	1.28e-01	-0.054	1.20e-01
Benefits Econ	848	-0.012	7.22e-01	-0.005	8.75e-01
Antitrust Econ	844	-0.073	3.30e-02	-0.07	4.13e-02
Population	838	-0.074	3.17e-02	-0.082	1.75e-02
Energy Crisis	826	0.008	8.11e-01	0.024	4.84e-01
Public Ed.	797	0.198	1.67e-08	–	–
Child Standard	791	0.235	1.99e-11	–	–
China vs US	823	-0.135	1.03e-04	-0.118	6.71e-04
Race Relations	812	-0.215	6.47e-10	–	–
Climate Change	854	0.045	1.86e-01	0.068	4.77e-02
Alzheimer Cure	844	0.167	1.14e-06	0.169	8.48e-07
Military Cost	851	-0.041	2.38e-01	-0.042	2.20e-01
Religion	833	-0.076	2.80e-02	–	–

Table 13: Correlations of Pew Opinion responses with Enticing primal. For questions with only 2 answer options, Spearman rank correlation is unavailable.

	Alive				
Question	n	r	p of r	ρ	p of ρ
Medical Costs	789	-0.018	6.16e-01	–	–
Animal Research	753	-0.05	1.70e-01	–	–
Organic Foods	805	0.022	5.26e-01	0.028	4.22e-01
Gene Risks	773	0.048	1.79e-01	0.042	2.43e-01
Gene Disease	753	-0.105	3.80e-03	–	–
Meat Hormone	793	0.129	2.75e-04	0.143	5.51e-05
New Treatments	801	0.129	2.40e-04	0.132	1.72e-04
Science Funding	784	-0.155	1.28e-05	–	–
Food Additives	796	0.018	6.11e-01	–	–
Evolution	805	0.32	1.44e-20	–	–
Govt Retirement	805	-0.099	5.09e-03	–	–
Church Econ	788	0.073	4.07e-02	0.078	2.93e-02
Immigrant Econ	779	0.108	2.44e-03	0.11	2.18e-03
Gas Prices	801	0.308	4.22e-19	0.301	3.15e-18
House Prices	799	0.117	9.39e-04	0.123	4.86e-04
Job Confidence	800	0.025	4.76e-01	0.036	3.08e-01
Race Econ	794	-0.144	4.50e-05	-0.134	1.50e-04
Corporate Econ	786	0.06	9.10e-02	0.049	1.71e-01
Benefits Econ	794	-0.114	1.28e-03	-0.113	1.40e-03
Antitrust Econ	795	-0.1	4.63e-03	-0.097	6.10e-03
Population	789	-0.042	2.40e-01	-0.053	1.35e-01
Energy Crisis	782	-0.01	7.81e-01	-0.005	8.79e-01
Public Ed.	762	0.137	1.51e-04	–	–
Child Standard	756	0.141	9.75e-05	–	–
China vs US	775	-0.104	3.61e-03	-0.112	1.85e-03
Race Relations	773	-0.079	2.88e-02	–	–
Climate Change	803	-0.197	1.70e-08	-0.192	4.02e-08
Alzheimer Cure	795	0.151	1.96e-05	0.145	4.21e-05
Military Cost	799	-0.113	1.42e-03	-0.096	6.56e-03
Religion	782	-0.15	2.56e-05	–	–

Table 14: Correlations of Pew Opinion responses with Alive primal. For questions with only 2 answer options, Spearman rank correlation is unavailable.

Code	Statement
ed1	In life, there's way more beauty than ugliness.
am1	It often feels like events are happening in order to help me in some way.
sd1	I tend to see the world as pretty safe.
am2	What happens in the world is meant to happen.
ed2x	While some things are worth checking out or exploring further, most things probably aren't worth the effort.
ed3x	Most things in life are kind of boring.
ed4	The world is an abundant place with tons and tons to offer.
ed5	No matter where we are or what the topic might be, the world is fascinating.
ed6x	The world is a somewhat dull place where plenty of things are not that interesting.
sd2x	On the whole, the world is a dangerous place.
sd3x	Instead of being cooperative, the world is a cut-throat and competitive place.
am3x	Events seem to lack any cosmic or bigger purpose.
sd4x	Most things have a habit of getting worse.
am4	The universe needs me for something important.
sd5	Most things in the world are good.
am5	Everything happens for a reason and on purpose.
sd6	Most things and situations are harmless and totally safe.
ed7	No matter where we are, incredible beauty is always around us.

Table 15: The 18 item Primal World Belief Inventory (PI-18). Response options are on a six point 0-5 scale: (5) Strongly agree, (4) Agree, (3) Slightly Agree, (2) Slightly Disagree, (1) Disagree, and (0) Strongly disagree. Items whose codes include "x" are reverse scored.

Primal	Equation
Good	$(sd1 + sd2x + sd3x + sd4x + sd5 + sd6 + ed1 + ed2x + ed3x + ed4 + ed5 + ed6x + ed7 + am1 + am4)/15$
Safe	$(sd1 + sd2x + sd3x + sd4x + sd5 + sd6)/6$
Enticing	$(ed1 + ed2x + ed3x + ed4 + ed5 + ed6x + ed7)/7$
Alive	$(am1 + am2 + am3x + am4 + am5)/5$

Table 16: Equations for calculating high-level Primal scores from survey responses.

C PI-18 Primal World Belief Inventory

The PI-18 consists of 18 multiple choice questions which assess worldview. Table 15 shows the exact statements used in this survey. Participants rate their agreement with each statement on a scale from "Strongly Agree" to "Strongly Disagree". The responses are converted to high-level scores for each Primal using the equations in Table 16.

D Model prompts and instructions

Figure 4 shows the general prompt template for the opinion prediction experiments. For generating synthetic explanations from the FINETUNED model, the prompt in Figure 5 is used.

E Model Configuration

Prediction experiments were conducted via API calls. Each model processed somewhere in the range of 500-750M tokens for these experiments. Hyper-parameters "temperature= 0" and "max_tokens= 1" were used in the final results. We explored other max_tokens settings $\in \{1, 2, 10\}$ to ensure this parameter wasn't impacting model outputs.

The FINETUNED opinion predictor was GPT-4o finetuned via OpenAI API on 174,399 tokens. The TRAINED Primals predictor was fineuned on 24,920,748 tokens.

Explanations helpfulness calculations were done with Mistral 7B Instruct v0.3 on 8 A100 40GB GPU and took less than 24 hours.

System Message:
You are a person described as follows::

<demographic information>

You have the following opinions:

1. Question: <question>
Answer choices: <answer choices>
Your answer: <user selected response>
Reason: <explanation>
2. ...

User Message
Based on your demographic and opinion information above, which answer would you select for the following question?

Question: <question>
Answer choices: <answer choices>
Your answer:

Figure 4: General prompt template for opinion prediction. Settings without demographics, opinions, or reasons omit these fields.

System Message:
You are a person described as follows:
<demographic information>

User Message
You hold the following opinion:
Question: <question>
Answer choices: <answer choices>
Your answer: <user selected response>

Please explain your answer to the question above. Provide 2-4 sentences which could help someone understand why you have the opinion you have. Your explanation can draw on any aspect of your personal history, social life, experiences, thoughts, feelings, beliefs, or values. Please don't simply repeat your opinion; try to explain *why* you have that opinion.

Your explanation:

Figure 5: Prompt for generating FINETUNED explanations, which is the same as the prompt given to survey respondents.