PRIMEX: A Dataset of Worldview, Opinion, and Explanation

Rik Koncel-Kedziorski[1], **Brihi Joshi**[2], **Tim Paek**[1]
[1]Apple, [2]University of Southern California
Correspondence: rikka@apple.com

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 858 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, opening 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 can significantly 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 persistent behavioral patterns and correlates with personality, well-being, political, religious, and demographic variables (Clifton et al., 2019). As such, worldview can be seen as a powerful, compact, and predictive model of the individual's belief system.

Simultaneously, advancements in language modeling have made it possible to incorporate higher-level user beliefs into predictive models (Sun et al., 2025). A better understanding of individual belief systems can improve personalization of language models (LMs), 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 by 858 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). This enables 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 from the respondent of why they hold a particular opinion. These explanations often draw on the respondent's higher-level beliefs about the world. We show that explanations improve a PA-LM's ability to 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 writ-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24748–24773

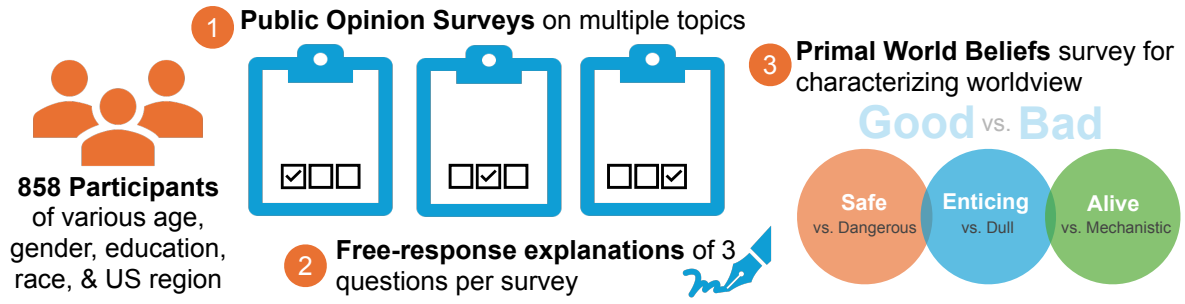

Figure 1: Overview of the PRIMEX data. We collect three types of responses from a diverse pool of participants: Opinions from 3 Pew Research surveys; explanations for 3 opinions per survey; and Primal World Belief survey of participant worldview.

ten explanations. Including worldview and explanations in user representations for PA-LMs can improve opinion prediction, and we develop an analysis of the utility of a given explanation to PA-LMs. 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 (Clifton et al., 2019). 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 explana-

[1] https://github.com/apple/ml-primex

tion 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 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 personas. 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., 2022), including preference modeling and opinion prediction (Sun et al., 2025; Do et al., 2025; Joshi et al., 2025). We analyze human-written explanations and model-generated explanations for prediction, and identify characteristics of explanations which models can use to improve prediction.

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., 2025). These works make use of personas to adapt language models to an individual user's preferences. PRIMEX provides rich user data and can serve as training and testing data for personalization methods.

Psychology and LMs Popular psychological theories in NLP literature include Big 5 personality traits (Goldberg, 1993) and Schwartz theory of basic values (Schwartz, 1992). Recent works have analyzed the presence of personality traits in machine generated text (Hilliard et al., 2024), and have even administered psychometric evaluations to various LMs to identify their default synthetic personality traits (Karra et al., 2022; Serapio-García et al., 2023). Some works have incorporated psychological traits into PA-LMs (Moon et al., 2024; Park et al., 2024; Li et al., 2025). Primal World Beliefs have been shown to explain broad aspects of personality (Clifton et al., 2019), but remain underexplored in this space. To our knowledge, ours is the first work to collect worldview data for the purposes of analyzing and developing PA-LMs.

3 Dataset Construction

The goal of PRIMEX 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 personas 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 persona development. Lastly, we consider a source of user information which has not yet been brought to bear on PA-LM development: individual worldview. Hence, we collect responses to the 18 question Primal World Beliefs survey to capture an individual's worldview. In total, PRIMEX includes responses from 858 individuals, similar in size to recent works in LM personalization (Park et al. (2024), $N = 1052$) and personality psychology (Ludwig et al. (2022), $N = 529$)

3.1 Survey Questions

Our data consists of three types of questions: public opinion, free-response explanations, and Primals. We use multiple-choice 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 use 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 10 questions that meet two characteristics: they ask about personal opinions rather than biological

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.

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 11 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 (see Figure 7 for the full instructions). 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 (2022) (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 858 US 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] A condensed version of the demographic distribution of this data is provided in Table 10 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 a more nationally representative demographic distribution if necessary.

Primal	*N*=	Avg	Std	min	max	US Avg.
Good	785	3.09	0.68	0.53	4.93	2.9
Safe	827	2.50	0.90	0.00	5.00	2.5
Alive	780	2.66	1.10	0.00	5.00	2.8
Enticing	830	3.73	0.75	0.57	5.00	3.7

Table 1: Primal Belief scores of our respondents.

Each participant was offered payment above local minimum 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 design and collection process was guided by an institutional review board to ensure compliance with regulatory and ethical standards. After the survey, participants had a chance to review their responses and given the opportunity to opt-out and have their data removed from the dataset. We removed all of the data from participants who had opted-out.

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

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

	Good		Safe		Alive		Enticing	
	↑	↓	↑	↓	↑	↓	↑	↓
+	0.29	0.09	0.23	0.08	0.04	0.01	0.10	0.04
−	0.25	0.56	0.19	0.43	0.62	0.71	0.31	0.48
∅	0.46	0.35	0.58	0.49	0.35	0.28	0.60	0.48

Table 2: Primal Beliefs in the explanations from users with highest ↑ and lowest ↓ scores. Rows correspond to evidence of high (+) or low (−) belief in the explanation, or no evidence (∅)

Primal	↑	↓	Δ
Good	171	183	6.84%
*Safe**	154	187	21.43%
*Alive**	160	254	58.91%
*Enticing**	189	167	-11.92%

Table 3: Length of explanations from users with highest ↑ and lowest ↓ Primal scores. Rows marked * indicate significant differences at $p < 0.01$.

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. From each participant we collected 30 opinion question responses, 9 explanations, answers to the Primal World Beliefs survey comprising 18 scalar ranked questions, and 11 demographic attributes.

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-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). We report Spearman's rank coefficient ρ for ordinal (non-binary) and Pearson's correlation coefficient r for all correlations. 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 Pearson 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 ($r = 0.338$ and $r = 0.326$ 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 ($r = 0.322$). Interestingly, we also observe relatively large correlations between how much gas prices impact one's view of the economy (GAS PRICES) and *Alive* scores ($r = 0.310$).

Overall, we observe small or stronger correlations ($r > 0.1$) between all Primals and at least some questions from each topic. 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

Is a person's worldview evidenced in their explanations for their opinions? We test this using an LLM (zero-shot GPT-4o) to judge whether each explanation indicates the user has high (+) or low (−) Primal, or if there is no evidence (∅). The judge is provided a definition of the Primal taken from (Clifton, 2018), shown in Figure 8. We analyze the explanations from the 50 respondents with the highest (↑) and lowest (↓) scores for each Primal (900 explanations per Primal in total). Percentages of explanations with evidence for each group are shown in the columns of Table 2. This analysis shows that

indications of a respondent's Primals can be identified in their explanations. Across the + and − rows we see that indications of high (or low) primal belief appear more often in explanations from users with high (or low) scores for each Primal. Note that we intend this as a preliminary analysis. While the LLM-as-judge paradigm has been validated in some domains (Zheng et al., 2023), additional work is needed to validate its use for predicting Primals. We explore the problem of predicting Primals by training models on PRIMEX in Section 5.2.

As a more stable analysis, Table 3 shows variations in length of explanations based on the respondent's Primal scores. The Δ column indicates the change moving from the high to low group for each Primal. We observe significant differences in average explanation length; respondents with lower *Alive* scores give over 50% longer responses than their high *Alive* counterparts.

An interpretable vocabulary analysis of explanations from these groups is complicated by the underlying topicality of the explanations, but predictive results of LMs conditioned on explanations in Section 5.1 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 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 in text format and instructions to predict the user's response to unseen test questions one at a time. Test questions are provided in multiple-choice format. The general prompt format is shown in Figure 6. We report model accuracy, which is computed by comparing the user's true response token to a single token generated by the model.[3] See Appendix D for additional details.

User Representation	GPT-4o	Mistral
All Topics		
DEMOGRAPHICS	42.36	42.40
DEMO & OPINIONS	45.15	44.37
+ Primals	46.21	41.92
+ Explanations	48.12	46.20
+ Generated Explanations	46.30	45.13
PRIMEX PERSONA	48.31	45.44
Cross Topic		
DEMOGRAPHICS	39.24	39.72
DEMO & OPINIONS	39.82	40.12
+ Primals	40.42	41.03
+ Explanations	40.21	40.40
+ Generated Explanations	40.02	40.37
PRIMEX PERSONA	40.68	40.78

Table 4: Predicting user opinions from PRIMEX. Underlined results are significantly different from DEMO & OPINIONS from the same model at $p < 0.05$.

Primal	Correlated	Uncorrelated
Good	51.07	40.05
Safe	48.15	51.54
Alive	54.88	41.07
Enticing	53.27	44.89

Table 5: Accuracy of model predictions for most and least correlated questions.

We study two forms of this task: in the *All Topics* setting, test questions are drawn from all waves of the ATP survey. For user representations including seed opinions, we use the 9 explained opinions and test on the remaining 21 for each user. In the *Cross Topic* setting, seed opinions include 3 opinions with explanations and 7 additional opinions from Wave 34 and the 20 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 use a train/test split with 430 train and 428 test users. We explore the capability of 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

[3]We confirmed that the generated token always indicates an answer choice rather than reasoning or refusal.

data from PRIMEX: 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); an example of this information is shown in Figure 8. For short context models (Mistral) we provide the question/response pairs from the user's PI-18 (see Table 16). 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).

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 4 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. Mistral can use worldview to make significant improvements in prediction, but GPT-4o requires both worldview and explanations to achieve statistically significant improvements.

Primals and Accuracy Continuing the analysis of Section 4, we compare the model prediction accuracy of five opinion questions from the *All Topics* test set which are most and least correlated with different Primals. Table 5 shows the average accuracy of GPT-4o with DEMO & OPINIONS + *Explanations* user representation for these questions. We see that for *Good*, *Alive*, and *Enticing* Primals the model is much better at predicting opinions for strongly correlated questions, but this is not true for *Safe world* belief. These trends hold for other user representations. 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 6 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, which also demonstrates better accuracy for opinion prediction. This indicates the additional user data in PRIMEX helps the LLMs to overcome their default values and personalize more effectively.

	Good	**Safe**	**Alive**	**Enticing**
GPT-4o				
DEMOGRAPHICS	0.15	0.13	-0.04	0.10
PRIMEX PERSONA	0.06	0.05	-0.07	0.03
Mistral				
DEMOGRAPHICS	0.15	0.11	-0.12	0.13
PRIMEX PERSONA	0.10	0.09	-0.06	0.06

Table 6: Correlation of model accuracy and user Primal score. Underlined values are significant at $p < 0.05$

	Good	**Safe**	**Alive**	**Enticing**
DEMO & OPIN.	0.56	1.13	1.46	0.61
+ *Explanations*	0.55	1.40	1.07	0.62
TRAINED	0.46	0.69	1.22	0.65

Table 7: Predicting a user's Primals (MSE). Underlined results are significantly better than both other methods at $p < 0.05$

5.2 Primals Prediction

In Section 4 we found evidence of Primals in written explanations. Can we predict a person's Primals based on their demographics, opinions, and explanations? Table 7 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 7 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 a significant 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 Utility of Explanations

Explanations have been shown to improve the predictive accuracy of PA-LMs overall, but do some explanations provide more information about a user's beliefs than others? What characteristics of an explanation allow a PA-LM to learn generalizable information about the user?

To study this, we define a utility function $\mathcal{M}$ which expresses how useful a user's explanation of a seed opinion is to the PA-LM when predicting test opinions. We model this as the expected log-likelihood gain in the true class when predicting a user's opinions conditioned on the explanation compared to a baseline.[4] For the baseline, we condition the LM on a simple user representation consisting of demographic information plus a single unexplained seed opinion. See Appendix H for the full details of how we compute $\mathcal{M}$. Note that $\mathcal{M}$ may be negative, as some explanations cause the language model to move probability mass away from the user's test responses.

We compute $\mathcal{M}$ for explanations of answered seed opinion questions. The distribution of scores is shown in Figure 3. Scores on this dataset range from -0.503 to 0.317 with a mean of 0.024. This is a non-normal distribution, so we report the interquartile range (IQR) of scores as a measure of variability. The IQR for utility of PRIMEX explanations is 0.054, indicating that most scores are clustered close to the mean.

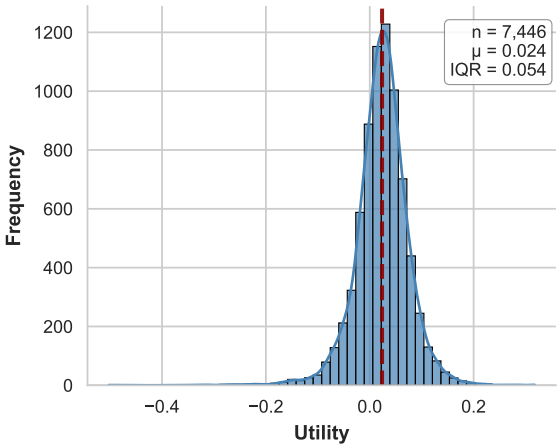

Figure 3: Distribution of Utility Scores for Explanations in PRIMEX.

	Length	Good	Safe	Alive	Enticing
High	333	0.58	0.44	0.59	0.36
Low	144	0.61	0.53	0.69	0.53

Table 8: Length and evidence of Primals in high and low utility explanations.

Quantitative Characterization To characterize the textual difference between the highest and lowest utility explanations, we group the explanations with utility values beyond $1.5\times$ IQR from the first and third quartiles. In Table 8 we repeat the LLM-as-judge experiment introduced in Section 4. Here we report the percentage of explanations which indicate any valence of the Primal (i.e. $+$ and $-$ versus $\varnothing$). Notably, the low utility explanations are marked as more indicative of a user's worldview across Primals. Developing PA-LMs that can utilize these worldview-revealing texts may be a fruitful avenue for future work.

We also see that high utility explanations are on average over twice as long as low utility ones. Longer explanations may include more overlapping information with test questions and provide additional signal to the PA-LM for predicting user responses to these. To test this, we measure the average cosine similarity of user explanation with the test questions and user responses.[5] Table 9 shows that the highest utility explanations are more similar to both the seed questions they explain and the user's test set.

[4]This is equivalent to the reduction in cross-entropy loss.

[5]Texts are embedded using the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019).

	Seed Question	Test Questions
High	0.609	0.182
Low	0.484	0.151

Table 9: High and low utility explanation similarity with seed and test question/response pairs.

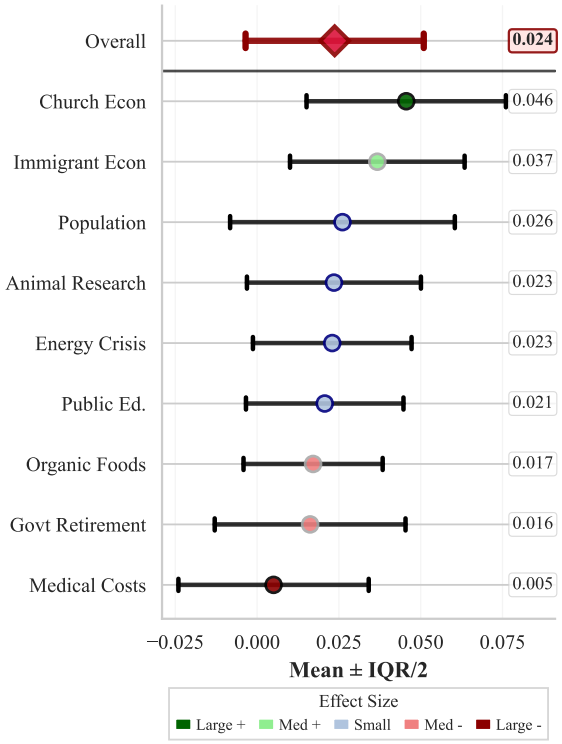

Figure 4: Utility Scores by Question. This figure is best viewed in color.

Effect of Question on Explanation Utility Certain questions in PRIMEX seem to elicit higher utility explanations. A plot of explanation utility by question is shown in Figure 4. We compare the distribution of utility scores for each question against the average for all questions using Welch's T-test and report effect size in the figure. The question eliciting the highest utility explanations is CHURCH ECON; the lowest utility explanations are in response to MEDICAL COSTS. On average, each question yields explanations with positive utility, but there is large variance within groups. Based on these results, future work should consider the expected utility of the collected explanations when crafting elicitation materials.

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 a detailed analysis of the utility of user beliefs in PA-LMs. The analyses described here are only some of what is possible with PRIMEX.

PRIMEX is the first dataset with opinions, explanations, and worldview data from the same individuals and can facilitate the development of methods for realistic LLM-based simulation. Importantly, PRIMEX allows for the modeling of user behavior based on individual beliefs rather than demographic attributes. Primal World Beliefs are validated constructs with strong explanatory power for a range of human behaviors. The quality of a simulated persona with a given set of Primals can be assessed based on the similarity of its opinions and explanations to comparable respondents in PRIMEX. Likewise, the quality of a simulated explanations can be improved by considering the PRIMEX data. In personality psychology, PRIMEX can be combined with existing data for continued development of the theory of Primal World Beliefs; it can also enable further analysis, including connections between demographics and worldview, qualitative coding of explanations, or the study of groups based on shared opinions.

In general, we hope that PRIMEX will inspire future research modeling not only user behavior but underlying user beliefs, as we observe that these correlate with the behaviors studied here. We encourage the NLP and psychological research communities to make use of this resource.

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 United States' 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 model 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 may be offensive to some people. Our data was collected under the guidance of an internal review board to ensure participant safety. Participants gave their informed consent before participating in the survey. Participants had the option to refuse to answer any question. After completing the survey, participants had the option to withdraw their responses from the data release.

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 Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.

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

Jeremy Clifton. 2022. Primals inventories administration instructions.

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. In *The Twelfth International Conference on Learning Representations*.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting personality traits in large language models. *ArXiv*, abs/2402.08341.

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 language model personas via rationalization with psychological scaffolds. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Saketh Reddy Karra, Son Nguyen, and Theja Tulabandhula. 2022. Ai personification: Estimating the personality of language models. *ArXiv*, abs/2204.12000.

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. In *Advances in Neural Information Processing Systems*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping llm personalities through training on human-grounded data. In *ACL*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

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 Var-

avva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, 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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

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, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

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? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

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 Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296, Abu Dhabi, UAE. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

	US Region
South	269
West	263
Midwest	183
Northeast	143
	Age
30 to 49	231
50 to 64	214
65 or older	212
18 to 29	200
	Gender
Female	489
Male	348
Non-binary	18
Prefer not to say	3
	Education
High school	311
Undergraduate degree (Bachelor's)	271
Graduate or Professional degree	142
Associate's degree	130
Other responses	4
	Race
White or Caucasian	554
Black or African American	122
Asian	79
Spanish, Hispanic, or Latino	65
Other responses	38

Table 10: Demographic distribution of PRIMEX.

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. In *International Conference on Learning Representations (ICLR)*, Vienna, Austria.

A Data Details

A.1 Demographics

Demographic distribution is shown in Table 10

A.2 Pew Survey Questions

Table 11 lists the public opinion questions in PRIMEX. 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. We made slight formatting changes compared to Pew presentation to accommodate our survey software.

Table 11: Public opinion survey questions in PRIMEX. For questions marked with † we elicit explanations of participant responses.

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

Continued on next page

Table 11 – continued from previous page

Name	Question Text	Options
†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

Continued on next page

Table 11 – continued from previous page

Name	Question Text	Options
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

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 5: Examples of opinions with user explanations.

A.3 Explanation Examples

Figure 5 provides additional examples of explanations from PRIMEX.

B Full Correlations

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

Question	Good				
	n	r	p of r	ρ	p of ρ
Medical Costs	770	0.24	1.39e-11	–	–
Animal Research	732	0.108	3.42e-03	–	–
Organic Foods	785	0.076	3.36e-02	0.078	2.85e-02
Gene Risks	754	0.103	4.57e-03	0.103	4.75e-03
Gene Disease	736	0.082	2.54e-02	–	–
Meat Hormone	776	0.024	5.12e-01	0.027	4.50e-01
New Treatments	781	-0.032	3.73e-01	-0.018	6.18e-01
Science Funding	764	0.058	1.10e-01	–	–
Food Additives	778	-0.042	2.48e-01	–	–
Evolution	785	0.037	3.01e-01	–	–
Govt Retirement	785	-0.098	5.83e-03	–	–
Church Econ	768	0.116	1.30e-03	0.121	7.58e-04
Immigrant Econ	759	-0.095	9.18e-03	-0.102	5.11e-03
Gas Prices	780	0.043	2.35e-01	0.051	1.52e-01
House Prices	782	-0.028	4.41e-01	-0.009	7.92e-01
Job Confidence	782	-0.01	7.88e-01	0.006	8.64e-01
Race Econ	775	-0.023	5.31e-01	-0.031	3.90e-01
Corporate Econ	768	-0.122	7.21e-04	-0.119	9.55e-04
Benefits Econ	779	-0.039	2.80e-01	-0.048	1.79e-01
Antitrust Econ	777	-0.144	5.85e-05	-0.154	1.66e-05
Population	773	-0.169	2.30e-06	-0.184	2.58e-07
Energy Crisis	762	-0.063	8.42e-02	-0.062	8.79e-02
Public Ed.	741	0.286	2.05e-15	–	–
Child Standard	741	0.338	3.27e-21	–	–
China vs US	766	-0.178	7.42e-07	-0.173	1.45e-06
Race Relations	751	-0.274	1.96e-14	–	–
Climate Change	783	-0.02	5.70e-01	-0.008	8.26e-01
Alzheimer Cure	779	0.183	2.63e-07	0.181	3.91e-07
Military Cost	781	-0.047	1.88e-01	-0.046	2.03e-01
Religion	770	-0.125	5.35e-04	–	–

Table 12: Correlations of Pew Opinion responses with Good primal. Table shows number of responses *n*, Pearson correlation coefficient r with p-value p of r, Spearman rank correlation coefficient ρ with p-value p of ρ. For questions with only 2 answer options, Spearman rank correlation is unavailable.

Question	Safe n	r	p of r	ρ	p of ρ
Medical Costs	806	0.259	8.62e-14	–	–
Animal Research	768	0.144	6.43e-05	–	–
Organic Foods	827	0.029	4.00e-01	0.031	3.80e-01
Gene Risks	787	0.087	1.42e-02	0.076	3.26e-02
Gene Disease	767	0.053	1.40e-01	–	–
Meat Hormone	815	-0.128	2.59e-04	-0.125	3.64e-04
New Treatments	822	-0.132	1.41e-04	-0.122	4.40e-04
Science Funding	801	0.051	1.52e-01	–	–
Food Additives	817	-0.144	3.68e-05	–	–
Evolution	827	0.012	7.24e-01	–	–
Govt Retirement	827	-0.129	2.02e-04	–	–
Church Econ	802	0.058	1.02e-01	0.059	9.63e-02
Immigrant Econ	798	-0.088	1.25e-02	-0.075	3.43e-02
Gas Prices	822	-0.036	3.02e-01	-0.031	3.70e-01
House Prices	820	-0.108	1.96e-03	-0.099	4.43e-03
Job Confidence	823	-0.056	1.10e-01	-0.057	9.95e-02
Race Econ	815	-0.069	5.02e-02	-0.075	3.16e-02
Corporate Econ	805	-0.184	1.35e-07	-0.188	7.75e-08
Benefits Econ	816	-0.082	1.94e-02	-0.104	2.89e-03
Antitrust Econ	814	-0.166	2.06e-06	-0.177	3.73e-07
Population	810	-0.193	3.16e-08	-0.206	3.29e-09
Energy Crisis	798	-0.132	1.83e-04	-0.146	3.63e-05
Public Ed.	774	0.284	7.10e-16	–	–
Child Standard	766	0.326	1.84e-20	–	–
China vs US	796	-0.177	4.89e-07	-0.174	8.41e-07
Race Relations	784	-0.289	1.49e-16	–	–
Climate Change	824	-0.04	2.53e-01	-0.04	2.57e-01
Alzheimer Cure	816	0.067	5.72e-02	0.069	4.80e-02
Military Cost	820	-0.037	2.92e-01	-0.033	3.46e-01
Religion	805	-0.09	1.10e-02	–	–

Table 13: Correlations of Pew Opinion responses with Safe primal. Table shows number of responses n, Pearson correlation coefficient r with p-value p of r, Spearman rank correlation coefficient ρ with p-value p of ρ. 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	810	0.193	3.26e-08	–	–
Animal Research	770	0.074	4.07e-02	–	–
Organic Foods	829	0.089	1.06e-02	0.088	1.10e-02
Gene Risks	794	0.051	1.48e-01	0.051	1.53e-01
Gene Disease	771	0.109	2.43e-03	–	–
Meat Hormone	817	0.078	2.54e-02	0.071	4.25e-02
New Treatments	826	0.028	4.20e-01	0.04	2.50e-01
Science Funding	804	0.078	2.75e-02	–	–
Food Additives	820	0.054	1.25e-01	–	–
Evolution	830	-0.03	3.87e-01	–	–
Govt Retirement	830	-0.08	2.10e-02	–	–
Church Econ	809	0.088	1.24e-02	0.087	1.30e-02
Immigrant Econ	799	-0.113	1.37e-03	-0.123	4.72e-04
Gas Prices	825	0.03	3.94e-01	0.022	5.33e-01
House Prices	826	0.027	4.45e-01	0.037	2.93e-01
Job Confidence	827	0.049	1.60e-01	0.06	8.25e-02
Race Econ	819	0.026	4.59e-01	0.018	6.04e-01
Corporate Econ	810	-0.071	4.31e-02	-0.065	6.24e-02
Benefits Econ	822	-0.013	7.06e-01	-0.006	8.59e-01
Antitrust Econ	820	-0.073	3.58e-02	-0.074	3.44e-02
Population	815	-0.083	1.75e-02	-0.087	1.29e-02
Energy Crisis	802	0.019	6.00e-01	0.031	3.75e-01
Public Ed.	771	0.19	1.10e-07	–	–
Child Standard	769	0.225	2.71e-10	–	–
China vs US	799	-0.124	4.34e-04	-0.111	1.64e-03
Race Relations	788	-0.216	9.41e-10	–	–
Climate Change	828	0.05	1.47e-01	0.074	3.43e-02
Alzheimer Cure	820	0.149	1.73e-05	0.156	6.98e-06
Military Cost	825	-0.048	1.67e-01	-0.048	1.68e-01
Religion	809	-0.076	3.00e-02	–	–

Table 14: Correlations of Pew Opinion responses with Enticing primal. Table shows number of responses *n*, Pearson correlation coefficient r with p-value p of r, Spearman rank correlation coefficient ρ with p-value p of ρ. For questions with only 2 answer options, Spearman rank correlation is unavailable.

Question	Alive				
	n	r	p of r	ρ	p of ρ
Medical Costs	765	-0.022	5.50e-01	–	–
Animal Research	728	-0.044	2.31e-01	–	–
Organic Foods	780	0.018	6.22e-01	0.023	5.28e-01
Gene Risks	748	0.047	1.96e-01	0.042	2.56e-01
Gene Disease	731	-0.101	6.35e-03	–	–
Meat Hormone	768	0.13	3.05e-04	0.144	5.97e-05
New Treatments	776	0.129	3.04e-04	0.131	2.62e-04
Science Funding	759	-0.157	1.46e-05	–	–
Food Additives	772	0.016	6.50e-01	–	–
Evolution	780	0.322	2.71e-20	–	–
Govt Retirement	780	-0.104	3.60e-03	–	–
Church Econ	764	0.074	4.14e-02	0.08	2.72e-02
Immigrant Econ	754	0.108	3.01e-03	0.109	2.76e-03
Gas Prices	776	0.31	9.08e-19	0.301	1.04e-17
House Prices	774	0.115	1.33e-03	0.123	6.24e-04
Job Confidence	775	0.026	4.65e-01	0.038	2.96e-01
Race Econ	770	-0.145	5.54e-05	-0.134	2.03e-04
Corporate Econ	762	0.049	1.75e-01	0.036	3.28e-01
Benefits Econ	770	-0.117	1.14e-03	-0.115	1.34e-03
Antitrust Econ	771	-0.106	3.08e-03	-0.103	4.29e-03
Population	766	-0.043	2.34e-01	-0.053	1.45e-01
Energy Crisis	759	-0.012	7.43e-01	-0.008	8.35e-01
Public Ed.	737	0.136	2.12e-04	–	–
Child Standard	735	0.138	1.78e-04	–	–
China vs US	752	-0.1	6.21e-03	-0.108	2.99e-03
Race Relations	750	-0.075	3.88e-02	–	–
Climate Change	778	-0.202	1.23e-08	-0.198	2.63e-08
Alzheimer Cure	771	0.15	2.95e-05	0.143	6.56e-05
Military Cost	774	-0.113	1.60e-03	-0.098	6.62e-03
Religion	759	-0.145	6.02e-05	–	–

Table 15: Correlations of Pew Opinion responses with Alive primal. Table shows number of responses *n*, Pearson correlation coefficient r with p-value p of r, Spearman rank correlation coefficient ρ with p-value p of ρ. 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 16: 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 17: 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 16 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 17.

D Opinion prediction task details

Let S and T denote the set of seed and test multiple-choice opinion questions respectively. Response options are denoted with letters, and all questions include a "Prefer not to answer" option as required by our internal review board.

In the *All Topics* setting $|\ S\ | = 9$ and $|\ T\ | = 21$ across 3 topics. In the *Cross Topic* setting $|\ S\ | = 10$ from Wave 34 and $|\ T\ | = 20$ from Waves 41 and 45. For a given user u, let $S^u = \{(a_i^u, e_i^u)|q_i \in S\}$ denote the user's responses a and explanations e for questions in S, and $T^u = \{a_j|q_j \in T\}$ be the user's responses for questions in T.

We predict unseen user responses for opinion questions by prompting and off-the-shelf language model $LM_{\mathcal{T}}$ with task-specific instructions $\mathcal{T}$. Prediction is conditioned on a test question $q_j \in T$ and user representation U.

$$\hat{a_j} = LM_{\mathcal{T}}(q_j, U) \tag{1}$$

We restrict $\hat{a}$ to a single output token. Manual observations and experiments with longer outputs have shown that the models studied here produce a valid answer choice. The model output is correct when $\hat{a_j} = a_j^u$. Accuracy is the percentage of correct answer across all T^u for users in the test split.

E Model prompts and instructions

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 6: 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 7: Prompt for generating FINETUNED explanations, which is the same as the prompt given to survey respondents.

Figure 6 shows the general prompt template for the opinion prediction experiments. For generating synthetic explanations from the FINETUNED model, the prompt in Figure 7 is used.

F 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 20 most *helpful* explanations for each question from the training set are used as training data for this model (see Section 6 for details on how helpfulness is calculated). The TRAINED Primals predictor was finetuned 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.

G Contextual information for Primal scores

Example Primal Analysis Representation

Good World Belief (vs. bad)
This primal concerns arguably the single most basic question anyone could ever ask: Is the world a good place?
What Does Good Predict?
When people are in places they see as really bad (or really good) we would expect them to behave in certain ways. When people see the world as a really bad place (or a really good place), we see a similar pattern. Good is highly related to optimism, life satisfaction, curiosity, agreeableness, and virtually all aspects of psychological well-being—even subjectively reported physical health. Likewise, scoring low on Good (i.e., seeing the world as a bad place) is a risk factor for depression and many other mental health issues. Though we don't yet know where primals come from, we can definitely rule one thing out: seeing the world as Good is not a product of our material circumstances. Income and Good are related, but the relationship is very small. This small relationship may also be explained by Good causing someone to experience more professional success, instead of professional success causing one to see the world as Good.
You have a Good World Belief score of 3.0, which corresponds to seeing an OK world The average American falls at 2.9. Those in this range are the most flexible in how they interpret ambiguity and see both pessimism and optimism as rational and appropriate – different strokes. However, even within this range, where you fall may matter a lot. Because scores may vary considerably on Safe, Enticing, and Alive, different profiles suggest quite different approaches to life. For example, high Enticing and low Safe could make for "neurotic explorers;" low Enticing high Safe could make for "satisfied provincials," though both may score equally on Good.

Safe World Belief (vs. dangerous)
This primal concerns to what degree we see the world as generally a safe place where threats are often overblown or a dangerous place where major threats really are everywhere. In their foundational 2019 research article, Clifton and the research team describe Safe in the following way: Those low on Safe see a Hobbesian world defined by misery, decay, scarcity, brutality, and dangers of all sorts. Base rates for hazards—from germs to terrorism to getting stabbed in the back—are generally higher. In response to chronic external threats, they remain on high alert, often viewing the non-vigilant as irresponsible. Those high on Safe see a world of cooperation, comfort, stability, and few threats. To them, things are safe until proven otherwise, vigilance appears neurotic, risk is not that risky, and, in general, people should calm down.
What Does Safe Predict?
When people are in places they see as really dangerous, we would expect them to behave in certain ways, such as being irritable, tense, and so forth. When it comes to seeing the world as safe or dangerous, we see a similar pattern. Safe is highly related to a host of personality and well-being variables. These include agreeableness, extraversion, interpersonal trust, and life satisfaction. Likewise, scoring low on Safe is related to neuroticism, depression, and loneliness. Though we don't yet know where primals come from, we are starting to realize that Safe scores are not simply determined by our environment. They appear more like a lens through which we view the world than a direct result of our experiences. For example, though people living in more affluent households are objectively safer than those living in poorer households because they can absorb various catastrophes, large medical bills, and so forth, there is no relationship between household income and Safe. Also, even though a case could be made that the world is a safer place for men than women, gender is completely unrelated to Safe. Family members, friends, and colleagues, who score differently on Safe may have different baseline assumptions for vague situations. This may lead to a variety of misunderstandings and differing opinions. For example, those scoring high on Safe may find it easier to relax, be less suspicious of others, and assume nothing that bad will happen. Those scoring low on Safe may find it harder to relax, be more suspicious of others, and assume the worst.
You have a Safe World Belief score of 2.0, which corresponds to seeing a somewhat dangerous world The average American scores a 2.5. Those in the middle are least likely to see themselves as holding a belief about the world. They can relate to everyone but also may be a bit baffled by the behavior of those on both extremes. Even within this large group (50% of the population), how one scores may still matter a great deal.

Figure 8: An example user representation with contextualizing information about Primal Beliefs and the user's score for each belief.

Figure 8 shows the contextual information included for GPT-4o models in the *+Primals* and PRIMEX PERSONA user representations. This information is taken from (Clifton et al., 2019).

Alive World Belief (vs. mechanistic)
This primal is about how much we see the world as full of intention and purpose that is interacting with us and wants our help. In their foundational 2019 paper, Clifton and his research team describe Alive like this: Those low on Alive inhabit inanimate, mechanical worlds without awareness or intent. Since the universe never sends messages, it makes no sense to try to hear any. Those high on Alive sense that everything happens for a purpose and are thus sensitive to those purposes. To them, life is a relationship with an active universe that animates events, works via synchronicity, communicates, and wants help on important tasks.
What Does Alive Predict?
When people are in places thought to be overseen by something alive, we would expect them to behave in certain ways, such as trying to determine its thoughts and desires, seeking to interact with it, being responsive to perceived requests, and so forth. When it comes to seeing the world as alive, we see a similar pattern. Alive is related to a variety of personality and well-being variables, for example, finding meaning in life, optimism, and life satisfaction. Also closely related to Alive are a tendency to have transcendent experiences, being a spiritual person, being more religious, and even being more extroverted. Likewise, scoring low on Alive is related to depression and ingratitude. However, compared to Safe and Enticing, Alive is less related to personality and well-being variables while still playing an important role. For example, gratitude is very strongly related to Enticing. However, there is also a small but important unique relationship to Alive. In other words, to score high in gratitude, it's critical to see the world as full of things to be grateful for, but also helps to have someone to be grateful to. Interestingly, though Alive is strongly related to being religious, lots of non-religious people, even atheists, see the world as Alive. Within religious groups, there is also a lot of variation. Family members, friends, and colleagues, who score differently on Alive may have different baseline assumptions for vague situations. This may lead to a variety of misunderstandings and differing opinions. For example, those scoring high on Alive may be more likely to read intentionality, meaning, and purpose into events that others perceive are not there. But they will also enjoy some mental health benefits that elude those who insist on seeing the world, and events that happen within the world, as mechanical and indifferent to them.
You have a Alive World Belief score of 2.8, which corresponds to seeing a not quite alive world The average American scores a 2.8. This means they tend to see the world as slightly Alive, but not by much. Those in the middle are least likely to see themselves as holding a belief about the world at all, but they do. They can relate to everyone but may be a bit baffled by the behavior of those on both extremes. Even within this large group (50% of the population), where one scores may still matter a great deal. These folks see the world as often animate, but often not. The universe has desires and intentions, but they don't necessarily influence lots of events.

Enticing World Belief (vs. dull)
This primal concerns how much we see the world as generally full of interesting and beautiful things or dull and ugly things. In their foundational 2019 paper, Clifton and his research team describe Enticing like this: Those low on Enticing inhabit dull and ugly worlds where exploration offers low return on investment. They know real treasure—truly beautiful and fascinating things—is rare and treasure-hunting appropriate only when it's a sure bet. Those high on Enticing inhabit an irresistibly fascinating reality. They know treasure is around every corner, in every person, under every rock, and beauty permeates all. Thus, life is a gift, boredom a misinformed lifestyle choice, and exploration and appreciation is the only rational way to live.
What Does Enticing Predict?
When people are in places they see as really enticing, we would expect them to behave in certain ways, such as being grateful, curious, open to experience, extroverted, and so forth. When it comes to seeing the world as enticing, we see a similar pattern. Enticing is highly related to a host of personality and well-being variables. These include curiosity, gratitude, openness to experience, agreeableness, extroversion, engagement with life, positive emotion, meaning in life, a sense of accomplishment, and even having friends. Likewise, scoring low on Enticing is related to psychopathy, depression, and not trying. Identifying Enticing is the single most important discovery of primals research so far. While Safe has been studied to a small degree, Enticing is both wholly unexplored and of immense practical importance. Enticing may play an enormous role in one's own well-being and mental health; it's just as important as Safe. Often Safe and Enticing correlate (increase and decrease together). After all, they are the two main reasons for seeing the world as Good (Alive is icing on the cake). However, sometimes Safe and Enticing work very differently. For example, it appears almost impossible to be a curious person or a grateful person without scoring high on Enticing. But, when it comes to those traits, Safe scores don't matter much. Family members, friends, and colleagues, who score differently on Enticing may have different baseline assumptions for vague situations. This may lead to a variety of misunderstandings and differing opinions. For example, those scoring high on Enticing may be able to spot amazing opportunities, but they may also be sucked in by bad ones. Those scoring low on Enticing may be less likely to be fooled or waste their time, but they may be more likely to miss opportunities and have life pass them by.
You have a Enticing World Belief score of 4.0, which corresponds to seeing a somewhat interesting world The average American scores a 3.7 and sees the world as somewhat interesting. Those in the middle are least likely to see themselves as holding a belief about the world. They can relate to everyone but may be a bit baffled by the behavior of those on both extremes. Even within this large group (50% of the population), where one scores may still matter a great deal.

Figure 8: An example user representation with contextualizing information about Primal Beliefs and the user's score for each belief (cont.)

H Utility Computation

Let S, S^u, T, and T^u be as defined in Appendix D. We compute the utility of an explanation $e_i^u \in S^u$ to a LM by first establishing a baseline performance for the LM given a simple user representation $U = D + (q_i, a_i^u)$ for user demographics D. Here (q_i, a_i^u) are the seed question and user response which the user has explained with explanation e_i^u. The baseline for computation is the expected log-probability assigned to T^u by the LM conditioned on U:

$$\mathbb{E}_{q_j \in T, a_j^u \in T^u} \mathcal{P}(a_j^u | q_j, U)$$

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 $\mathcal{P}(a_j^u|\cdot)$ to represent the probability assigned by the language model to the user's true choice.

The utility of e_i^u is defined as the expected log-likelihood gain of a_i^u when P is conditioned on e_i^u:

$$\begin{aligned}\mathcal{M}(e_i^u) = \mathbb{E}_{q_j \in T, a_j^u \in T^u} [&\log(\mathcal{P}(a_j^u | q_j, U + e_i^u)) \\ &- \log(\mathcal{P}(a_j^u | q_j, U))]\end{aligned}$$

$\mathcal{M}(e_i^u)$ then represents the change in probability of the true user answers 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.