Improving Language Model Personas via Rationalization with Psychological Scaffolds

Brihi Joshi[♥] **Xiang Ren**[♥] **Swabha Swayamdipta**[♥] **Rik Koncel-Kedziorski**[♦] **Tim Paek**[♦]
[♥]University of Southern California [♦]Apple
brihijos@usc.edu

Abstract

Language models prompted with a user description or *persona* have been used to predict the user's preferences and opinions. However, existing approaches to building personas mostly rely on a user's demographic attributes and/or prior judgments, but not on any underlying *reasoning* behind a user's judgments. We introduce PB&J (**P**sychology of **B**ehavior and **J**udgments), a framework that improves LM personas by incorporating potential rationales for *why* the user could have made a certain judgment. Our rationales are generated by a language model to explicitly reason about a user's behavior on the basis of their experiences, personality traits, or beliefs. Our method employs *psychological scaffolds*: structured frameworks such as the Big 5 Personality Traits or Primal World Beliefs to help ground the generated rationales in existing theories. Experiments on public opinion and movie preference prediction tasks demonstrate that language model personas augmented with PB&J rationales consistently outperform personas conditioned only on user demographics and / or judgments, including those that use a model's default chain-of-thought, which is not grounded in psychological theories. Additionally, our PB&J personas perform competitively with those using human-written rationales, suggesting the potential value of synthetic rationales guided by existing theories[1].

1 Introduction

Recent advances in language modeling for user *persona* (i.e, characteristics, preferences, and behavior of a user) offer straightforward yet powerful ways to predict user behavior and decision-making (D'Onofrio, 2020). Personas have enabled simulation capabilities in LMs, such as helping survey design by simulating a wide range of human responses (Argyle et al., 2023; Santurkar et al., 2023; Tjuatja et al., 2023), simulating communities to study behavior (Park et al., 2022; Zhou et al., 2024; Park et al., 2024), and producing diverse and large-scale synthetic data (Moon et al., 2024; Ge et al., 2024). Yet, simulated LM personas still struggle to align well with intended user behavior (Gupta et al., 2024; Liu et al., 2024).

A typical task for evaluating LM personas is opinion prediction: how accurately does the LM persona reflect the user's real opinion to questions such as "*For your job or career aspects of your life, would you say that you are where you expected to be at this point in your life?*" Fine-tuning LMs on user conversations and interaction history to create personas does not scale well due to the lack of adequate user-specific data (Mazaré et al., 2018; Madotto et al., 2019; Li et al., 2024). Zero-shot prompting approaches, where a persona is specified using demographic attributes in a system prompt, avoid these issues but often produce inconsistent responses and expose known biases in current LMs (Santurkar et al., 2023; Hu and Collier, 2024; Cheng et al., 2023a; Gupta et al., 2024). To improve zero-shot prompting, recent approaches incorporate user *judgments* (prior user responses and history) in the prompt (Hwang et al., 2023; Sun et al., 2024). For example, a user judgment like "*I'm not very focused on my professional life right now*" could inform their response to questions about their career expectations. While these methods provide additional context, they still fail to address a critical gap: the ability to *rationalize why* a user might have a specific judgment, an essential component to understand behavior. In our example, *one plausible rationale* for the user's judgment could be that they are prioritizing their family over their career at this time.

In this work, we hypothesize that incorporating *plausible post-hoc rationales for a user's judgment* could improve LM personalization, bridging the gap between surface-level personas and deeper,

[1]Code available at https://github.com/apple/ml-scaffold

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21747–21770

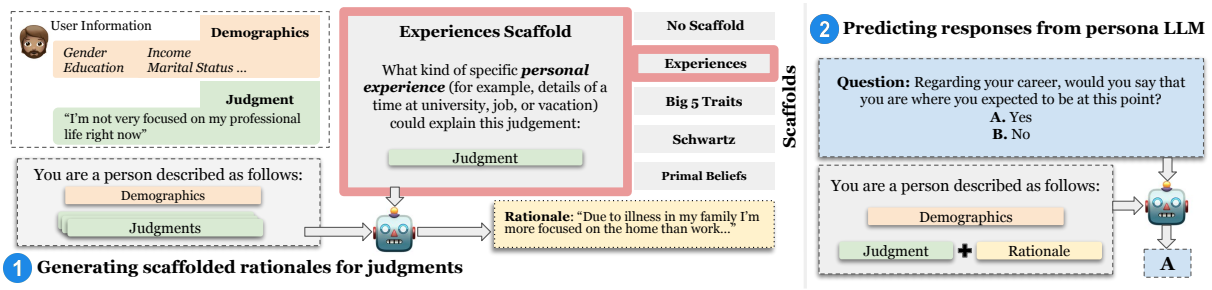

Figure 1: **Overview of the PB&J framework:** A base persona comprising user Demographics and Judgments is augmented with post-hoc rationales of various Scaffolds generated by an LM. The updated persona is integrated into the system prompt of an LM to align predictions with user behavior.

more nuanced simulated behavior. However, collecting such rationales is expensive and existing datasets for building personas only contain judgments (Kirk et al., 2024; Durmus et al., 2024), and in rare cases, user demographics (Santurkar et al., 2023; Harper and Konstan., 2015), but no rationales for judgments. To address this limitation, we introduce PB&J (**P**sychology of **B**ehavior and **J**udgments), a framework that uses post-hoc plausible, yet synthetic, LM-generated rationales for user judgments, to improve LM personas. The rationales for user judgments in our method stem from psychological theories, such as personality traits (e.g., Big 5 traits (Goldberg, 1993), Schwartz's value theory (Schwartz, 1992)) and belief systems (e.g., Primal World Beliefs (Clifton et al., 2019)); we call these *psychological scaffolds*. Each psychological scaffold offers additional context to produce one potential (albeit generated) path of reasoning to support a given user judgment. This is in contrast to chain-of-thought rationales (Wei et al., 2023) where no explicit psychological scaffold is employed; the LM relies on its own world knowledge to rationalize a test input prediction ad hoc.

It is important to note that while our synthetic rationales may not reflect the *true* reasoning behind a user's judgment, they may contain plausible markers of real behavioral reasoning due to the presence of such reasoning in the LM's training data (Binz and Schulz, 2023; Hagendorff, 2023; Balepur et al., 2025). Indeed, humans use everyday reasoning to rationalize behavior; this is known as *folk psychology* (Churchland and Haldane, 1988; Malle and Knobe, 1997; Malle, 2004).

Given a user's demographics and prior judgments, PB&J produces LM-generated rationales for user's judgments to construct a richer, more comprehensive user description for persona prompting. Our experiments show that when prompted with such richer descriptions, LMs can result in more accurate personas, as evaluated on a set of test instances. We evaluate PB&J on two tasks that represent different aspects of personalization: opinion prediction in OpinionQA (Santurkar et al., 2023) and movie preference prediction in MovieLens (Harper and Konstan., 2015). On both benchmarks, we show that LM personas prompted with PB&J rationales are significantly better at predicting a user's response to test instances than those prompted with user demographics and/or prior judgments alone, including those with additional chain-of-thought reasoning. Improvements from PB&J are seen across nearly all demographic categories, and PB&J can improve personas with or without demographic information. Finally, we present a pilot experiment where we collect human-written rationales, along with their demographics and judgments on a subset of OpinionQA; LM-rationales nearly match the performance of human-written rationales in PB&J, showing the efficacy of rationales generated with our approach.

At a high level, our work highlights the value of synthesizing data with the appropriate structural context to produce plausible rationalizations that improve the modeling of user opinions and preferences in capable language models.

2 Background

Humans commonly use explanations to rationalize each other's actions (Malle, 2004). We hypothesize that rationales are common enough in our discourse (e.g., narrative text of novels, biographies, social media) to be prevalent in training corpora for LMs to generate effective rationales in zero-shot settings (Binz and Schulz, 2023; Hagendorff, 2023). In order to better guide LMs, we employ various *psychological scaffolds*, which are templates based on

more structured theories of psychology, to rationalize user judgment.

2.1 Rationalization

Psychologists have long studied how people use everyday language to rationalize and anticipate behavioral patterns in social interactions, also known as folk psychology (Churchland and Haldane, 1988; Malle and Knobe, 1997; Malle, 2009; Gordon, 1986). Building on the example introduced in Section 1, suppose that we have to rationalize *why* a user believes that *they are not focused on their professional life and career at the moment* (See Figure 1). We could assume that this behavior arises due to some belief, such as *life is dull and uninteresting and chasing career progressions is a waste of time* (Dennett, 1981). Or we may offer life experiences that shape beliefs, such as *personal life issues that get prioritized over one's career at the moment* (Malle, 2004). For our purposes, we define **rationalization** as an attempt to understand the behavior of others using explanations based on assumed mental states.

Unlike prior work, which uses rationalization as reasoning chains to explain model decisions ad hoc in order to improve model task performance (Wei et al., 2023; Zelikman et al., 2022; Ramnath et al., 2024), or to enhance human utility (Joshi et al., 2023; Si et al., 2024; Chaleshtori et al., 2024), we focus on post-hoc rationalization of user judgments to improve LM personas for accurate user predictions and preferences. Formally, given a user persona $\mathcal{Q}$ and judgment j as input, we instruct an LM to generate "a reasonable explanation that the user would provide for holding that judgment." The *basic rationale* r_{basic} is defined as

$$r_{\text{basic}} = LM_{\mathcal{R}}(j, \mathcal{Q}), \quad (1)$$

where LM is an off-the-shelf pretrained language model instructed to elicit a rationale.

2.2 Psychological scaffolds

While Equation (1) helps us generate rationales from LMs, it is important to control *what* LMs generate as plausible rationales for human behavior. Can LMs generate more informed rationales based on well-established psychological theories? To find out, we employ *psychological scaffolds* to enhance the LMs' reasoning capabilities about user beliefs and preferences by providing coherent scaffolds as instructions to follow for reasoning about the underlying causes of observed behavior.

Scaffold Name	Description
NO SCAFFOLD	Free-form rationales without any structured guidance which rely on the model's internal knowledge of folk psychology.
EXPERIENCES (McAdams, 1993)	Rationales describing the life experiences and societal norms that would motivate a user's judgment, such as relationships, work, or community.
BIG 5 PERSONALITY TRAITS (Goldberg, 1993)	Rationales use personality traits such as openness, conscientiousness, extraversion, agreeableness, and neuroticism to explain judgments.
SCHWARTZ THEORY OF BASIC HUMAN VALUES (Schwartz, 1992)	Rationales that ground judgments in the user's relationship to stimulation, hedonism, self-direction, universalism, and security, among others.
PRIMAL WORLD BELIEFS (Clifton et al., 2019)	Rationales which use worldview — whether the world is good, safe, alive, or enticing — to explain user judgments.

Table 1: Summary of the scaffolds used in PB&J, highlighting their theoretical underpinnings and practical relevance.

The specific scaffolds we study in this work are shown in Table 1. In the context of our previous example, the EXPERIENCES scaffold might promote a rationale based on the bad experience of a family member who prioritized career goals over personal life (McAdams, 1993). Rationales generated from the PERSONALITY TRAITS scaffold may posit the user's openness to new experiences, including bold career choices (Goldberg, 1993; Schwartz, 1992). A rationale using the PRIMAL WORLD BELIEFS scaffold may surmise that the user sees the world as dull and uninteresting and thus not worth the effort of seeking professional success (Clifton et al., 2019). More details about the theories underlying each scaffold can be found in Appendix A. While the scaffolds that we select in this work are common psychological constructs used in prior work on personality and social intelligence (Zhou et al., 2024; Vu et al., 2022b; Moon et al., 2024), any scaffold that helps reason about a user's belief can be used. Our selected scaffolds offer diverse yet complementary ways to structure psychological rationales, anchoring rationales in personal experiences, personality traits, and belief systems.

To generate a rationale from an LM with a particular scaffold, we include additional scaffold information in the rationale generation instructions to arrive at scaffold-specific instructions ψ. Using $\mathcal{Q}$ and j as above, the *scaffold rationale* r_ψ is defined

OpinionQA
Question: How well, if at all, do the following words or phrases describe you? Interested in visiting other countries?
Options:
A. Describes me well
B. Does not describe me well
C. Refused to answer
User-selected Answer: A. Describes me well

MovieLens
Question: Out of 5, what would 'To Kill a Mockingbird (1962)' be rated?
Synopsis: In small-town Alabama in 1932, Atticus Finch (Gregory Peck) is a lawyer and a widower. He has two young children, Jem and Scout. Atticus Finch is currently defending Tom Robinson, a Black man accused of raping a white woman. Meanwhile, Jem and Scout are intrigued by their neighbors, the Radleys, in particular the mysterious, seldom-seen Boo Radley.
Directors: Robert Mulligan
Cast: Gregory Peck, John Megna, Frank Overton, Rosemary Murphy, Ruth White
User-selected Answer: I would rate 'To Kill a Mockingbird (1962)' a 2 out of 5.

Figure 2: **Example Task:** Shown here are example inputs and outputs for OpinionQA and MovieLens datasets respectively. Each of these instances have corresponding user selected answers.

as follows:

$$r_\psi = LM_\mathcal{R}(j, \mathcal{Q}, \psi) \tag{2}$$

The specific instructions used for each scaffold are given in Appendix B.

3 LM Personas using PB&J

Consider a task where an LM has to adopt the persona of a given user, and perform a task (e.g. answer opinion-based questions) as their proxy. In this section, we describe how we leverage LM rationalization to construct LM personas, which are used as prompts to generate personalized responses as a proxy for the user. We begin by setting up a basic persona description composed of user demographics and augmented with a set of user judgments, just as in previous work (Hwang et al., 2023). We then present PB&J, which enriches persona descriptions by providing additional context in the form of rationales, and demonstrate how PB&J is used to generate personalized responses as a proxy for the user.

3.1 Persona Building

Base Persona Setup. Following Hwang et al. (2023), we begin with a *base persona* $\mathcal{Q}_B$ for an arbitrary user consisting of demographic attributes $\mathcal{D}$ and a set of N *seed judgments* $\mathcal{J}$. Demographic attributes $\mathcal{D}$ capture the sociological characteristics and group identity traits of a user, such as age, gender, education, and race (Santurkar et al., 2023). While these attributes can provide useful information on user preferences and opinions, prior work has shown that relying solely on demographics often results in stereotypical and biased responses from LMs (Hwang et al., 2023; Cheng et al., 2023b; Gupta et al., 2024). To address these limitations, Hwang et al. (2023) augment demographic attributes with a set of seed judgments $\mathcal{J}$. Seed judgments often represent user-selected answers from a held out set of questions, or a prior interaction history which aims to provide additional context to enrich the persona description (Hwang et al., 2023; Do et al., 2024). In other tasks, a user's interaction history, such as the movies they watched beforehand along with their reviews, constitute the seed judgments. It is also possible to construct alternate base personas with just the demographic attributes $\mathcal{Q}_D$ or seed judgments $\mathcal{Q}_J$.

Post-hoc Rationales for Seed Judgments. In PB&J, we augment the description of the above persona $\mathcal{Q}_B$ with additional context in the form of post-hoc generated rationales r for each seed judgment $j \in \mathcal{J}$ following Equation 1. Each post-hoc rationale is generated independently for the corresponding seed judgment and is added to the persona alongside the judgment itself. Our *rationalized persona* is defined as:

$$\mathcal{Q}_\mathcal{R} = (\mathcal{D}; \{(j, r^j_{\text{basic}}) | j \in \mathcal{J}\}). \tag{3}$$

Building on Section 2.1, post-hoc rationales operationalize how $LM_\mathcal{R}$ generates rationales to explain *why* a user might hold a given judgment.

Using rationales from Equation 2, we arrive at the *scaffold rationalized persona*,

$$\mathcal{Q}_\psi = (\mathcal{D}; \{(j, r^j_\psi) | j \in \mathcal{J}\}). \tag{4}$$

3.2 Generating responses from Persona LMs

Persona descriptions $\mathcal{Q}$ obtained in §3.1 can be used as prompts to steer responses towards a given user persona, following prior work (Santurkar et al., 2023; Hwang et al., 2023). Consider a task where an individual user has provided several input-output pairs for a test instance $\{x, y\}$ (e.g., opinion question and answer pairs). The persona description can provide additional context to predict responses

Judgment: I am somewhat dissatisfied with my current job or career.
NO SCAFFOLD At my age, I've had many experiences in my working life and feel that I could have achieved more. While I continue to work because I believe in keeping busy and maintaining a sense of purpose, my current job doesn't give me the satisfaction and fulfillment I'd hoped for...
EXPERIENCES As a man in my senior years, I've spent decades working. I began in manual labor, but found a calling in community work and completed an associate's degree in social work. While this transition has been rewarding, I've grown somewhat dissatisfied with my job mainly because of ever-increasing paperwork and bureaucracy. They are detracting from the direct, meaningful interactions that I so enjoy and that rekindled my career passion in the first place. As I age, I've found that what I value most in work is not status or power...
BIG 5 PERSONALITY TRAITS This dissatisfaction may stem from extroversion and my desire for more social interaction that my current job may not provide. My conscientiousness also comes into play as I am a productive and organized individual who might feel stifled if these traits are not fully utilized in my career. At the same time, my neuroticism may cause me to overanalyze situations and magnify any negative aspects of my job...
SCHWARTZ THEORY OF BASIC HUMAN VALUES This opinion reflects my value of Achievement and Self-Direction. I value being successful and competent, and currently, my job is not allowing me to feel that sense of accomplishment. Additionally, I value independent actions and the ability to choose my own path, thus being dissatisfied with my current job also relates back to my sense of autonomy and need for personal fulfilment...
PRIMAL WORLD BELIEFS Despite being on the older side, I still have an enticement for experiencing new things and learning more. Sticking to the same old job with monotonous routines doesn't appeal to me because it does not satisfy my curiosity or hunger for novelty. Additionally, my safe world belief also comes into play. I believe in stability and minimal threats but my current job doesn't provide that comfort and ease that I expect at this point in my life. The constant worries and pressures create a sense of dissatisfaction. Lastly, I also believe in the alive world stance. I want a sense of purpose in what I do. I believe there is a reason for my dissatisfaction, and it's perhaps because this isn't the job I was meant to stay in for the rest of my life...

Table 2: Snippets of rationales generated by GPT-4 for a career satisfaction judgment according to the scaffolds used in PB&J. Full examples are shown in Appendix D, along with user demographics

for inputs, enabling an LM to generate accurate predictions for a user.

We predict a user's output for an input x by prompting an off-the-shelf language model $LM_{\mathcal{T}}$ with task-specific instructions $\mathcal{T}$.

$$\hat{y} = LM_{\mathcal{T}}(x, \mathcal{Q}), \tag{5}$$

To evaluate a persona building method, we vary $\mathcal{Q}$ over the base, rationalized, and scaffold rationalized personas $\mathcal{Q}_B, \mathcal{Q}_{\mathcal{R}}, \mathcal{Q}_{\psi}$.

4 Experiments

Task and Datasets. We conduct experiments on two datasets - OpinionQA (Santurkar et al., 2023) and MovieLens (Harper and Konstan., 2015). OpinionQA is a collection of American public opinion surveys (Pew Research Center, 2025) conducted by PEW Research containing user-selected answers to multiple-choice questions in 15 different topics ranging from food safety to guns, with rich accompanying demographic attributes. The task is to predict a user's answer for a question. For OpinionQA, we use a subset of 750 users, and 10 test questions for each user. MovieLens contains timestamped movie ratings (between 1-5) corresponding to individual users, with gender, age, occupation and location of each user. The task is to predict a user's rating for a movie. We use a subset of 100 users and 10 test movies for each user. Further details and splits of the dataset are presented in Appendix C. Example inputs-outputs of each of these datasets are shown in Figure 2.

Seed Judgments. Our persona description generates rationales for a set of seed judgments. For OpinionQA, these seed judgments are a 'train' set of questions and users' corresponding answers to them. The questions and answers are then converted into declarative forms (e.g., If a user responds *No, not really* to *"Are you currently focused on your professional life and career?"*, the declarative form would be *I'm not very focused on my professional life right now*). For MovieLens, we use movie ratings provided by the user with earlier timestamps for predicting ratings by the same user with later timestamps. Each judgment consists of a movie, its rating, and a short description of the movie consisting of its plot synopsis, actors, and directors. For our main experiments, a fixed set of 8 seed judgments are provided in the persona description.

Evaluation. We use a 'test' set of opinion questions and movies to evaluate the personas. Since our personas are customized to each real user, we calculate accuracies and standard deviations for both OpinionQA and MovieLens, macro-averaged for each user.

Language Models. We use two LMs of varying sizes — GPT-4[2](OpenAI et al., 2024) and Mistral

[2]Last accessed on 2 December 2024, GPT-4 points to `gpt-4-0613`.

Prompting Approach	OpinionQA		MovieLens	
	GPT-4	Mistral 7B	GPT-4	Mistral 7B
NO PERSONA	20.57 ± 15.15	32.08 ± 16.83	21.89 ± 13.01	06.30 ± 9.34
ONLY DEMOGRAPHICS	45.69 ± 17.46	32.97 ± 16.46	35.30 ± 18.68	22.90 ± 14.65
ONLY JUDGMENTS	33.63 ± 16.76	38.40 ± 16.25	32.50 ± 19.10	15.30 ± 17.97
DEMOGRAPHICS + JUDGMENTS	49.63 ± 16.84	42.17 ± 16.70	34.80 ± 17.52	21.20 ± 17.65
DEMOGRAPHICS + JUDGMENTS$_{CoT}$	49.17 ± 17.10	31.67 ± 15.59	30.20 ± 18.11	24.20 ± 18.30
PB&J$_{\text{NO SCAFFOLD}}$	53.71* ± 17.92	46.96* ± 16.44	24.80 ± 15.20	22.40 ± 14.01
PB&J$_{\text{EXPERIENCES}}$	54.12* ± 17.59	**47.61* ± 16.55**	30.50 ± 16.45	22.40 ± 15.37
PB&J$_{\text{BIG 5 PERSONALITY TRAITS}}$	53.59* ± 17.14	46.09* ± 16.75	35.79 ± 17.84	29.30* ± 16.02
PB&J$_{\text{SCHWARTZ THEORY OF BASIC HUMAN VALUES}}$	53.45* ± 17.17	45.00* ± 16.71	**39.89* ± 17.46**	26.70* ± 17.03
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	**54.43* ± 17.01**	45.52* ± 16.19	38.00* ± 16.91	**30.50* ± 16.15**

Table 3: **Improved Persona Alignment with PB&J**: Shown here are accuracy scores and standard deviations macro-averaged across users, for PB&J with different psychological scaffolds (NO SCAFFOLD, EXPERIENCES, BIG 5 PERSONALITY TRAITS, SCHWARTZ THEORY OF BASIC HUMAN VALUES, and PRIMAL WORLD BELIEFS). PB&J consistently outperforms baselines, demonstrating the effectiveness of scaffolded rationales in improving persona alignment. Best performing method is **bolded** and best performing baseline is underlined. * represents results that are significantly better ($p < 0.05$) than the best baseline. Full significance results are in Appendix E.

0.2 Instruct 7B (Jiang et al., 2023). All results are based on prompting these two LMs, without any fine-tuning.

Baselines. Following prior work, we experiment with different variants of building persona descriptions as baselines, Santurkar et al. (2023) propose two methods – one where LMs are prompted without any user information (NO PERSONA) and with demographics (ONLY DEMOGRAPHICS). Hwang et al. (2023) add judgments to the persona descriptions without (ONLY JUDGMENTS) and with demographics (DEMOGRAPHICS + JUDGMENTS). For the latter setting, we also include results using Chain-of-thought reasoning (Wei et al., 2023) as DEMOGRAPHICS + JUDGMENTS$_{CoT}$, where $LM_{\mathcal{T}}$ is prompted to reason about the test question using the persona information in order to arrive at the answer. This is $LM_{\mathcal{T}}$'s default rationale that is not explicitly grounded in any psychological scaffold. In Section 3.1, ONLY DEMOGRAPHICS, ONLY JUDGMENTS and DEMOGRAPHICS + JUDGMENTS correspond to $\mathcal{Q}_D$, $\mathcal{Q}_J$ and $\mathcal{Q}_B$ respectively.

Existing persona LM baselines are (almost) equivalent. Table 3 presents macro-averaged accuracies for various baseline methods across models and datasets. Without any persona information (NO PERSONA condition), both GPT-4 and Mistral 7B perform significantly worse, highlighting the necessity of incorporating some form of user context. The performance drop is particularly stark for GPT-4; due to its safety guardrails, GPT-4 abstains frequently from expressing opinions and preferences (Chen et al., 2023). This leads to responses such as *"This is a subjective question and the answer will vary"* or *"As an AI, I don't have personal experiences"*[3]. Demographic attributes (ONLY DEMOGRAPHICS) are generally more predictive than judgments alone (ONLY JUDGMENTS) with DEMOGRAPHICS + JUDGMENTS as the best performing baseline, especially for GPT-4. However, it is interesting to see that providing additional tokens to reason via Chain-of-Thought rationales in DEMOGRAPHICS + JUDGMENTS$_{CoT}$ also yields no significant improvements over DEMOGRAPHICS + JUDGMENTS.

Approach	Accuracy
NO PERSONA	08.61 ± 05.20
ONLY DEMOGRAPHICS	24.28 ± 15.81
ONLY JUDGMENTS	21.47 ± 08.50
DEMOGRAPHICS + JUDGMENTS	39.42 ± 11.43
DEMOGRAPHICS + JUDGMENTS$_{CoT}$	39.30 ± 11.82
PB&J$_{\text{NO SCAFFOLD}}$	44.62* ± 11.42
PB&J$_{\text{EXPERIENCES}}$	43.76* ± 11.60
PB&J$_{\text{BIG 5 PERSONALITY TRAITS}}$	44.61* ± 10.98
PB&J$_{\text{SCHWARTZ THEORY OF BASIC HUMAN VALUES}}$	45.33* ± 11.50
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	46.71* ± 11.52
PB&J$_{\text{HUMAN WRITTEN}}$	**48.52* ± 12.30**

Table 4: **Incorporating human-written rationales in PB&J:** Human-written rationales for OpinionQA judgments consistently outperform baselines and all LM-generated rationales except PRIMAL WORLD BELIEFS.

[3]Such behavior does not occur for all questions but has a substantial impact on overall performance.

PB&J provides significant improvements for personas. In Table 3, we also demonstrate variations of PB&J, where $\mathcal{Q}_R$ in Section 3.1 corresponds to the personas created using rationales without scaffolding (NO SCAFFOLD) and $\mathcal{Q}_\psi$ corresponds to all other personas with scaffolded rationales, where ψ relates to the scaffold-specific instructions. Across both datasets (Table 3), PB&J consistently outperforms the baselines, demonstrating the effectiveness of plausible, synthetic rationales (example generations in Table 2). On OpinionQA, scaffolded rationales yield significant improvements over DEMOGRAPHICS + JUDGMENTS, even after generating Chain-of-Thought rationales, with PRIMAL WORLD BELIEFS achieving the highest accuracy, closely followed by EXPERIENCES. Even unstructured rationales (NO SCAFFOLD) surpass DEMOGRAPHICS + JUDGMENTS$_{CoT}$. These trends hold across LMs.

For MovieLens, psychological scaffolds are even more critical. While GPT-4 performs better overall, smaller LMs like Mistral 7B benefit substantially from SCHWARTZ THEORY OF BASIC HUMAN VALUES and PRIMAL WORLD BELIEFS scaffolds. PRIMAL WORLD BELIEFS consistently ranks among the top scaffolds across datasets, however, SCHWARTZ THEORY OF BASIC HUMAN VALUES excels in MovieLens, highlighting that different tasks may benefit from different psychological frameworks. Lastly, we also experimented with combinations of scaffolds, including synthesizing summary rationales from all scaffolds, but these were less performant (Appendix F).

5 Using PB&J with human-written rationales

Table 3 demonstrates that plausible yet synthesized LM rationales effectively improve user personas. However, we hypothesize that the observed improvements stem not from the LM-generated rationales themselves, but from the additional context they provide, enabling more accurate generalization for each user. To test this, we conducted a pilot study using a subset of 30 questions from OpinionQA. We recruited 100 users to respond to the subset of questions, collected their demographics and asked them to provide rationales for the first nine responses. These rationales were unconstrained, allowing participants to explain their judgments based on personal experiences, beliefs, or personality traits without any imposed structure.

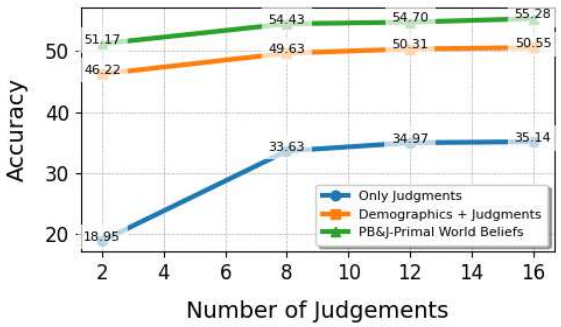

Figure 3: **Performance as a function of the number of user judgments:** PB&J outperforms baselines across all settings, providing substantial gains even with minimal judgments. All results use GPT-4.

The remaining 21 questions were used for evaluation. Details can be found in Appendix C.3.

Table 4 reports the results on this subset, with the inclusion of HUMAN WRITTEN rationales using GPT-4. HUMAN WRITTEN rationales outperform all baselines and PB&J variants, with the exception of PRIMAL WORLD BELIEFS, where the difference is not statistically significant. This highlights that even though PB&J generates synthetic rationales, plausible and carefully selected scaffolds could exhibit similar predictive power to HUMAN WRITTEN rationales.

6 Discussion

LM-generated rationales help even with a limited budget of judgments. Unlike demographics, user judgments take time to collect, making it crucial to assess personas with limited data (Section 4). As shown in Figure 3, PB&J with PRIMAL WORLD BELIEFS outperforms both baselines (ONLY JUDGMENTS and DEMOGRAPHICS + JUDGMENTS) across varying number of judgments. Even with just two judgments, PB&J far surpasses baselines, highlighting the value of LM-generated rationales in low-data settings. As the number of judgments increases, the performance of all methods improves. However, the rate of improvement for all diminishes after 8 judgments. This indicates that while all methods benefit from additional user judgments, PB&J maximizes its potential gains earlier due to the contextual richness provided by LM-generated rationales.

PB&J improves performance across demographics. A robust personalization method should improve performance across diverse user

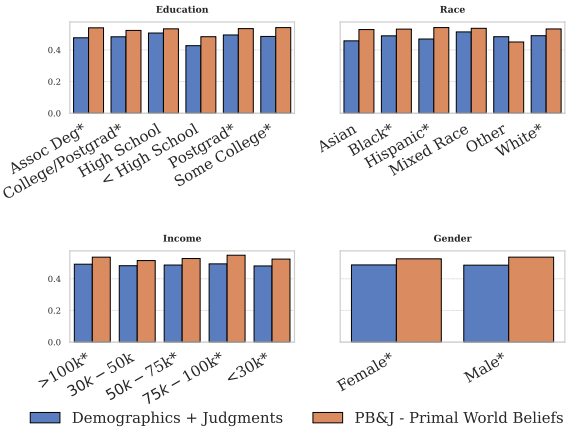

Figure 4: **PB&J's improvements over DEMOGRAPHICS + JUDGMENTS across education, race, income, and gender:** Subgroups marked with * indicate significant improvements ($p < 0.05$). All results use GPT-4.

groups rather than relying on gains from specific demographics. To assess this, we compare PB&J (with PRIMAL WORLD BELIEFS scaffolds) to DEMOGRAPHICS + JUDGMENTS across education, race, income, and gender in Figure 4 (splits across all demographics in Figure 5). PB&J consistently outperforms the baseline across all demographics. While gains are notable for users with postgraduate education and those identifying as Asian or Mixed Race, improvements extend across all education levels, racial groups, income brackets, and genders. By incorporating scaffolded rationales, PB&J improves with diverse user perspectives, demonstrating broad effectiveness with synthetic rationales.

PB&J improves performance *without* demographics. In cases where user demographics are unavailable or where the risks of bias outweigh the utility of demographic information, PB&J can improve performance of personalization from judgments alone. Rationalizing ONLY JUDGMENTS with PRIMAL WORLD BELIEFS results in an 11.64% absolute improvement on OpinionQA and 1.1% on MovieLens (GPT-4). While still underperforming demographic-based methods, this result indicates that deeper reasoning about user behavior alone may obviate the need for coarse and potentially biasing demographic information.

Effect of reasoning length on performance. While variants of PB&J improve performance over the default reasoning offered by DEMOGRAPHICS + JUDGMENTS$_{CoT}$, we investigate where this improvement comes from, using reasoning length as our control. On the pilot subset reported in Section 5, we observe that HUMAN WRITTEN rationales are much shorter in length (40.73 ± 25.66 tokens), as compared to chain-of-thought rationales generated by DEMOGRAPHICS + JUDGMENTS$_{CoT}$ (59.70 ± 16.90 tokens) and PB&J$_{\text{PRIMAL WORLD BELIEFS}}$ rationales (124.12 ± 38.16 tokens); however, HUMAN WRITTEN rationales outperform both of these variants. Across all PB&J variants in Section 5, we observe that the correlation between the accuracy and length of rationales (in terms of tokens) for corresponding users is negligible (Pearson's r=0.03). This suggests that performance gains are not simply a result of longer rationales, but stem from the quality and relevance of the information provided in the rationales.

Approach	Rationale PoV	Answer PoV	Accuracy
DEMOGRAPHICS + JUDGMENTS	-	first	32.05
DEMOGRAPHICS + JUDGMENTS	-	third	49.63
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	first	first	**54.43**
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	first	third	50.42
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	third	first	53.08
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	third	third	51.12

Table 5: **Ablations with different PoVs in PB&J:** We experiment with different PoVs to generate answers and rationales, for the best performing baseline, and variant of PB&J. All experiments are shown for GPT-4.

Effect of PoV in persona prompts. Our analysis (Table 5) examines how the point of view (PoV) in rationale and answer generation affects PB&J. All baselines benefit from using a *third*-person PoV for answers, suggesting that distancing the model from a subjective stance improves alignment. In contrast, PB&J performs best when both the rationale and answer are generated in *first*-person. Performance declines when either the rationale or answer shifts to *third*-person, indicating that while baselines may benefit from objective framing, *first*-person perspectives enhance persona consistency when paired with rationale-augmented personas. Therefore, all baselines depicted in Table 3 are prompted in *third*-person and all PB&J variants are prompted in *first*-person, for both the rationale and answer prompts.

7 Related Work

Personalizing LM. Recent works have used LM personas to simulate behavior in psycholinguistic and other social science experiments (Aher et al., 2023; Karra et al., 2022b; Filippas et al., 2023; Argyle et al., 2023). Specifically, the use of LMs

to simulate user responses to surveys, using existing user information like demographic background has been gaining increasing attention (Santurkar et al., 2023; Hwang et al., 2023; Durmus et al., 2024; Chuang et al., 2024; Do et al., 2024; Sun et al., 2025; Moore et al., 2024; Dillion et al., 2023; Tjuatja et al., 2023; Balepur et al., 2025) Recently, that attention has shifted more towards synthetically augmenting persona information (Moon et al., 2024; Simmons, 2022) or completely synthesizing personas generated from seed human data (Park et al., 2022, 2024, 2023; Ge et al., 2024).

Psychology and Personas. Psychologists have long investigated how different sociological and psychological aspects influence personality (McAdams, 1993; Bruner, 1991; Pennebaker and King, 1999). Recently, researchers have focused on eliciting psychological markers for evaluating LMs (Hilliard et al., 2024; Karra et al., 2022a; Serapio-García et al., 2023). Very few works actually incorporate these principles in an LM persona itself (Moon et al., 2024; Park et al., 2024).

Reasoning and Rationalization. Previous works have focused on generating reasoning chains or rationales by either prompting language models (Wei et al., 2023; Saha et al., 2023) or learning to generate rationales by fine-tuning on such data (Wiegreffe et al., 2022; Ramnath et al., 2024). Recently, there has been a surge in built-in reasoning capabilities in language models via extensive steps (Zelikman et al., 2022; DeepSeek-AI et al., 2025).

8 Conclusion

This work introduces PB&J, a framework that improves LM personas by incorporating plausible, yet synthetic rationales to explain user judgments. By leveraging psychological scaffolds, PB&J improves LM persona accuracy across diverse opinion prediction and preference modeling tasks, while performing, at best, close to human-written rationales. Additionally, PB&J remains effective even with a limited number of user judgments, highlighting its potential for real-world personalization applications, with scarce user history.

9 Acknowledgments

We thank anonymous reviewers and lab members at Apple and USC NLP for their feedback on this work. This work was also supported in part by the National Science Foundation under grant IIS-2403437, the Simons Foundation, and the Allen Institute for AI. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was partially done while B. Joshi was at Apple. and S. Swayamdipta was a visitor at the Simons Institute for the Theory of Computing. B. Joshi was also supported by the Apple Scholars in AL/ML PhD Fellowship.

Limitations

While PB&J improves personas through plausible yet synthetic LM-generated rationales, it relies solely on zero-shot prompting for both rationale generation and downstream predictions. While this allows for flexibility and adaptability across users, it may not fully capture the complexity and depth of individual reasoning. Fine-tuning LMs on human-written rationales could further improve personalization by enabling models to learn user-specific patterns rather than relying solely on generated rationales. Additionally, while these rationales improve performance, we cannot validate their fidelity to actual user reasoning, as no ground-truth rationales are available. This limitation is inherent to synthetic-data based persona modeling (Moon et al., 2024; Park et al., 2024), where plausible explanations generated by LMs may align well with observed user behavior but not necessarily reflect the true underlying motivations. Hence, we emphasize and caution that synthesized rationales can be *a plausible* reason for a judgment, but may not represent the user's exact reason. We provide further analyses about this in Appendix I and motivate future work in this space (Koncel-Kedziorski et al., 2025).

Ethics Statement

Our study primarily evaluates PB&J on U.S.-based user populations, as both OpinionQA and our human pilot study consist of participants located in the United States. Our study was conducted under the guidance of an ethics review board. Additionally, the subset of users selected from MovieLens also resides in the U.S. While this ensures consistency in evaluation, it limits the generalizability of our findings to more diverse global populations. Future work should explore the effectiveness of PB&J across different cultural and linguistic con-

texts to ensure broader applicability. Since our datasets involve personal judgments on opinion-based questions, some generated responses may reflect viewpoints that could be offensive or controversial. While we do not directly intervene in the LMs' generation of rationales, it is crucial to recognize that models can inherit biases present in both training data and user-generated inputs. Finally, as with any system that models human behavior, there are concerns around user privacy. While our work does not use real user data beyond voluntary survey responses, deploying such approaches in real-world settings would require careful consideration of data collection practices, consent mechanisms, and safeguards against potential misuse.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *Preprint*, arXiv:2208.10264.

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.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Whose boat does it float? improving personalization in preference tuning via inferred user personas. *Preprint*, arXiv:2501.11549.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Jérôme Seymour Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18:1 – 21.

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasović. 2024. On evaluating explanation utility for human-ai decision making in nlp. *Preprint*, arXiv:2407.03545.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *Preprint*, arXiv:2307.09009.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. *Preprint*, arXiv:2311.09618.

Paul Churchland and John Haldane. 1988. Folk psychology and the explanation of human behaviour. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 62:209–254.

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.

PT Costa and RR McCrae. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01):1999.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi,

Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Daniel Clement Dennett. 1981. *The Intentional Stance*. MIT Press.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27:597–600.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. 2024. Choire: Characterizing and predicting human opinions with chain of opinion reasoning. *Preprint*, arXiv:2311.08385.

Annette D'Onofrio. 2020. Personae in sociolinguistic variation. *Wiley Interdiscip. Rev. Cogn. Sci.*, 11(6):e1543.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Kostya Esmukov and contributors. 2025. geopy: Python geocoding toolbox. `https://github.com/geopy/geopy`. Accessed: 2025-02-15.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? In *ACM Conference on Economics and Computation*.

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.

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

Robert M. Gordon. 1986. Folk psychology as simulation. *Mind and Language*, 1(2):158–71.

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

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*.

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

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

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

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022b. Estimating the personality of white-box language models.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human

feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Rik Koncel-Kedziorski, Brihi Joshi, and Tim Paek. 2025. Primex: A dataset of worldview, opinion, and explanation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. *Preprint*, arXiv:2405.20253.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Bertram F. Malle. 2004. *How the mind explains behavior*.

Bertram F. Malle. 2009. Folk theories of consciousness. In William P. Banks, editor, *Encyclopedia of Consciousness*, pages 251–263. Elsevier.

Bertram F Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of experimental social psychology*, 33(2):101–121.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

D.P. McAdams. 1993. *The stories we live by: Personal myths and the making of the self*. William Morrow.

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

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

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

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

Alessandro Pasotti and contributors. 2025. Cinemagoer: Python package to access imdb's database. `https://github.com/cinemagoer/cinemagoer`. Accessed: 2025-02-15.

James W Pennebaker and Anna Graybeal. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77 6:1296–312.

Pew Research Center. 2025. The American Trends Panel. Accessed May 19, 2025.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2024. Tailoring self-rationalizers with multi-reward distillation. *Preprint*, arXiv:2311.02805.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via personalization. *Preprint*, arXiv:2306.09299.

Nikita Salkar, Thomas Trikalinos, Byron Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 341–350, Online only. Association for Computational Linguistics.

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

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

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.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

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.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III au2, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. *Preprint*, arXiv:2310.12558.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *ArXiv*, abs/2209.12106.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

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

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. 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. *Preprint*, arXiv:2402.11060.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Huy Vu, Salvatore Giorgi, Jeremy D. W. Clifton, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022a. Modeling latent dimensions of human beliefs. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1064–1074.

Huy Vu, Salvatore Giorgi, Jeremy DW Clifton, Niranjan Balasubramanian, and H Andrew Schwartz. 2022b. Modeling latent dimensions of human beliefs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1064–1074.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2022. Measuring association between labels and free-text rationales. *Preprint*, arXiv:2010.12762.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

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

A More Background on Psychological Scaffolds

A rich body of research has extended, debated, and validated each of the psychological scaffolds presented earlier. Among these, PRIMAL WORLD BELIEFS stands out as the most linguistically motivated framework, which may explain its strong performance in our experiments. As described in (Clifton et al., 2019), researchers analyzed historical texts and over 80K tweets using topic modeling and concept extraction to identify statements about how people perceive the world. These statements were then categorized through expert coding, consultation with social scientists, and discussions with religious focus groups, leading to the identification of 26 fundamental "primal world beliefs". These beliefs encapsulate deep-seated assumptions individuals hold about the world, such as whether the world is inherently safe or dangerous, simple or complex, and abundant or limited. Expanding on this work, (Vu et al., 2022a) introduced a Latent Beliefs Model, which leverages transformer-based embeddings and a modified GPT-2 model to automatically infer latent dimensions of human beliefs from social media text. This data-driven discovery of worldviews underscores the linguistic basis of primal beliefs and their connection to naturally occurring human rationales.

Another well-established framework in social psychology is the SCHWARTZ THEORY OF BASIC HUMAN VALUES (Schwartz, 1992; Schwartz et al., 2012). This theory posits that human values, which are fundamental guiding principles, are organized along universal motivational dimensions that drive behavior. Schwartz identifies ten broad value categories, such as self-direction (independence of thought and action), benevolence (concern for others' welfare), and power (desire for dominance or control). These values are structured in a circular model, where adjacent values are more compatible, and opposing values (e.g., security vs. stimulation) tend to be in tension. A key aspect of Schwartz's values is their cross-cultural validation; extensive empirical studies have shown that these value dimensions hold across diverse populations, making them a robust framework for modeling user judgments in an LM setting. Unlike primal beliefs, which describe broad worldviews, Schwartz's values provide a structured way to infer decision-making tendencies and moral considerations, making them particularly useful for understanding user

preferences in ethical or societal questions.

Similarly, the BIG 5 PERSONALITY TRAITS (also known as the OCEAN model) (Goldberg, 1993; Costa and McCrae, 1999; Pennebaker and Graybeal, 2001; Schwartz et al., 2013) offer a comprehensive framework for describing individual differences in personality. The Big Five dimensions—Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—have been extensively validated through psychometric studies and natural language analysis. These traits predict a wide range of behaviors, from political preferences to purchasing decisions, and have been found to correlate with linguistic patterns in social media and personal narratives. For instance, individuals high in Openness to Experience tend to use more abstract and imaginative language, while those high in Neuroticism are more likely to express negative emotions. Given these correlations, Big Five traits serve as a useful scaffold for generating rationales that reflect personality-driven reasoning processes, such as why a highly conscientious user might favor structured decision-making or why an extraverted user might prioritize social considerations.

Each of these psychological scaffolds offers a unique perspective on human behavior: Primal World Beliefs focus on fundamental assumptions about the world, Schwartz's values provide a structured way to model decision-making, and the Big Five capture stable personality traits that influence judgment and preference formation.

B PB&J prompts

PB&J is primarily an inference-time prompt-based strategy to improve LM personas, without requiring any fine-tuning. We pick the basic structure of our prompt from Hwang et al. (2023). For all psychological scaffolds, we try multiple variations of prompts with varying levels of instructions, definitions and examples. Varying instructions for scaffolds does not lead to significant changes in performance. For example, adding more detailed instructions for BIG 5 PERSONALITY TRAITS leads to a small, yet insignificant increase of 0.6 points for the OpinionQA HUMAN WRITTEN subset.

In Table 11 and Table 12, we provide the final prompts that we use to generate scaffolded rationales and predictions from both GPT-4 and Mistral 7B. These are used to generate the results displayed in Table 3 and Table 4.

C Dataset Details

C.1 OpinionQA

OpinionQA (Santurkar et al., 2023) contains fifteen topics, with multiple questions in each topic. Users answer several questions for a topic. There is no 1:1 correspondence between users in different topics. To this end, we select 50 users per topic, resulting in 750 unique users. For each user, we separate 8 answered questions and use them as prior judgments (this number changes when we run evaluations with varying number of judgments as in Figure 3). All remaining questions are used as a 'test set', out of which we randomly select 10 questions for evaluation. All configurations and demographic setup are similar to Hwang et al. (2023).

C.2 MovieLens

MovieLens (Harper and Konstan., 2015) contains user ratings (on a scale of 1-5) for movies belonging to different genres. These ratings are timestamped. While there are limited user demographic attributes, the dataset contains information about a user's age, occupation, location (zipcode) and gender. We convert the zipcode of every participant to a string-based city/state location (Esmukov and contributors, 2025). For curating judgments, we order movie ratings based on their timestamp, and pick the first 8 ratings as judgments in the persona, and from the remaining ratings, sample 10 to be used for evaluation. The judgments include the movie name and a user's rating. We augment the judgments with the movie's synopsis and key cast and crew using the IMDb API (Pasotti and contributors, 2025).

C.3 Human Pilot Dataset

For the human pilot experiment, we recruit 100 participants from a third-party user study company called User Research International. We selected a subset of 30 questions from OpinionQA belonging to 3 different topics: food, economics and America in 2050. Participants who consent to the study are requested to answer all 30 questions, but can refuse any question they want. For 9 questions, we also ask participants to provide a free-text rationale justifying their selection.

D More examples of PB&J-generated rationales

We present examples of personas constructed by PB&J. For given users with demographics, we

display a representative judgment provided by the user, and present rationales generated by PB&J using different scaffolds for the same judgment. We add examples in Table 6, Table 7 and Table 8.

E Significance Tests

We conduct statistical tests to assess whether the improvements of PB&J over the best-performing baseline are statistically significant. Table 9 presents significance test results, comparing each PB&J variant with the strongest baseline in its respective setting. We use a one-tailed independent t-test (Student, 1908) to evaluate the null hypothesis that PB&J does not provide a significant improvement over the baseline. To ensure robustness, we compute significance in two ways. The first approach, user-wise significance, examines whether PB&J improves performance on a per-user basis, assessing whether predictions for individual users show meaningful gains. The second, question-wise significance, evaluates improvements across all instances of a user, aggregating performance over multiple questions answered by the same user. For each comparison, we report the test statistic and p-value in Table 9.

F Combining Psychological Scaffolds

Approach	Accuracy
PB&J$_{\text{PRIMAL WORLD BELIEFS+SCHWARTZ THEORY OF BASIC HUMAN VALUES}}$	45.19 ± 10.74
PB&J$_{\text{EXPERIENCES+BIG 5 PERSONALITY TRAITS}}$	42.67 ± 11.45
PB&J$_{\text{CONCATALL}}$	41.40 ± 11.21
PB&J$_{\text{COMBINEALLTOONE}}$	44.28 ± 11.20

Table 10: **Combining Scaffolded Rationales:** Given that scaffolds are key to improve LM persona, we ask *to what extent can scaffolded rationales help?* We experiment with four varying settings where scaffolded rationales are combined.

Given that psychological scaffold-based rationales help improve LM personas, we also investigated settings where we combined rationales from multiple scaffolds for a user. On the OpinionQA subset containing the HUMAN WRITTEN rationales, we concatenate the top two (PRIMAL WORLD BELIEFS and SCHWARTZ THEORY OF BASIC HUMAN VALUES), bottom two (EXPERIENCES and BIG 5 PERSONALITY TRAITS) and all rationales (CONCATALL) as context for the LM persona. This leads to subpar performance; we posit that adding rationales from all scaffolds is too noisy for the LM to be able to select reasonable justifications to support user judgments. In order to mitigate this, inspired by Self-Consistency (Wang et al., 2023), we consolidate rationales from all scaffolds into a single rationale (COMBINEALLTOONE). An additional LM is used for this post hoc processing, where the LM is provided the following instructions: "*For a given judgment, you will be provided multiple rationales for why this person holds this judgment. Your job is to consolidate these rationales into one concise rationale. If the rationales are not consistent with each other or present diverging perspectives, you are allowed to pick a perspective, or also allowed to keep multiple perspectives for that judgment, based on what you think best reflects the person.*" While this improves over the CONCATALL setting, this is still not enough signal for the LM persona, unlike single scaffolds like PRIMAL WORLD BELIEFS.

G Model Details

We provide model configurations for both GPT-4 and Mistral 7B for rationale generation and answer generation steps in Table 13. GPT-4 is under proprietary license, and Mistral 7B is subject to the Apache 2.0 license.

H Additional Discussion

I Analysis of LM generated rationales

Context analysis of LM-generated rationales To better understand how different scaffolds influence the content and structure of LM-generated rationales, we analyze the rationales along three key dimensions: demographic reliance, lexical diversity, and structural repetition (Table 14). The percentage of rationales mentioning demographics and the density of demographic keywords capture how often rationales explicitly reference user identity. The average 2-gram diversity per user and per judgment reflects the linguistic variability of rationales, with lower values indicating greater diversity (i.e., less repetition). This score is also known as self-repetition (Salkar et al., 2022; Shaib et al., 2024).

Our findings reveal distinct patterns across scaffolds. Experience-based rationales (PB&J$_{\text{EXPERIENCES}}$) exhibit the strongest reliance on demographic cues, with the highest demographic mention rate (99.25%) and density (1.57), suggesting that user experiences are often framed explicitly through identity attributes. Trait- and value-based scaffolds (PB&J$_{\text{BIG 5 PERSONALITY TRAITS}}$,

User Demographics:
Age: 18-29, Citizenship: Yes, Region: South, Education: Postgraduate, Income: $75,000-$100,000, Marital status: Married, Political ideology: Liberal, Political party: Democrat, Race: White, Religion: Nothing in particular, Frequency of religious attendance: Never, Gender: Female

Selected Judgment:
I am somewhat concerned that global climate change will harm me personally at some point in my lifetime.

PB&J-generated rationales:

NO SCAFFOLD
I believe that the effects of global climate change are increasingly evident and impactful. The rise in natural disasters, unpredictable weather, food and water shortages, even though they might seem distanced, have the potential to directly affect my lifestyle, health, and well-being. It's a concern not just for the distant future, but for my lifetime as well, hence my concern.

EXPERIENCES
As a postgraduate scholar majoring in environmental sciences, I am constantly exposed to alarming research findings on global climate change. Personally, I love outdoor activities like hiking and camping, and it's heartbreaking seeing the devastation in forests due to wildfires reportedly linked to climate change. This, coupled with evidence of rising sea levels and increased frequency of extreme weather events, propels my concern about how this can directly impact me and our planet.

BIG 5 PERSONALITY TRAITS
My concern about global climate change affecting me personally stems from my high level of Conscientiousness. As someone responsible and oriented toward future planning, I understand the significant risks and consequences posed by climate change. Moreover, my Openness to experience allows me to accept and internalize scientific evidence and reports that underline the severity of the issue. My slight Neuroticism also plays a role, making me more prone to worry about potential negative outcomes, hence my concern about personal harm due to global climate change.

SCHWARTZ THEORY OF BASIC HUMAN VALUES
This opinion is grounded in a few of Shwartz' basic human values. Primarily, it expresses my value of Universalism, as I am aware of the interconnections and mutual dependencies in our global ecosystem and am concerned for the welfare of nature. It also links to Security, as the threat of climate change disrupts the stability of my personal life and the wider community. Lastly, the opinion slightly touches on Self-direction. My awareness and concern reflect my informed understanding and independent thought on global matters.

PRIMAL WORLD BELIEFS
This viewpoint aligns with my low Safe World Belief, as one can view global climate change as a significant and imminent threat to stability and personal safety. Despite the advancements in technology and modern conveniences, our safety is still compromised by environmental degradation. Furthermore, my high Enticing World Belief encourages me to explore and understand more about this global issue, its impacts, and the ways to mitigate it. However, my low Alive World Belief also reinforces this concern as climate change does not seem to occur with comprehensible intent or purpose, instead contributing to an unpredictable threat to my life.

Table 6: Example 1

PB&J$_{\text{SCHWARTZ THEORY OF BASIC HUMAN VALUES}}$) show less reliance on demographics, producing rationales that generalize beyond identity markers. Belief-driven rationales (PB&J$_{\text{PRIMAL WORLD BELIEFS}}$) are highly structured but more repetitive, showing the lowest lexical diversity across users and judgments. Notably, unstructured rationales (PB&J$_{\text{NO SCAFFOLD}}$) generate the most varied rationales, highlighting that scaffolding provides consistency at the cost of lexical diversity. These findings suggest that structured reasoning, rather than demographic emphasis or linguistic diversity, is key to improving persona alignment. Scaffolds like PRIMAL WORLD BELIEFS highlight how value-based rationales can enhance LM personalization without overfitting to surface-level demographic traits.

Comparing LM-generated rationales with human-written. While we emphasize that PB&J generates plausible yet synthetic rationales that *may* not be the actual reasoning user by the user, we acknowledge in Section 9 that this is indeed a problem with any synthetic data used in an LM pipeline. However, we compare and contrast rationales generated by PB&J with HUMAN WRITTEN ones on the OpinionQA HUMAN WRITTEN subset. Table 15 presents examples that fall into two broad categories: cases where human and LM rationales reflect similar reasoning, and cases where they diverge. In the former, LM rationales often elaborate on the human rationale, mapping it onto structured beliefs such as the Safe or Enticing World. In the latter, human rationales sometimes cite specific personal anecdotes or cultural references that are absent in LM generations, which tend to rely on general psychological framing.

Despite this variability, PB&J still benefits from the structure provided by scaffolds like PRIMAL

User Demographics:
Age: 65+,Citizenship: Yes, Region: Northeast, Education: High school graduate, Income: $100,000 or more, Marital status: Divorced, Political ideology: Very conservative, Political party: Other, Race: White, Religion: Protestant, Frequency of religious attendance: Never, Gender: Female

Selected Judgment:
I feel no pressure from family members to marry my partner.

PB&J-generated rationales:

NO SCAFFOLD
I believe in independence and making my own decisions. Also, having been through a divorce, my family understands and respects my perspective about not wanting to remarry.

EXPERIENCES
After my divorce, my family has been exceptionally supportive of my decisions regarding my love life. With my grown children doing well and leading their own lives, they are more focused on my happiness than societal expectations. They have always wanted me to be comfortable and happy, rather than rush into any decisions. That's why I haven't felt any pressure from them to marry my current partner.

BIG 5 PERSONALITY TRAITS
As a very conservative person who is also divorced, I've learned the importance of self-reliance and making decisions guided by my own perceptions and experiences (relating to the trait of conscientiousness). My family respects my conservative beliefs and my decision not to remarry, thereby causing no pressure for me to marry my partner. Furthermore, my trait of extroversion makes me capable of effectively communicating my decisions and viewpoints to my family, reinforcing their understanding and respect for my personal choices.

SCHWARTZ THEORY OF BASIC HUMAN VALUES
This opinion is rooted in the values of Self-direction and Achievement. My high-income bracket has allowed me a certain degree of independence and control over my life choices, which in turn alleviates any external pressure from family to marry my partner. Moreover, being of a mature age, I have lived through the social norms and expectations often imposed on younger generations. Hence, my achievements in life so far have given me the confidence and autonomy to make my own decisions.

PRIMAL WORLD BELIEFS
As someone with a very high Safe World Belief, I live in a cooperative and comforting environment where my decisions are respected. This harmonious setting also allows for open conversation about my personal choices, including my decision to stay unmarried to my partner. Thus, my family members do not pressure me to marry.

Table 7: Example 2

WORLD BELIEFS, often yielding rationales that are plausible and useful for prediction, even when the reasoning path differs from that of the original user. This highlights both the promise and limitations of structured, synthetic rationalization for modeling human behavior.

J Wasserstein Distance Metrics

Approach	GPT-4	Mistral 7B
NO PERSONA	0.95 ± 0.50	2.66 ± 0.66
ONLY DEMOGRAPHICS	0.75 ± 0.42	0.97 ± 0.53
ONLY JUDGMENTS	0.83 ± 0.44	1.31 ± 0.62
DEMOGRAPHICS + JUDGMENTS	0.81 ±0.45	1.03 ± 0.56
DEMOGRAPHICS + JUDGMENTS$_{CoT}$	0.88 ±0.39	0.97 ± 0.27
PB&J$_{\text{NO SCAFFOLD}}$	1.33 ± 0.69	1.41 ± 0.69
PB&J$_{\text{EXPERIENCES}}$	0.99 ± 0.61	1.45 ± 0.72
PB&J$_{\text{BIG 5 PERSONALITY TRAITS}}$	0.84 ± 0.63	0.64 ± 0.31
PB&J$_{\text{SCHWARTZ THEORY OF BASIC HUMAN VALUES}}$	0.61 ± 0.33	1.14 ± 0.67
PB&J$_{\text{PRIMAL WORLD BELIEFS}}$	0.64 ± 0.34	0.61 ± 0.32

Table 16: **Wasserstein Distance Metrics for MovieLens**

In addition to accuracy, we also present the Wasserstein Distance (WSD) between predicted ratings and user-provided ratings for the MovieLens dataset. We present these results in Table 16. Note that lower values are better for Wasserstein Distance.

User Demographics:
Age: 30-49, Citizenship: No, Region: South, Education: College graduate/some postgrad, Income: $50,000-$75,000, Marital status: Married, Political ideology: Conservative, Political party: Independent, Race: Black, Religion: Protestant, Frequency of religious attendance: More than once a week, Gender: Male

Selected Judgment:
The promotion of racial and ethnic diversity in the workplace is somewhat important for companies and organizations.

PB&J-generated rationales:

NO SCAFFOLD
While I believe in the importance of individual skill and merit, I also acknowledge the role racial and ethnic diversity plays in fostering a broad range of perspectives in the workplace. This, in turn, can drive innovation, create a more inclusive environment, and better represent the diverse markets that today's companies and organizations serve. Nevertheless, I don't regard it as the most crucial factor for success. Hiring should always focus primarily on capability and qualifications.

EXPERIENCES
As someone who has spent a significant portion of my career working in a corporate setting, I have observed firsthand how diversity contributes to a more innovative and productive work environment. Companies with a wide range of perspectives—from people of all races, ethnicities, and backgrounds—tend to solve problems more creatively and be more adaptable in the face of change. While I believe many factors play into a business's success, I have come to appreciate that variety in cultural background and experience certainly enriches the workplace and can provide a competitive edge.

BIG 5 PERSONALITY TRAITS
With my level of conscientiousness, I believe that recruiting a workforce that is diverse allows companies and organizations to draw upon a broader range of insights, skills, and experiences, thereby enhancing their productivity and competitiveness. My openness to experience influences my belief that this diversity brings in new and varied perspectives which leads to innovative problem solving. However, it's only 'somewhat' important because I believe merit should be the primary determining factor for hiring and promotions, rather than meeting a quota system.

SCHWARTZ THEORY OF BASIC HUMAN VALUES
This opinion aligns with my value of Universalism, as promoting racial and ethnic diversity in the workplace fosters an environment of tolerance and respect for all people, regardless of their background. Additionally, from a Power perspective, a diverse workplace may offer more opportunities for differing perspectives, which could yield more balanced and inclusive decision-making processes, thereby amplifying social status and control over resources. Furthermore, such diversity contributes to Achievement by elevating a organization's competency and credibility in an increasingly globalized world.

PRIMAL WORLD BELIEFS
As someone who has spent a significant portion of my career working in a corporate setting, I have observed firsthand how diversity contributes to a more innovative and productive work environment. Companies with a wide range of perspectives—from people of all races, ethnicities, and backgrounds—tend to solve problems more creatively and be more adaptable in the face of change. While I believe many factors play into a business's success, I have come to appreciate that variety in cultural background and experience certainly enriches the workplace and can provide a competitive edge.

Table 8: Example 3

Dataset	Model	Best Baseline	Approach	User-wise Significance	Question-wise Significance
OpinionQA	GPT-4	DEMOGRAPHICS + JUDGMENTS	NO SCAFFOLD	statistic=-8.9082, p=1.94e-18	statistic=-9.0335, p=1.04e-19
			EXPERIENCES	statistic=-9.8689, p=5.57e-22	statistic=-9.8824, p=3.42e-23
			BIG 5 PERSONALITY TRAITS	statistic=-9.1648, p=2.34e-19	statistic=-9.0564, p=8.47e-20
			SCHWARTZ THEORY OF BASIC HUMAN VALUES	statistic=-9.0304, p=7.14e-19	statistic=-8.8108, p=7.62e-19
			PRIMAL WORLD BELIEFS	statistic=-9.9742, p=2.19e-22	statistic=-9.8010, p=7.61e-23
	Mistral 7B	DEMOGRAPHICS + JUDGMENTS	NO SCAFFOLD	statistic=-6.1466, p=6.43e-10	statistic=-6.1429, p=4.26e-10
			EXPERIENCES	statistic=-6.8153, p=9.68e-12	statistic=-7.2184, p=2.89e-13
			BIG 5 PERSONALITY TRAITS	statistic=-4.4573, p=4.78e-06	statistic=-4.6148, p=2.00e-06
			SCHWARTZ THEORY OF BASIC HUMAN VALUES	statistic=-3.5916, p=1.75e-04	statistic=-3.7695, p=8.24e-05
			PRIMAL WORLD BELIEFS	statistic=-3.9861, p=3.69e-05	statistic=-4.1223, p=1.90e-05
MovieLens	GPT-4	ONLY DEMOGRAPHICS	BIG 5 PERSONALITY TRAITS	statistic=-0.2627, p=0.397	statistic=-0.3260, p=0.372
			SCHWARTZ THEORY OF BASIC HUMAN VALUES	statistic=-2.6476, p=0.00472	statistic=-3.1586, p=0.00082
			PRIMAL WORLD BELIEFS	statistic=-1.7573, p=0.04098	statistic=-1.8790, p=0.03027
	Mistral 7B	ONLY DEMOGRAPHICS	BIG 5 PERSONALITY TRAITS	statistic=-3.0672, p=0.00139	statistic=-3.5543, p=0.00020
			SCHWARTZ THEORY OF BASIC HUMAN VALUES	statistic=-1.8034, p=0.03718	statistic=-2.1552, p=0.01569
			PRIMAL WORLD BELIEFS	statistic=-3.8985, p=8.81e-05	statistic=-4.1427, p=1.86e-05
Human Pilot	GPT-4	DEMOGRAPHICS + JUDGMENTS	NO SCAFFOLD	statistic=-4.9186, p=1.74e-06	statistic=-5.5086, p=2.03e-08
			EXPERIENCES	statistic=-2.4163, p=0.00876	statistic=-2.9501, p=0.00161
			BIG 5 PERSONALITY TRAITS	statistic=-3.0124, p=0.00165	statistic=-3.5183, p=0.00022
			SCHWARTZ THEORY OF BASIC HUMAN VALUES	statistic=-3.4111, p=0.00047	statistic=-4.0207, p=3.00e-05
			PRIMAL WORLD BELIEFS	statistic=-7.2335, p=5.05e-11	statistic=-7.8278, p=3.90e-15
			Human Pilot	statistic=-8.7457, p=2.96e-14	statistic=-9.2996, p=1.71e-20

Table 9: **Statistical significance of PB&J's improvements for both GPT-4 and Mistral 7B:** Results from one-sided paired t-tests comparing PB&J variants to the best-performing baselines. We report both user-wise and question-wise significance.

System Message:
A person can be described as follows:

<demographic information>

The person has the following judgements:
1. <judgement_1>
2. <judgement_2>
3. <judgement_3>

User Message:

NO SCAFFOLD
For a given judgement, what would be a reasonable explanation that the person would provide for holding that judgement?

Judgement: <judgement_i>

EXPERIENCES
For a given judgement, what would be a reasonable explanation that the person would provide for holding that judgement? The explanation should contain a specific experience or personality trait (for example, fill in details of the university, or food, or any other detail, that the person would use to better explain their judgment).

Judgement: <judgement_i>

BIG 5 PERSONALITY TRAITS
For a given judgement, what would be a brief, reasonable explanation that the person would provide for holding that judgement? The explanation should be grounded in the big five personality traits listed below:

1. Openness to experience (includes aspects such as intellectual curiosity and creative imagination)
2. Conscientiousness (organization, productiveness, responsibility)
3. Extroversion (sociability, assertiveness; its opposite is Introversion)
4. Agreeableness (compassion, respectfulness, trust in others)
5. Neuroticism (tendencies toward anxiety and depression)

Judgement: <judgement_i>

SCHWARTZ THEORY OF BASIC HUMAN VALUES
For a given judgement, what would be a brief, reasonable explanation that the person would provide for holding that judgement? The explanation should be grounded in the Shwartz Theory of Basic Human Values listed below:

1. Power: Refers to the pursuit of social status, dominance, and control over people and resources.
2. Achievement: Personal pursuit of success, demonstrating competence according to social standards.
3. Hedonism: Pursuit of pleasure, enjoyment, and sensory and emotional gratification.
4. Stimulation: Seeks novelty and challenge in life, valuing excitement, variety, and adventure.
5. Self-direction: Independent thought and action — choosing, creating, and exploring.
6. Universalism: Understanding, appreciation, tolerance, and protection for the welfare of all people and nature.
7. Benevolence: Preserving and enhancing the welfare of those with whom one is in frequent personal contact (the 'in-group').
8. Tradition: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide the self.
9. Conformity: Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
10. Security: Safety, harmony, and stability of society, relationships, and the self.

Judgement: <judgement_i>

PRIMAL WORLD BELIEFS
For a given judgement, what would be a brief, reasonable explanation that the person would provide for holding that judgement? The explanation should be grounded in the three primal world beliefs listed below:

1. Safe World Belief: Those low on Safe see a Hobbesian world defined by misery, decay, scarcity, brutality, and dangers of all sorts. Those high on Safe see a world of cooperation, comfort, stability, and few threats.
2. Enticing World Belief: Those low on Enticing inhabit dull and ugly worlds where exploration offers low return on investment. Those high on Enticing inhabit an irresistibly fascinating reality.
3. Alive World Belief: Those low on Alive inhabit inanimate, mechanical worlds without awareness or intent. Those high on Alive sense that everything happens for a purpose and are thus sensitive to those purposes.

Judgement: <judgement_i>

Table 11: **Prompts used to generate rationales for judgments:** We use a common system message that includes a user's demographic information and all prior judgments held by the user. The user message then includes scaffold specific instructions ψ to generate rationales for a specific judgment.

System Message:
You are the following person:

<demographic information>

You have the following opinions:
1. <judgement_1>+<rationale_1>
2. <judgement_2>+<rationale_2>
3. <judgement_3>+<rationale_3>

User Message:

Based on your demographic and opinion information above, which answer would you select for the question shown below?

Question: <question>
Answer choices: <choice>

User Message with Chain of Thought:

Based on the above list of opinions and the demographic information, what would you choose for the question shown below? Provide your answer in the following format - "Reason: <reason>, Answer: <answer>". Only answer amongst the provided options, nothing else. Do not abstain from answering.

Question: <question>
Answer choices: <choice>

Table 12: **Prompts used to predict answers, given a persona:** We use a common system message that includes a user's demographic information and all prior judgments, along with generated rationales. The user message then includes the exact question and answer choices, with or without rationales.

Config	GPT-4	Mistral 7B
model	**GPT-4-0613** Number of parameters: Unknown	**Mistral 7B 0.2 Instruct** Number of parameters: 7 billion
Rationale Generation		
new_tokens	256	256
temperature	1	1
seed	6	6
GPU	N/A, openai api call	3 A100
Inferring time	2 hours	1 hour
Answer Generation		
new_tokens	10	275
temperature	0	0
seed	6	6
GPU	N/A, openai api call	3 A100
Inferring time	2 hours	1 hour

Table 13: Model Configurations for Rationale and Answer Generation

Approach	% of rationales that have atleast one demographic mentioned (↑)	Demographic Keyword Density (# of demographic keywords / rationale) (↑)	Avg. diversity of 2-grams in rationales per user (↓)	Avg. diversity of 2-grams in rationales per judgment (↓)
PB&J$_{\text{No Scaffold}}$	94.66	1.26 ±0.50	4.15 ±0.12	4.18 ±0.34
PB&J$_{\text{Experiences}}$	99.25	1.57 ±0.47	4.66 ±0.11	4.64 ±0.29
PB&J$_{\text{Big 5 Personality Traits}}$	94.96	1.41 ±0.57	4.60 ±0.12	4.54 ±0.33
PB&J$_{\text{Schwartz Theory of Basic Human Values}}$	89.03	1.14 ±0.54	4.61 ±0.11	4.51 ±0.36
PB&J$_{\text{Primal World Beliefs}}$	89.33	1.22 ±0.61	4.83 ±0.13	4.75 ±0.34

Table 14: **Analysis of LM-generated explanations across psychological scaffolds, measuring demographic mentions, keyword density, and 2-gram diversity:** Results suggest that strong persona alignment does not necessarily require high demographic reliance or linguistic variation. All analysis on rationales generated by GPT-4.

Example Type: Similar Reasoning
Selected Judgment:
The growing number of legal immigrants working in the U.S. does not contribute too much to economic inequality in the country.

PRIMAL WORLD BELIEFS rationale:

I firmly stand on the opinion that the growing number of legal immigrants in the U.S. does not contribute significantly to economic inequality. Grounded in the Safe World Belief, I perceive the world as one of cooperation where diverse people can coexist harmoniously. Immigrants often fill roles in our workforce that are frequently overlooked, thus contributing to the economy and overall stability of our country.
From the Enticing World Belief perspective, our world is fascinating and enriched by the presence of individuals of different cultures and skillsets. Legal immigrants bring unique perspectives and cultural norms that can actually enhance our socioeconomic environment.
Finally, looking through the lens of the Alive World Belief lens, I understand that people immigrate to the U.S. for a better life, not to contribute to economic disparity. They serve a purpose in the economy, thereby contributing to the betterment of society as a whole. *They are not the ones who create or increase economic inequality. Instead, it's systemic policies and structures that need addressing.*

HUMAN WRITTEN rationale:

With or without them, capitalism is designed to make economic inequality.

Example Type: Similar Reasoning
Selected Judgment:
All in all, I favor the use of animals in scientific research.

PRIMAL WORLD BELIEFS rationale:

My belief is rooted in the Enticing World belief. I see the world as a place of progress and innovation. The use of animals in scientific research enables us to make strides in understanding diseases and developing treatments. While I acknowledge it's a controversial topic due to ethical considerations, I believe the benefits outweigh the negatives. This kind of research promises great return, potentially leading to medical breakthroughs that can enhance human health and longevity.

HUMAN WRITTEN rationale:

Although I do not endorse animal cruelty or abuse, I think some form of scientific research using animals is necessary for the greater human good. Even though these animals might suffer or die, they are helping to advance human society.

Example Type: Not Similar Reasoning
Selected Judgment:
Medical treatments today are considered worth the costs because they enable individuals to live longer and enjoy better quality lives.

PRIMAL WORLD BELIEFS rationale:

I hold this opinion mainly because of my high score on the safe world belief. I believe in a world of cooperation and comfort, where people come together to ensure the wellbeing of everyone. Today's medical treatments, despite their costs, allow individuals to live longer, healthier lives, which contributes to a safer, more comfortable world. I also have a high enticing world belief because medical advancements hold the possibility of new discoveries, further enhancing our quality of life and making the world more enticing. Lastly, my moderate alive world belief stresses the significance of purpose in every happening, and I see the purpose of these treatments in promoting longer and healthier lives.

HUMAN WRITTEN rationale:

I loved listening to A Prairie Home Companion on NPR. A while back, the host, Garrison Keillor, had a heart valve replacement. Many of his ancestors had just died at his age, but he got to entertain us for another 20 and is still touring at 82.

Example Type: Not Similar Reasoning
Selected Judgment:
Medical treatments today often create as many problems as they solve.

PRIMAL WORLD BELIEFS rationale:

I believe that we are living in a Safe World, where medical technologies are advancing rapidly. However, sometimes, this progression can be seen as a double-edged sword. While it offers the possibility to treat illnesses, it can also lead to unforeseen side effects and complications, potentially causing more harm than good. This is a result of the complexity of the human body and the inherent risks associated with medical treatments. We must balance the desire for healing with an understanding of the potential risks involved.

HUMAN WRITTEN rationale:

Over priced medical care

Table 15: We provide examples of PRIMAL WORLD BELIEFS and HUMAN WRITTEN rationales for the same judgment and user. We observe that sometimes, LM-generated rationales reflect the same reasoning conveyed by the human-written ones, however, in cases where the human-written rationales are very niche, it is hard for the LM-generated rationales to mimic the same reasoning.

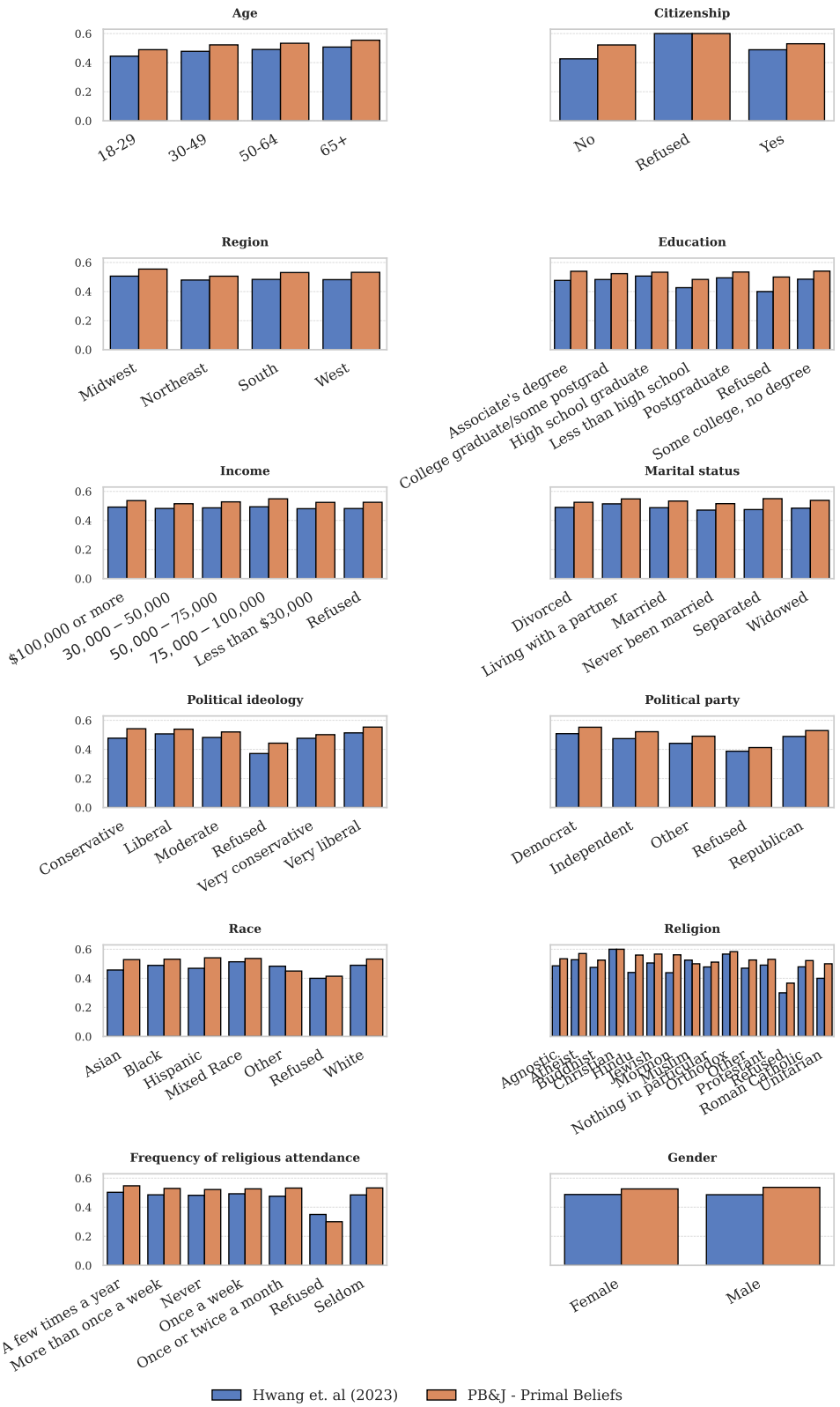

Figure 5: **PB&J's improvements over DEMOGRAPHICS + JUDGMENTS across all demographics:** Subgroups marked with * indicate significant improvements ($p < 0.05$). All results use GPT-4.