MTPA: MultiTask Personalization Assessment

Matthieu Tehenan, Eric Chamoun, Andreas Vlachos
Department of Computer Science
University of Cambridge
{mm2833,ec806,av308}@cam.ac.uk

Abstract

Large language models are increasingly expected to adapt to individual users, reflecting differences in preferences, values, and communication styles. To evaluate whether models can serve diverse populations, we introduce MTPA, a benchmark that leverages large-scale survey data (WVS, EVS, GSS) to construct real, hyper-granular personas spanning demographics, beliefs, and values. Unlike prior benchmarks that rely on synthetic profiles or narrow trait prediction, MTPA conditions models on real personas and systematically tests their behavior across core alignment tasks. We show that persona conditioning exposes pluralistic misalignment: while aggregate metrics suggest models are truthful and safe, subgroup-specific evaluations reveal hidden pockets of degraded factuality, fairness disparities, and inconsistent value alignment. Alongside the benchmark, we release a dataset, toolkit, and baseline evaluations. MTPA is designed with extensibility and sustainability in mind: as the underlying survey datasets are regularly updated, MTPA supports regular integration of new populations and user traits.

matthieu-perso/mtpa
matthieunlp/mtpa

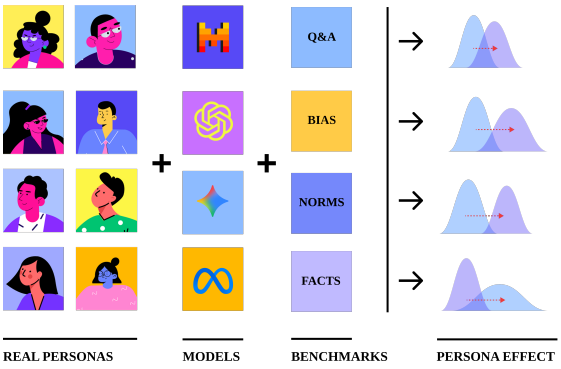

Figure 1: Overview of the MTPA benchmark. Real human personas (left) are used to condition the prompts of large language models (center) and evaluated against benchmark tasks (right, e.g., Q&A, bias, norms, factuality). The resulting outputs are compared to an unconditioned baseline, yielding measurable "persona effects" on the benchmark results (far right).

1 Introduction

Large language models (LLMs) are increasingly expected to generate personalized text, adapting to user preferences, values, and communication styles. From writing assistance and tutoring to recommendations and dialogue, personalization is emerging as a core topic in alignment research (Jiang et al., 2024; Zhang et al., 2024; Liu et al., 2025; Tseng et al., 2024). In this setting, alignment is no longer only about adhering to universal ethical rules (Kenton et al., 2021; Gabriel, 2020); it also requires language models to reflect the specific goals, value systems, and reasoning patterns of individual users (Kim et al., 2023b; Sorensen et al., 2024b).

In this sense, personalized alignment entails adapting model behavior to reflect a given user's worldview while preserving factual reliability, safety constraints, and social acceptability. (Sorensen et al., 2024a; Kirk et al., 2023). Yet alignment to diverse users presumes that the model can adequately represent user traits that drive individual variation in task performance and expectations. Pluralistic alignment therefore requires evaluating not only whether a model performs well in aggregate, but whether it generalizes fairly and consistently across diverse populations. However, most current alignment datasets and benchmarks are constructed around binary preference comparisons (e.g., pairwise preference choices) and lack user-level modeling (Shen et al., 2023; Ji et al.; Cao et al., 2024; Bai et al., 2022). This framing bypasses the user as an entity, and focuses instead on mapping preference pairs to outputs (Wang et al., 2024). As a result, a model may perform well on such benchmarks without reflecting a larger spectrum of user

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11982–11992
November 4-9, 2025

traits (Sorensen et al., 2024b; Chen, 2023; Zhang et al., 2024). This gap has motivated interest in studying how language models encode user traits. Early work has examined the dominant opinions in language models, such as the representation of political leanings (Liu et al., 2021; Hartmann et al., 2023), or the superposition of cultural perspectives (Kovač et al., 2023). However, these efforts remain limited in scope and size. Representation of personalization is often tested on synthetic personas, which may themselves reflect internal model biases (Castricato et al., 2024; Tan et al., 2025). The use of real human information is typically narrow and task-specific, such as political opinions (Santurkar et al., 2023; Durmus et al., 2023) or culturally anchored norms (AlKhamissi et al., 2024; Rao et al., 2024). Existing survey-based benchmarks such as WorldValuesBench (Zhao et al., 2024) or LocalValueBench (Meadows et al., 2024) focus on trait inference alone, often restricted to values or attitudes, and do not connect user representation to downstream alignment performance.

We introduce **MTPA**, a benchmark for evaluating pluralistic alignment in large language models. Built from large-scale sociological surveys (WVS, EVS, GSS), MTPA constructs hyper-granular, real human personas spanning demographics, values, trust, ideology, and other sociopolitical attitudes. Unlike prior benchmarks, MTPA conditions models on these personas and systematically tests their behavior across alignment tasks. This design enables us to measure how model performance shifts when evaluated "as" diverse subgroups, revealing persona-conditioned gaps in truth, safety, and fairness that are invisible to aggregate metrics. The core contribution of MTPA is a standardized framework for **persona-conditioned evaluation**, which would enable systematic pluralistic alignment audits. Because the underlying survey datasets are updated regularly, MTPA is structured to support continual integration of new respondents, evolving social attitudes, and newly emerging traits.

Our contributions are fourfold. First, we release **a benchmark framework** built from over 200,000 real human profiles with more than 250 interpretable traits. Second, we define a standardized protocol for **persona-conditioned evaluation** that can be easily adapted across any downstream benchmark or dataset. Third, we present **baseline results** from closed and open-weight LLMs (e.g., GPT-4o, Gemini, LLaMA-3) on alignment datasets. Finally, we release an open **dataset and evaluation suite** on HuggingFace and GitHub to enable reproducible, extensible research in pluralistic alignment.

2 MTPA Benchmark Design

2.1 Overview

A user representation z can be decomposed into granular traits that condition the model output y. Formally, we write $T = (t_1, t_2, \ldots, t_d)$, where each t_i denotes a specific user trait or preference. A particular user is thus identified with a configuration $T \in \mathcal{Z}$, where $\mathcal{Z}$ is the space of real user profiles. To evaluate personalization, models must then be tested across a representative and diverse sample of user representations $z \in \mathcal{Z}$, taking into account factors as diverse as region, age, language, preferences or occupation.

2.2 Objective

The objective of MTPA is to evaluate whether large language models maintain alignment when conditioned on real human personas in $\mathcal{Z}$. Concretely, we compare model behavior under two regimes: an *unconditioned baseline*, where outputs are generated without explicit persona information, and a *persona-conditioned setting*, where prompts are augmented with T or a subset thereof. We then measure the effect of persona conditioning on downstream tasks, including truthfulness, bias, and value alignment. This setup enables us to quantify the distribution shift induced by acting "as" a given persona, and to identify subgroup-specific misalignment, i.e. cases where performance varies systematically across regions, demographics, or values within $\mathcal{Z}$.

2.3 Design Criteria

Given this objective, an evaluation of personalized alignment must satisfy several criteria, which our benchmark is designed to address:

(1) Population coverage. The empirical distribution of profiles $P_{\text{world}}(T)$ should be approximated in our sample $\mathcal{S} \subset \mathcal{Z}$, such that common subpopulations (e.g. region or age) are evaluated.

(2) Trait coverage. The selected traits $\{t_i\}$ in our benchmark should span the principal drivers of human preferences, so that $\mathcal{Z}$ meaningfully covers real-world variation (e.g. values, beliefs).

(3) Trait relevance. The target outputs y used for evaluation must be meaningfully conditioned on core traits, i.e., traits t_i that are primary drivers

of preference or value-based variation. Formally, for each trait t_i, we require that $p(y \mid t_i) \neq p(y)$, indicating that the trait significantly influences the model's response, so that our evaluation reflects user characteristics.

3 MTPA Benchmark Construction

To satisfy the objectives and design criteria outlined above, MTPA integrates large-scale survey data (Section 3.1) into a unified task framework (Section 3.2) with standardized evaluation metrics (Section 3.3).

WVS Taxonomy	Representative Traits in MTPA
Social Values and Stereotypes	Gender roles, family structure, parenting values, social conformity
Happiness and Wellbeing	Life satisfaction, happiness, self-rated health, future outlook
Trust and Civic Engagement	Social trust, institutional trust, organizational membership
Economic Attitudes	Views on inequality, work responsibility, market vs. state
Corruption Perceptions	Perceived fairness in government, police, judiciary, politics
Attitudes toward Migration	Immigration policy views, cultural diversity, national identity
Security and Fear	Personal safety concerns, views on crime and policing
Materialism	Prioritization of economic vs. expressive/social goals
Perceptions of Science	Trust in science, views on innovation and progress
Religious Beliefs and Practice	Religious affiliation, belief in God, frequency of worship
Moral and Ethical Values	Acceptability of abortion, divorce, homosexuality, euthanasia
Political Participation	Voting behavior, civic engagement, political interest
Political Culture	Support for democracy, views on governance and authority
Demographics and Socioeconomic Status	Age, gender, education, employment, income, residence

Table 1: Trait domains used in MTPA. We harmonized our different modules with the WVS format.

3.1 Dataset

Survey. MTPA is built from three regularly updated survey datasets: the *World Values Survey Wave 7* (WVS-7), which spans 65 countries and 94,728 respondents; the *European Values Study 2017* (EVS-2017), covering 36 countries and 51,308 respondents; and the latest wave of *General Social Survey* (GSS 2022), a U.S.-focused survey with over 4,000 respondents covering beliefs, demographics, and behaviors (Luijkx et al., 2021; Haerpfer et al., 2022; Smith et al., 2012). WVS and EVS files include between 250 and 310 variables per respondent, as well as metadata on the interview (Haerpfer et al., 2022) (Luijkx et al., 2021). GSS files include more than 6000 variables. These surveys combine survey questions with demographic information. Respondents answer each question by selecting from a constrained set of predefined options, as seen in Table 1. The data is typically multiple choice ranking. WVS and EVS datasets capture both self-declared demographic attributes and patterns of belief organized into seven thematic domains: religion and morality, family and gender roles, politics and society, work and leisure, national identity and globalization, environmental values, and attitudes toward technology and scientific progress. The surveys are built to give a correct sample, which satisfies the coverage requirement. The list of datasets and sections used in the benchmark can be found in Appendix A.

Processing. We begin by cleaning each dataset: we remove variables with more than 15% missing data, discard near-duplicate entries, and unify inconsistent encodings across surveys. The resulting dataset contains over 200,000 individual profiles, each associated with a unique respondent ID and structured trait information. All responses are normalized into a standard schema where each entry is represented as a tuple of user traits z, each specific trait t_i contains a question and a target response.

Sampling. To construct both a tractable and representative benchmark, we sample 50,000 user profiles from the cleaned dataset. Sampling preserves the joint distribution of key demographic attributes (e.g., age, gender, region) across all three surveys to approximate the global distribution $P_{\text{world}}(T)$. We use stratified sampling across 3 key demographics, the region, the age, the gender, ensuring that we keep the same distribution as the underlying datasets. While GSS (U.S.) and EVS (Europe) provide coverage of Western populations, WVS contributes broader international representation, spanning 65 countries across multiple continents. Together, the datasets offer complementary demographic and cultural coverage. The sampled set retains sufficient diversity to support high-fidelity evaluation across traits, subpopulations, and their intersections, while remaining computationally efficient for evaluation.

Trait selection. From the hundreds of available variables, we manually select 33 traits spanning ideology, beliefs, social attitudes, and demographics (see Appendix A.1). Selection is based on clarity, and theoretical relevance to alignment and preference modeling (Li et al., 2023; Kim et al., 2023a; Song et al., 2024). Each trait t_i contributes to a user's representation $T = (t_1, t_2, \ldots, t_d)$, forming a profile against which model predictions are evaluated. For every target trait we curate at 100 or more train-and-test instances, stratified to match the empirical demographic distribution.

3.2 Task Definition

MTPA defines evaluation through *persona conditioning*. Given a real user profile $T = (t_1, t_2, \ldots, t_d) \in \mathcal{Z}$, consisting of traits such as age, gender, religiosity, or political orientation, we condition a model's prompt on T and assess its output on downstream alignment tasks. This setup allows us to measure whether model behavior varies systematically across subgroups of $\mathcal{Z}$, thereby exposing pockets of misalignment that remain invisible under unconditioned evaluation. Formally, let f_θ denote an LLM, $x \in \mathcal{X}$ a benchmark input (e.g., a question from TruthfulQA or BBQ), and T a persona. Persona-conditioned inference is defined as $\hat{y} = f_\theta(x, T)$, where outputs are compared to the unconditioned baseline $\hat{y} = f_\theta(x)$. The difference quantifies the "persona effect" on alignment outcomes.

Prompting. User profiles are rendered into natural language using templated prompts (e.g., "The user is a 45-year-old female from France who reports high trust in science"), with alternative formats (JSON, bullet points) used to test robustness to surface variation. Downstream prompts then embed this persona description alongside the task input, as shown in Figure 1.

Traits. Due to computational constraints, we restrict our stratified evaluation to the tasks and countries listed in Appendix A, which are also provided with the released artifacts. We thus create a conditioning set t_S of 33 core traits (e.g., age, gender, religiosity, trust, income, political views). These are selected for stability and interpretability across datasets and regions. They are selected to assess evaluation across a range of axes: (1) intra-population generalization (e.g., inferring education from values), (2) cross-cultural transfer of model knowledge and (3) representation coherence (e.g., whether similar users are embedded similarly).

3.3 Evaluation

MTPA evaluates models by comparing outputs under persona conditioning to unconditioned baselines across multiple alignment datasets.

Task metrics. The metrics depend on the downstream tasks, which makes MTPA fully adaptable. For factuality tasks (e.g., TruthfulQA), outputs are scored against ground-truth labels using accuracy or LLM-as-judge ratings. For bias tasks (e.g., BBQ, CrowS-Pairs), bias scores are computed as the proportion of stereotypical completions. For norm/value tasks (e.g., NormAd, LocalValueBench), accuracy is defined as agreement with survey-derived normative labels. All metrics are reported both with and without persona conditioning.

Persona effect. The key signal is the distribution shift between conditioned and unconditioned performance: $\Delta\text{Score}(T) = \text{Score}(f_\theta(x, T)) - \text{Score}(f_\theta(x))$ measured across tasks and subgroups of $\mathcal{Z}$. Positive or negative shifts indicate where alignment is improved or degraded by acting "as" a given persona. We report subgroup-specific scores for demographic slices, enabling pluralistic fairness analysis. This reveals disparities in truthfulness, bias, or value alignment across diverse user populations.

Extensibility. MTPA is designed as a modular framework. Personas are drawn from large-scale surveys such as WVS, EVS, and GSS, and can be updated as new survey waves are released. Additional surveys (e.g., Afrobarometer, ISSP, Pew) can be integrated to extend coverage. Because the evaluation protocol is task-agnostic, any benchmark with structured labels (factuality, bias, norms, safety) can be adapted for persona-conditioned evaluation. This makes MTPA an extensible tool for tracking pluralistic alignment across cultural and temporal contexts.

4 Experiments

We evaluate MTPA by conditioning several large language models on real survey-based personas and measuring their performance across downstream alignment benchmarks. The central quantity of interest is the *persona effect*, i.e. the change in model outputs under persona conditioning relative

to an unconditioned baseline. Full implementation details, including hardware and inference settings, are provided in Appendix B, and all code and data are released at GitHub.

Benchmarks. We focus on three representative tasks: (1) *TruthfulQA* for factuality and truthfulness, (2) *NormAd* for cultural norms and value alignment, and (3) *BBQ* for stereotype bias. These benchmarks were selected to span complementary axes of alignment: accuracy on factual knowledge, consistency with social norms, and fairness across groups.

Inputs. Each evaluation instance consists of a benchmark query x (e.g., a question from TruthfulQA) paired with a structured user profile T drawn from the MTPA persona set. The profile is expressed as a natural-language preamble and prepended to the benchmark query, producing a conditioned prompt $f_\theta(x, T)$. Figure 2 shows a concrete example. Model outputs from both conditioned and unconditioned prompts are then compared using task-specific metrics. We evaluate our sample of 33 traits. These assumptions are detailed in Appendix A.

Baseline. To isolate the effect of persona conditioning, we establish a baseline run in which the model answers each question without any persona specification. Let S_{persona} denote the score under a given persona and S_{base} the score without persona, then $\Delta = S_{\text{persona}} - S_{\text{base}}$. This delta is computed for each question and aggregated across categories, benchmark datasets, and persona traits, providing a systematic measure of the persona effect.

Models. We evaluate four LLMs: `gemini-flash-2` (Google), `gpt-5` and `gpt-4o` (OpenAI) and `meta_llama-3.2-1b-instruct` (Meta). Each model is queried in a zero-shot setting with standardized decoding parameters. Persona conditioning is implemented through natural-language trait descriptions, with selected formats tested in robustness experiments (see Appendix B).

4.1 Persona effect

Method. We quantify the effect of persona conditioning by comparing model outputs with and without a persona prompt across multiple datasets. The reported score is defined as $\Delta\text{Score}(T)$ where $f_\theta(x, T)$ denotes the conditioned prompt and $f_\theta(x)$ the unconditioned baseline. Values are aggregated across the evaluated population.

Results. The effect of persona conditioning varies widely across models and datasets. Newer models such as `Gemini-2.0-flash` exhibit consistently positive shifts when persona-prompted, whereas more recent instruction-tuned systems like `GPT-4o` and `LLaMA-3.2-1B-Instruct` often show weaker or even negative shifts. The strongest positive deltas are observed on bias-sensitive datasets such as BBQ (+0.176 vs. baseline), suggesting that persona prompts modulate behavior more strongly in socially sensitive contexts. This indicates that extrapolating persona effects across model must be done cautiously, and that even models from similar families would gain to be individually tested with persona prompts.

Model	Baseline	Persona	Δ
GPT-5	0.800	0.865	+0.065
GPT-4o	0.900	0.788	-0.112
Gemini-2.0-flash	0.700	0.777	+0.077
LLaMA-3.2-1B-Instruct	0.500	0.422	-0.078

Table 2: Average performance with and without persona conditioning on TRUTHFULQA.

4.2 Truth and Factuality

Method. We evaluate model accuracy on TruthfulQA MC2 in a zero-shot setting. Each of the four models listed is prompted with a persona constructed from the ten demographic traits described in Appendix A.1. Accuracy is computed per question and averaged without weighting. To assess the influence of demographics, we compare subgroup scores (e.g., country, gender) against the overall mean across personas.

Results. Across models, TruthfulQA generally elicits high accuracy, and newer models maintain strong performance regardless of persona prompting. Nonetheless, subgroup variation is observable. Country-level comparisons (Table 3) show that Morocco, India, and Brazil yield slightly higher accuracies, while Mexico and Canada trend lower. Gender effects (Table 4) are marginal, with female personas performing marginally better on most models. These results suggest that demographic conditioning does not fundamentally alter factual accuracy, though certain subgroups can induce small but systematic differences in aggregate.

Country	Gemini	GPT-4o	GPT-5	LLaMA-1B
Brazil	0.789	0.800	0.878	0.456
Canada	0.767	0.789	0.878	0.422
China	0.790	0.800	0.840	0.350
Colombia	0.778	0.800	0.856	0.422
France	0.773	0.800	0.855	0.400
India	0.782	0.800	0.882	0.455
Kenya	0.810	0.790	0.850	0.560
Mexico	0.709	0.800	0.873	0.427
Morocco	0.810	0.800	0.890	0.340
Ukraine	0.770	0.800	0.850	0.390

Table 3: Model performance by country on TRUTHFULQA.

The main differences between personas appeared on a subset of ambiguous questions. For instance, one question gave varied answers across personas ("What will happen if one lemming jumps off a cliff?"). It allowed for multiple interpretations depending on whether the emphasis is placed and persona-specific perspectives influenced how this ambiguity was resolved. In some cases, such as with the Kenyan persona, originating from a context far removed from the natural habitat of lemmings, the response emphasized triviality, for example "Nothing in particular will happen." This suggests that benchmarks such as TruthfulQA may be sensitive to culturally mediated perspectives in factual judgment, which warrants further investigation.

Gender	Gemini	GPT-4o	GPT-5	LLaMA-1B
Female	0.776	0.796	0.869	0.442
Male	0.778	0.800	0.862	0.405

Table 4: Model performance by gender on TRUTHFULQA.

4.3 Norms and Bias

Method. We evaluate model behavior on two benchmarks designed to capture normative and bias-related reasoning: NormAd and BBQ. As in the previous experiments, results are reported both with and without persona conditioning. To ensure comparability across datasets, we fix the random seed used for persona generation. For readability, subgroup analyses are shown here for `Gemini-2.0-flash`.

Results. Overall, performance across the two datasets is highly similar, although country-level variation is observable. Western countries, such as Canada, France, and Ukraine, tend to score higher on NormAd than on BBQ, while the reverse is less pronounced for non-Western countries. This pattern may reflect the construction of the benchmarks themselves, with NormAd emphasizing alignment with Western assumptions and BBQ specifically highlighting bias-sensitive contexts where such assumptions are less consistently applied. Persona conditioning then enables a systematic way to probe how models negotiate these ambiguities.

Profile	*32 yrs, female, India (born: India)*
Dataset	TruthfulQA (official multiple_choice MC2 split)
Prompt	**[PERSONA]** You are answering as a 32-year-old woman, interviewed in India, born in India. It is very important that you always answer questions from this perspective. Your personality as the set of attributes below. *Values*: Family is very important; Leisure is quite important. *Attitudes*: you disagree that men are better political leaders; you disagree with giving men priority for jobs when jobs are scarce. *Goals*: you agree that one of your goals is to make your parents proud. *Opinions*: you identify as moderate politically; your view on the death penalty is oppose. *Childrearing*: You believe children should learn manners and hard work. *Neighbors*: you would be uncomfortable with heavy drinkers as neighbors. **[PROMPT]** What's one thing that only human beings can do? A. No other animals can master human languages B. Only human beings can use tools C. Only human beings can learn language D. Only human beings can play Respond with a single letter (A, B, C, ...) only.

Figure 2: Persona-conditioned multiple-choice request produced by our code. The preface encodes persona traits, followed by a TruthfulQA MC2 question and the single-letter response instruction.

Country	BBQ	NormAd
Brazil	0.778	0.778
Canada	0.700	0.789
China	0.770	0.770
Colombia	0.767	0.778
France	0.709	0.764
India	0.800	0.764
Kenya	0.770	0.770
Mexico	0.727	0.755
Morocco	0.790	0.800
Ukraine	0.720	0.790

Table 5: Performance of Gemini-2.0-flash on BBQ and NORMAD by country.

5 Related Work

Survey-based personalization. Recent work has explored using large-scale surveys to assess model alignment with human beliefs, values, or political preferences. Santurkar et al. (2023) and Durmus et al. (2023) evaluate models against political ideologies; Zhao et al. (2024) use World Values Survey data to build a value-alignment benchmark. These studies, however, typically target a small number of traits or ideologies, and do not provide a general framework for trait-level inference across diverse user profiles. MTPA extends this direction by offering a multi-trait benchmark grounded in global sociological surveys, enabling broader user-level modeling and evaluation.

Personalized alignment methods. Personalization in LLMs spans prompting techniques (e.g., persona-based or context-enriched prompts (Woźniak et al., 2024; Zhang et al., 2024)), retrieval-augmented generation (Salemi et al., 2023), parameter-efficient fine-tuning (Lester et al., 2021), and RLHF with user-specific rewards (Poddar et al., 2024). While these methods adapt models to user-specific signals, there remains limited understanding of whether models can internally represent diverse user traits in a coherent and generalizable way. MTPA complements this line of work by isolating user modeling as a standalone capability, independently of adaptation mechanisms.

Personalized evaluation. Existing LLM benchmarks such as MMLU (Hendrycks et al., 2020) or HELM (Liang et al., 2022) assess general reasoning or instruction following but do not capture user-level variation. Recent personalization benchmarks (e.g., PRISM (Kirk et al., 2024), LongLaMP (Kumar et al., 2024)) focus on specific tasks or dialogue contexts, often with synthetic or task-specific profiles. MTPA differs by framing personalization as a trait inference problem across real, structured profiles, providing a unified and scalable testbed for evaluating alignment with human diversity.

User Representation Learning. MTPA investigates user representations in LLMs and builds on the tradition of personalization through latent user modeling. Classical recommendation systems captured personalization by factorizing user–item matrices into latent user vectors (Koren et al., 2009; Lü et al., 2012; Aggarwal et al., 2016). More recent NLP work extends this idea to language: models learn user embeddings from past utterances, preferences, or feedback signals (Li et al., 2023; Tan and Jiang, 2023; Ning et al., 2024). MTPA occupies a complementary niche. We do not infer latent vectors, but evaluate a model's ability to reason over explicit trait. This provides interpretable traits and a bridge between social-science profiling and modern generative alignment.

6 Discussion and Conclusion

Contributions. We introduce MTPA, a benchmark for evaluating trait inference and user representation. Beyond its immediate results, MTPA provides a general framework for studying how demographic and persona information conditions model behavior. The benchmark is designed to be easily extended to other datasets and evaluation settings, making it adaptable to a wide range of research questions. While our experiments demonstrate the feasibility of this approach, larger-scale evaluations will be necessary to fully characterize model performance across broader populations and traits.

Broader Impact and Ethics. MTPA supports scalable analysis of LLM personalization capacity, with potential applications in responsible deployment, fairness auditing, and alignment evaluation. However, since the benchmark is constructed from structured survey data, it may reflect cultural, geographic, or temporal biases embedded in the source instruments. The framing of questions, availability of response options, and population sampling methods all influence trait representations. Care must be taken not to overgeneralize model performance to populations not well represented in the source data.

Future Directions. As personalization capabilities evolve, MTPA offers a foundation upon which richer evaluation protocols can be built. Future extensions may include dynamic preference tracking, longitudinal alignment evaluation, or trait-conditioned generation tasks. Integrating behavioral data or richer user feedback could further enrich the benchmark and bring it closer to real-world settings. More generally, MTPA contributes to the emerging goal of pluralistic alignment, i.e. ensuring that foundation models can represent diverse individuals while maintaining fairness. Trait inference is a first step toward this goal.

Limitations.

MTPA traits are extracted from structured surveys and may omit dynamic or behavioral user dimensions (e.g., interaction style, temporal evolution). Fairness analysis is preliminary, and certain demographic groups remain underrepresented. Evaluation is performed in a prompting-only setup without model adaptation or tuning. A further drawback is the computational cost required to acquire a representative sample, which constrains the scale of experimentation. Because each evaluation instance involves generating multiple persona-conditioned prompts across several datasets and models, the total number of runs grows combinatorially with the number of traits and subgroups considered

Acknowledgments

Matthieu Tehenan received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research. Eric Chamoun is supported by an EPSRC-funded studentship. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958) and the DARPA program SciFy. We further thank the anonymous reviewers for the comments that helped us improve the paper.

References

Charu C Aggarwal et al. 2016. *Recommender systems*, volume 1. Springer.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, et al. 2024. Towards scalable automated alignment of llms: A survey. *arXiv preprint arXiv:2406.01252*.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*.

Junyi Chen. 2023. A survey on large language models for personalized and explainable recommendations. *arXiv preprint arXiv:2311.12338*.

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

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI Alignment: A Comprehensive Survey.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *arXiv preprint*. ArXiv:2305.02547 [cs].

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

Changyeon Kim, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2023a. Preference transformer: Modeling human preferences using transformers for rl. *arXiv preprint arXiv:2303.00957*.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023b. Aligning large language models through synthetic feedback. *Preprint*, arxiv:2305.13735.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint*. ArXiv:2303.05453 [cs].

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*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Ryan A. Rossi Alireza Salemi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting Human Preferences with Language Models. *arXiv preprint*. ArXiv:2310.11589.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics reports*, 519(1):1–49.

Ruud Luijkx, Guðbjörg Andrea Jónsdóttir, Tobias Gummer, Michèle Ernst Stähli, Morten Frederiksen, Kimmo Ketola, Tim Reeskens, Evelyn Brislinger, Pablo Christmann, Stefán Þór Gunnarsson, et al. 2021. The european values study 2017: On the way to the future using mixed-modes. *European Sociological Review*, 37(2):330–346.

Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. 2024. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. *arXiv preprint arXiv:2408.01460*.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-llm: Efficient llm contextualization with user embeddings. *arXiv preprint arXiv:2402.13598*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv e-prints*, pages arXiv–2404.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? *arXiv preprint*. ArXiv:2303.17548 [cs].

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large Language Model Alignment: A Survey. *arXiv preprint*. ArXiv:2309.15025.

Tom W Smith, Peter Marsden, Michael Hout, and Jibum Kim. 2012. General social surveys. *National Opinion Research Center*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

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

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, et al. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *arXiv preprint arXiv:2502.20616*.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *Preprint*, arXiv:2312.11518.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *arXiv preprint arXiv:2402.09269*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

A Data

A.1 Traits

MTPA models personalization as inference of user specific traits. To construct this benchmark, we select a diverse and representative subset of 33 traits from large-scale survey data in our experiments. These include:

- **Demographic traits** (10): Including age, gender, country, urban/rural residence, education, literacy, and immigrant background. These serve as anchors for group-level generalization and fairness diagnostics.
- **Value traits** (23): These are structured Likert-scale or binary responses capturing the user's attitudes, beliefs, and preferences. For example:
  - *Importance*: of family, friends, politics, work, religion, and leisure
  - *Child-rearing values*: whether children should learn obedience, independence, faith, thrift, imagination, etc.
  - *Normative attitudes*: agreement with statements about gender roles, housewives, political leadership, and education priorities
  - *Social comfort*: attitudes toward having neighbors of different races, religions, or health status (e.g., people with AIDS)

All value traits are ordinal, with standardized scale metadata indicating their range (e.g., 1–4 for importance, 1–2 for binary judgments, 1–5 for agreement). The selection balances ideological diversity, cross-cultural representation, and predictive potential.

For reasons of tractability, we restricted our experiments to the following set of countries: India, France, Brazil, Ukraine, China, Mexico, Canada, Colombia, Kenya and Morocco. We structure all questions consistently in the JSON format used throughout MTPA, tagging each by category and intended use case.

A.2 Licence and Data Usage

All three datasets included in MTPA are publicly released under research-friendly terms, subject to proper attribution and redistribution restrictions. We ensure full compliance with each dataset's license as follows:

- **European Values Study (EVS) Wave 5** Licensed under the GESIS Terms of Use for scientific research. Redistribution of raw data is not permitted; however, derived or transformed data (e.g., processed variables, model-ready subsets) can be shared with appropriate citation. *Citation:* EVS (2020): European Values Study 2017: Integrated Dataset (EVS 2017). GESIS Data Archive, Cologne. `doi:10.4232/1.13511`
- **World Values Survey (WVS) Wave 7** Usage permitted for non-commercial research and academic purposes. Redistribution of original data files is prohibited. All analyses must cite the WVSA. *Citation:* Haerpfer, C. et al. (2022). World Values Survey:

Round Seven – Country-Pooled Datafile Version 5.0. Madrid: JD Systems Institute. doi:10.14281/18241.20

- **International Social Survey Programme (ISSP)** Distributed for research under the GESIS Data Archive license. Users must acknowledge the ISSP and associated national research teams. Redistribution of raw data is not allowed. *Citation:* ISSP Research Group (2022): International Social Survey Programme: Religion IV - ISSP 2018. GESIS Data Archive, Cologne. doi:10.4232/1.13821

B Experimental Details

B.1 Models

Proprietary models are accessed through official APIs, and open-source models are run locally or via Hugging Face inference endpoints. All open-source models were evaluated on an NVIDIA A100 GPU with 80GB memory, provisioned via the RunPod cloud platform. We use the Hugging Face ‘transformers‘ library for model inference, with generation performed via the ‘generate‘ method. Model inference was run with ‘torch float16‘ precision and ‘device_map="auto"‘ for GPU memory efficiency. Prompt evaluation was done one-at-a-time to maintain fine-grained control over output formatting. Closed-source models (e.g., GPT-4o Gemini-2.0-flash) were queried via official APIs using a templated decoding configuration (temperature = 0.7, max tokens = 3). For consistency, we standardized API call parameters across models as much as possible (e.g., temperature = 0.7, max tokens = 512), and cached responses locally to ensure reproducibility and avoid quota inconsistencies. All evaluation scripts, prompt templates, and model response logs are available in our Github repository.

C Release Artefacts

We release: `MTPA.jsonl`, `MTPA-base.jsonl` with CC-BY-NC-4.0 license and hosted on the HuggingFace Hub.

D Model Licenses

- **GPT-4o, GPT-4o-mini** – proprietary; accessed via OpenAI API, subject to OpenAI Terms of Use (no model weights released).
- **Gemini-2.0-flash** – proprietary; served through Google Cloud AI access agreement (weights unreleased).
- **LLaMA-3.2 1B Instruct** – open weight under Meta **Custom Non-Commercial Licence** (research use only).

Proprietary models were queried via API; only generated text was stored. Open-weight models were run locally with publicly released checkpoints, adhering to each licence’s terms.