AlignSurvey: A Comprehensive Benchmark for Human Preferences Alignment in Social Surveys

Chenxi Lin[1], Weikang Yuan[1], Zhuoren Jiang[1,3*], Biao Huang[1,3], Ruitao Zhang[1], Jianan Ge[1], Yueqian Xu[2,3], Jianxing Yu[2,3]

[1]Zhejiang University
[2]Zhejiang Gongshang University
[3]Laboratory for Statistical Monitoring and Intelligent Governance of Common Prosperity
linchenxi@zju.edu.cn, yuanwk@zju.edu.cn, jiangzhuoren@zju.edu.cn, biaohuang@zju.edu.cn, zhangruitao24@zju.edu.cn, iamjianange@zju.edu.cn, xuyueqian@zjgsu.edu.cn, yujianxing@zju.edu.cn

Abstract

Understanding human attitudes, preferences, and behaviors through social surveys is essential for academic research and policymaking. Yet traditional surveys face persistent challenges, including fixed-question formats, high costs, limited adaptability, and difficulties ensuring cross-cultural equivalence. While recent studies explore large language models (LLMs) to simulate survey responses, most are limited to structured questions, overlook the entire survey process, and risks under-representing marginalized groups due to training data biases. We introduce **AlignSurvey**, the first benchmark that systematically replicates and evaluates the full social survey pipeline using LLMs. It defines four tasks aligned with key survey stages: social role modeling, semi-structured interview modeling, attitude stance modeling and survey response modeling. It also provides task-specific evaluation metrics to assess alignment fidelity, consistency, and fairness at both individual and group levels, with a focus on demographic diversity. To support AlignSurvey, we construct a multi-tiered dataset architecture: (i) the Social Foundation Corpus, a cross-national resource with 44K+ interview dialogues and 400K+ structured survey records; and (ii) a suite of Entire-Pipeline Survey Datasets, including the expert-annotated AlignSurvey-Expert (ASE) and two nationally representative surveys for cross-cultural evaluation. We release the SurveyLM family, obtained through two-stage fine-tuning of open-source LLMs, and offer reference models for evaluating domain-specific alignment. All datasets, models, and tools are available at github and huggingface to support transparent and socially responsible research.

Code — https://github.com/PiLab-ZJU/AlignSurvey
Datasets & Models — https://huggingface.co/PiLabZJU
Extended Version with Appendix — https://arxiv.org/abs/2511.07871

Introduction

Understanding human preferences, attitudes, and behaviors is central to both academic research and evidence-based policymaking. Social surveys, spanning qualitative interviews and quantitative questionnaires, have long served as a critical tool in this endeavor (Wright, Marsden et al. 2010; Tourangeau 2004), with an estimated $35.1 billion spent annually on survey research worldwide (ESOMAR 2024).

Despite this scale, traditional surveys face persistent challenges: fixed-question formats limit adaptability (Heffetz and Reeves 2019); high costs often force sampling compromises, introducing biases (Kalton 2009); slow turnaround hinders responsiveness to emerging issues (Moy and Murphy 2016; Evans and Mathur 2018; Prosser and Mellon 2018); and cross-cultural equivalence remains difficult despite translation efforts (Tsai et al. 2025).

Large language models (LLMs) offer a promising alternative. By learning from vast public corpora, LLMs can simulate human responses and reduce the burden of manual data collection (Thapa et al. 2025; Mellon et al. 2024; Zhang et al. 2025). However, their outputs often reflect the preferences of digitally active, well-educated users (Giorgi et al. 2025; Abeliuk, Gaete, and Bro 2025), reinforcing representational bias (Wang, Morgenstern, and Dickerson 2025; Hu et al. 2025; Hofmann et al. 2024) and marginalizing rural, low-income, or elderly populations. Group-level disparities are often underexplored, and many existing works lack systematic evaluation across demographic subgroups.

Existing benchmarks primarily focus on fixed-option quantitative tasks (Ji et al. 2023; Liu et al. 2025; Zhou et al. 2025; Lee et al. 2024), overlooking the full pipelines of professional surveys, such as qualitative interviewing and context-aware reasoning. This critical gap calls for a comprehensive framework that aligns LLMs with human preferences across the entire survey process.

To address these gaps, we introduce **AlignSurvey**, the first benchmark designed to systematically replicate and evaluate the **full pipeline** of professional social surveys using LLMs. AlignSurvey mirrors four core stages of professional social surveys (Ahmed, Pereira, and Jane 2024; Fetters, Curry, and Creswell 2013) by mapping role exploration, qualitative survey, attitude mining and quantitative survey to corresponding modeling tasks: *Social Role Modeling*, *Semi-structured Interview Modeling*, *Attitude Stance Modeling* and *Structured Response Modeling*. We introduce task-specific evaluation metrics that enable alignment assessment across tasks at individual and group levels, with

*Corresponding author

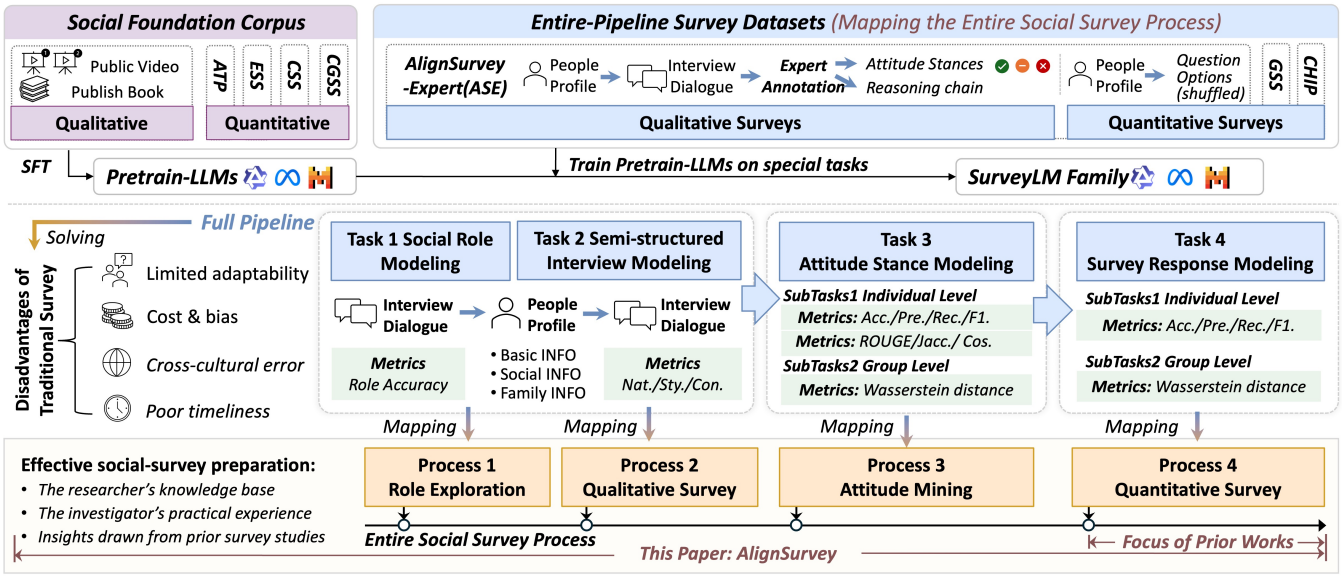

Figure 1: Overview of the **AlignSurvey**. AlignSurvey is a four-stage benchmark that mirrors the professional social survey process. The upper panel depicts a multi-tiered dataset: we pretrain on the Social Foundation Corpus and fine-tune on the Entire-Pipeline Survey Dataset for the four survey stages. AlignSurvey is the first to align LLMs across the entire social science survey, surpassing prior work limited to structured responses.

a particular focus on fairness and demographic diversity.

To support this pipeline, we construct a multi-tiered dataset architecture. The first component, the **Social Foundation Corpus**, is a cross-national resource comprising 44,000+ interview dialogues collected from publicly accessible video platforms and oral history books, and 400,000+ structured records from four authoritative surveys: ATP, ESS, CSS, and CGSS. These corpora provide foundational knowledge across diverse socio-cultural contexts, covering domains such as family dynamics, civic engagement, and inequality. We further introduce the **Entire-Pipeline Survey Datasets** for task-specific supervision. At the core is **AlignSurvey-Expert (ASE)**, an expert-annotated dataset consisting of 161 semi-structured interviews and 1,679 questionnaires, comprising 2,500+ dialogues and 16,000+ responses, and enriched with detailed demographic metadata and annotated reasoning chains. To enable cross-national validation, we include two nationally representative datasets: the U.S.-based GSS and the China-based CHIP.

Based on AlignSurvey, we fine-tune three representative open-source LLMs to develop the **SurveyLM** model family. The training follows a two-stage alignment strategy: we first adapt base model on the Social Foundation Corpus to equip it with general social knowledge and discourse patterns, then fine-tune them on each task using the Entire-Pipeline Survey Datasets. These models serve as strong reference baselines for evaluating alignment across survey tasks.

Experiments demonstrate that general-purpose LLMs fail to reliably reproduce survey outcomes, particularly for underrepresented groups. In contrast, SurveyLM models yield substantial improvements, 10–20% gains in demographic alignment and stance prediction, highlighting the need for domain-specific adaptation in social applications of LLMs.

Our contributions can be summarized as follows:

- We introduce AlignSurvey, the first benchmark that systematically replicates the full pipeline of professional social surveys using LLMs, including tasks for social role modeling, interview simulation, attitude modeling, and structured response prediction.
- We construct a multi-tier dataset architecture comprising (i) the Social Foundation Corpus (44K+ interviews, 400K+ survey records), and (ii) Entire-Pipeline Survey Datasets for full-pipeline alignment, enabling robust, cross-national evaluation and task-specific supervision.
- We release the SurveyLM model family, trained via a two-stage alignment strategy, as reference models for evaluating task-specific alignment across the survey pipeline.
- We contribute AlignSurvey-Expert (ASE), a multi-tiered dataset comprising expert-annotated interviews and thematic survey responses. It includes attitudinal stances, reasoning chains and demographic profiles, supporting supervision and evaluation across entire social survey process.
- All datasets, models, and evaluation code will be released at github and huggingface to support transparent, reproducible, and socially impactful research.

Related Work

Simulating Survey Respondents with LLMs. To reduce the cost and rigidity of traditional survey methods (Wright, Marsden et al. 2010; Heffetz and Reeves 2019; Kalton 2009), recent research explores using LLMs as virtual respondents (Zhou, Li, and Yu 2024; Zhu, Huang, and Sang 2025; Wang et al. 2024; Li et al. 2024; Santurkar et al. 2023b; Abdurahman et al. 2024a). Most approaches fall

Benchmark	Type	Size	Source	Demographic	Multi-Country	Availability
Psychology-related						
PhDGPT (De Duro et al. 2024)	Quant	756K	Synthetic	✓	✗	Dataset
Psych-101 (Binz et al. 2024)	Quant	10M	Public	✗	✓	Model + Dataset
Scenario (Cui, Li, and Zhou 2025)	Quant	/	Public	✗	✗	Dataset
Social Survey-related						
OpinionQA (Santurkar et al. 2023a)	Quant	80K	Public	✓	✗	Dataset
Anthology (Moon et al. 2024)	Quant	10K	Public	✓	✗	Dataset
SubPOP (Suh et al. 2025)	Quant	73K	Public	✓	✗	Dataset
AlignSurvey (Ours)	Qual + Quant	600K	Expert + Public	✓	✓	Model + Dataset

Table 1: Overview of benchmarks related to psychology and social surveys. Abbreviations are explained in the extended version.

into prompt engineering, that simulates demographic variation via persona-based instructions (Santurkar et al. 2023b; Hwang, Majumder, and Tandon 2023; Simmons 2022; Kim and Yang 2024; Sun et al. 2024; Moon et al. 2024) and fine-tuning on real-world corpora to model individual or group preferences (Chu et al. 2023; He et al. 2024; Kwon et al. 2023; Zhao, Dang, and Grover 2023; Suh et al. 2025). However, most of these efforts focus on isolated tasks and rarely address qualitative interviewing or full survey pipelines.

Benchmarks for Aligning LLMs with Social Survey Data. While aligning LLMs with human preferences remains a core challenge (Kopf et al. 2023; Aroyo et al. 2023; Lambert et al. 2024; Ethayarajh et al. 2024), benchmarks focused on social surveys are limited. Table 1 summarizes representative datasets across psychology and survey domains. Psychology-related resources have grown rapidly, often relying on synthetic scenarios to model cognition and responses (Binz and Schulz 2023; Binz et al. 2024; Abdurahman et al. 2024b; Dominguez-Olmedo, Hardt, and Mendler-Dünner 2024). However, they mostly focus on individual cognition, lack real-world demographic variation, and do not model full survey workflows. Survey-focused benchmarks such as OpinionQA, Anthology, and SubPOP (Kirk et al. 2024; Santurkar et al. 2023a; Moon et al. 2024; Suh et al. 2025) target structured, fixed-option questions, typically in U.S. contexts. They overlook qualitative interviewing, reasoning chains, and cross-cultural validity. None of these benchmarks replicates the entire process of professional social surveys or enables integrated evaluation across qualitative and quantitative components with expert annotations.

AlignSurvey

Design Principle

AlignSurvey is designed to systematically evaluate whether large language models can replicate the entire social survey process. As shown in Figure 1, the typical process of a professional social survey comprises four stages (Ahmed, Pereira, and Jane 2024; Fetters, Curry, and Creswell 2013): (1) Role Exploration: finding investigation targets, (2) Qualitative Survey: conducting semi-structured interviews, (3) Attitude Mining: extracting attitude stances and reasoning chains from dialogues, and (4) Quantitative Survey: collecting structured responses. AlignSurvey mirrors this process into four modeling and evaluation tasks: *Social Role Modeling*, *Semi-structured Interview Modeling*, *Attitude Stance Modeling*, and *Structured Response Modeling*. The Attitude and Response Modeling tasks include individual- and group-level subtasks to enable analysis of model alignment across multiple levels of social understanding. AlignSurvey provides stage-specific metrics to diagnose model strengths and limitations to support both rigorous benchmarking and responsible use in empirical social research.

ID	Task	Train	Test
Task1	Social Role	8 712	2 880
Task2	Semi–structured Interview	1 904	632
Task3.1	Attitude Stance(Individual)	13 239	3 338
Task3.2	Attitude Stance(Group)	108	36
Task4.1	Structured Response(Individual)	54 180	12 047
Task4.2	Structured Response(Group)	46 815	8 674

Table 2: Details of tasks within AlignSurvey.

Data Collection and Processing

AlignSurvey builds a multi-tiered dataset that combines large-scale public data and expert-curated resources to support comprehensive social contextual grounding and full pipeline alignment for LLM-based social surveys process.

Social Foundation Corpus. This corpus provides foundational training to equip models with broad social knowledge and cultural patterns before task-specific alignment. It comprises two components:

Qualitative Corpus contains 44,021 structured interview dialogues from publicly accessible video platforms and books, spanning diverse national and cultural contexts. Dialogues are segmented into conversational turns, with a 5-turn sliding window used to predict the next response.

Quantitative Corpus comprises 411,174 records from four authoritative cross-national surveys: the American Trends Panel (ATP), European Social Survey (ESS), Chinese Social Survey (CSS), and Chinese General Social Survey (CGSS). These datasets cover topics such as public

opinion, trust, inequality, and civic engagement. We selected five waves of ATP (40, 41, 54, 81, 103), Round 11 of ESS, and the latest releases of CSS and CGSS. To mitigate response biases (e.g., label-position bias (Dominguez-Olmedo, Hardt, and Mendler-Dünner 2024)), option labels and choices were randomly shuffled.

Entire-Pipeline Survey Datasets. As existing datasets seldom support full-pipeline alignment, we compile this suit of datasets, which comprises (1) AlignSurvey-Expert, an expert-annotated dataset, and (2) two national surveys from the U.S. and China.

AlignSurvey-Expert (ASE) is the core dataset for supervising and evaluating model performance across all four survey stages. It includes:

Qualitative Component includes 161 semi-structured interviews that comprise 2,500+ dialogues conducted by 15 social scientists, and each paired with rich demographic metadata (e.g., age, household size, occupation). The interviews are organized around eight core questions reflecting social perception. The first four gather general topic awareness, whereas the latter four target themes of service quality, social mobility, future expectations, and policy preferences. Each interview is divided into theme-specific dialogue segments, each comprising a sequence of utterances reflecting the respondent's views. Each segment is annotated by six domain experts with an attitude label (positive, neutral, or negative) and a reasoning chain explaining the stance. See Appendix F for annotation details and reliability. These chains capture underlying logic by incorporating contextual factors such as life experience, group identity, and perceived fairness. Grounded in demographic profiles and dialogue content, they enable transparent, interpretable alignment evaluation in assessing subgroup-level fidelity, and serve as key supervision signals for training attitude models.

Quantitative Component includes 1,679 questionnaires that comprise 16,000+ responses designed around the same themes as the interviews. These were collected via an online platform, with rigorous screening applied to ensure data quality. Each question contains multiple labeled answer choices (e.g., A/B/C), which were randomly shuffled to mitigate position bias. The correct answer for each question is explicitly marked to support supervised learning and evaluation in structured response modeling.

Supplementary National Datasets. To support robustness checks and enhance cross-cultural generalization, we include two complementary national questionnaire datasets: the General Social Survey (GSS) from the U.S. and the China Household Income Project (CHIP). Both underwent the same preprocessing procedures as the questionnaire corpus of **ASE**, including label shuffling and standardized formatting, enabling consistent cross-dataset evaluation.

See Appendix A for detailed dataset information and Appendix G.1 for licensing details.

Task Definition

AlignSurvey defines four core tasks aligned with key stages of the social survey pipeline, each guided by tailored prompts and evaluation metrics to ensure faithful modeling and assessment.

Task 1: Social Role Modeling. This task evaluates whether an LLM can predict an interviewee's demographic attribute (e.g., gender, education level) $c_i \in \boldsymbol{c}$ from a theme-specific dialogue segment $D^{(t)}$, where t denotes the theme. The model prediction is: $\hat{c}_i = f_{LLM}(D^{(t)})$. Model performance is measured by accuracy: the proportion of predicted $\hat{c}_i$ matching the ground-truth c_i:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}[\hat{c}_i = c_i],$$

where $\mathbf{I}[\cdot]$ is the indicator function.

Task 2: Semi-structured Interview Modeling. This task evaluates whether the model can generate coherent interview responses conditioned on a full demographic profile $\boldsymbol{c} = (c_1, c_2, .., c_i)$ and recent dialogue history $\mathcal{H}_\tau^{(t)}$, the last $\tau = 5$ utterances of a dialogue $D^{(t)}$. The model prediction is: $\hat{d}_{\text{next}}^{(t)} = f_{LLM}(\boldsymbol{c}, \mathcal{H}_\tau^{(t)})$. The generated response is evaluated along three dimensions: (i) *Naturalness* (fluency and topical coherence), (ii) *Style Match* (alignment with interview tone), and (iii) *Consistency* (logical and factual coherence).

Formally, let $\mathcal{J}$ be the set of evaluators [1], and $s_{j,k}^{(d)}$ be the score assigned by evaluator $j \in \mathcal{J}$ on dimension k for dialogue d. The final score is computed as:

$$S_k^{(d)} = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} s_{j,k}^{(d)}.$$

Task 3: Attitude Stance Modeling. This task involves predicting both individual- and group-level attitude stances $\mathcal{A}$ and reasoning chains $\mathcal{R}$.

Individual-level. Given individual demographic profile $\boldsymbol{c}$ and theme t, the model predicts the attitude stance and reasoning chain: $(\hat{a}^{(t)}, \hat{r}^{(t)}) = f_{LLM}(\boldsymbol{c}, t)$. Predictions are evaluated using macro-averaged accuracy, precision, recall, and F1 (for attitude stances), and ROUGE, Jaccard, and cosine similarity (for reasoning chains).

Group-level. For each question, the model produces individual stance predictions $\hat{a}i^{(t)}i = 1^N$ for respondents belonging to a demographic group t. We convert these predictions into an empirical stance distribution by counting label frequencies: $\hat{\mathbf{p}} = \text{Aggregate}(\{\hat{a}_i^{(t)}\}), \quad \hat{\mathbf{p}} \in \Delta^{|\mathcal{A}|-1}$. The predicted distribution is compared with the reference distribution $\mathbf{p}$ using the Wasserstein distance:

$$W_1(\hat{\mathbf{p}}, \mathbf{p}) = \min_{\Gamma\geq 0, \Gamma\mathbf{1}=\hat{\mathbf{p}}, \Gamma^\top\mathbf{1}=\mathbf{p}} \sum_{u,v\in\mathcal{A}} \Gamma_{uv}C_{uv},$$

where C_{uv} denotes the ground distance between stance labels $u, v \in \{\text{positive}, \text{neutral}, \text{negative}\}$.

[1] For Task 2 evaluation, to reduce bias and improve robustness, we use multiple strong LLMs (e.g., GPT-4o, DeepSeek-R1, Qwen-Max) as automated evaluators, with human validation confirming high agreement and consistency (see Appendix E.3).

Model	Task 1			Task 2			Task 3		
	$\text{Acc}_{\text{Basic}}$ ↑	$\text{Acc}_{\text{Social}}$ ↑	$\text{Acc}_{\text{Family}}$ ↑	Nat.↑	Sty.↑	Cons.↑	Acc.↑	Jaccard↑	WD↓
GPT-4o *(Zero)*	44.08%	35.28%	3.85%	3.45	2.85	2.89	38.12%	0.099	1.648
GPT-4o *(Few)*	40.59%	36.58%	17.43%	3.49	2.80	2.75	42.73%	0.103	1.180
Claude 3.7 Sonnet *(Zero)*	42.72%	34.17%	4.06%	3.41	2.81	2.85	28.12%	0.064	1.332
Claude 3.7 Sonnet *(Few)*	40.15%	33.18%	14.34%	3.48	2.83	2.90	31.48%	0.072	1.120
DeepSeek-R1 *(Zero)*	38.52%	34.31%	6.56%	3.40	2.81	2.83	36.88%	0.075	1.356
DeepSeek-R1 *(Few)*	36.36%	30.53%	18.27%	3.47	2.70	2.78	41.05%	0.081	1.100
R1-Distill-Qwen-14B *(Zero)*	30.25%	27.08%	2.81%	2.78	2.20	2.36	38.75%	0.105	1.795
R1-Distill-Qwen-14B *(Few)*	34.72%	31.18%	15.87%	2.85	2.25	2.30	33.81%	0.111	1.140
Qwen2.5-72B *(Zero)*	37.83%	33.34%	3.12%	3.40	2.72	2.85	35.00%	0.102	1.653
Qwen2.5-72B *(Few)*	38.65%	32.75%	16.90%	3.46	2.65	2.70	42.16%	0.113	1.050
Mistral-7B-v0.3 *(Zero)*	32.58%	35.14%	5.10%	3.21	2.57	2.44	38.44%	0.106	1.534
Mistral-7B-v0.3 *(Few)*	39.31%	32.73%	16.53%	3.28	2.50	2.40	39.77%	0.114	0.302
Meta-Llama-3.1-8B *(Zero)*	37.00%	33.89%	8.33%	3.30	2.69	2.79	39.38%	0.107	1.521
Meta-Llama-3.1-8B *(Few)*	35.42%	32.52%	16.02%	3.36	2.60	2.68	39.96%	0.119	0.447
Qwen2.5-7B *(Zero)*	36.08%	33.89%	2.71%	3.39	2.67	2.79	36.88%	0.104	1.700
Qwen2.5-7B *(Few)*	27.92%	37.83%	16.76%	3.44	2.60	2.85	44.69%	0.118	0.372
$SurveyLM_{\text{Mistral-7B-v0.3}}$	48.24%**	52.23%**	47.08%**	2.61	2.90*	2.90*	**57.13%****	0.131**	**0.297****
$SurveyLM_{\text{Meta-Llama-3.1-8B}}$	50.55%**	**58.32%****	44.73%**	3.77*	**3.00***	2.94*	55.09%**	**0.134****	0.458**
$SurveyLM_{\text{Qwen2.5-7B}}$	**54.65%****	57.23%**	**48.71%****	**3.98***	2.96*	**3.01***	56.83%**	0.128**	0.385**

Table 3: Multi-Task Evaluation Results. SurveyLM raises accuracy by about 10 to 15 percentage points and roughly halves Wasserstein distance. We conducted t-tests between each SurveyLM and its zero-shot model; $^{*}p < 0.05$ and $^{**}p < 0.01$. ↑ indicate that higher values are better, while ↓ indicate that lower values are better.

To mitigate bias, we ground predictions in expert-annotated reasoning chains and training with demographic fine-grained profiles. This enables interpretable, group-aware alignment without relying on stereotypes.

Task 4: Survey Response Modeling. This task predicts individual and group-level responses to questionnaire items.

Individual-level. Given a demographic profile $\boldsymbol{c}$ and question Q_i with labeled options (l_i, o_i), the model predicts the respondent's selected answer $(\hat{l}_i, \hat{o}_i) = f_{LLM}(\boldsymbol{c}, Q_i)$, where l_i denotes the label (e.g., A/B…) and o_i denotes the content. Predictions are evaluated using accuracy, macro-averaged precision, recall, and F1, based on the ground-truth answers.

Group-level. Aggregating individual responses, the model generates a distribution over possible answers: $\hat{\mathbf{p}}_q = \text{Aggregate}(\{\hat{o}_i\}), \quad \hat{\mathbf{p}}_q \in \Delta^{|L|-1}$. Alignment is measured via the Wasserstein distance (Similar to Task 3).

Table 2 summarizes the task definitions and dataset statistics. Further task-specific details and examples are provided in the Appendix D and E.

Constructing the SurveyLM Family. We adopt a two-stage alignment strategy to build the **SurveyLM** model family. First, we fine-tune three open-source LLMs (Mistral 7B, LLaMA 3.1 8B, Qwen 2.5 7B) on the Social Foundation Corpus to equip them with general social concepts and culturally diverse discourse patterns. Then, we fine-tune the adapted models on the four AlignSurvey tasks using the Entire-Pipeline Survey Datasets, enabling task-specific alignment with both qualitative and quantitative objectives. The resulting models serve as strong reference baselines for domain-aligned social survey modeling and will be publicly released to support reproducible research.

Experiments & Discussion

Setup and Implementation Details

We evaluate models under three settings: **zero-shot**, **few-shot** (with three in-context examples), and **supervised fine-tuning (SFT)**. SFT corresponds to the training procedure used to construct the **SurveyLM** family, which adapts LLMs through corpus-level and task-specific supervision. SFT first trains on the Social Foundation Corpus (1 epoch), followed by each task dataset (3 epochs). Prompts exceeding the model's context window are truncated by removing the middle segment, retaining the task instruction and query. Prompt templates are provided in Appendix D and E.

We apply LoRA (rank 8) targeting all attention and feed-forward layers. Training uses AdamW with learning rate 1×10^{-4}, cosine decay schedule, 10% linear warm-up, and bfloat16 precision. The effective batch size is 32 (per-device batch size of 4 with gradient accumulation of 8). All experiments fix the random seed to 42. Inference runs on 8× NVIDIA H20 GPUs or official APIs for proprietary models.

Evaluated Models

We benchmark three groups of models for comparative analysis. **Representative Top-tier Models:** GPT-4o, Claude 3.7 Sonnet, DeepSeek-R1, DeepSeek-R1-Distill-Qwen-14B, and Qwen2.5-72B, serve as references for general-purpose performance. **Open-source Base Models:** Meta-LLaMA-3.1-8B, Qwen2.5-7B, and Mistral-7B-v0.3 act as task-agnostic baselines, allowing analysis of architecture and scale effects on alignment. **SurveyLM Family:** $\text{SurveyLM}_{\text{Meta-LLaMA-3.1-8B}}$, $\text{SurveyLM}_{\text{Qwen 2.5-7B}}$, and $\text{SurveyLM}_{\text{Mistral-7B-v0.3}}$ specialized for the full survey pipeline and used to evaluate supervised domain alignment.

Model	ASE			CHIP			GSS		
	Acc.↑	F1↑	WD↓	Acc.↑	F1↑	WD↓	Acc.↑	F1↑	WD↓
GPT-4o *(Zero)*	46.56%	22.05%	2.1336	53.43%	26.24%	1.9325	32.79%	24.41%	1.1130
GPT-4o *(Few)*	47.91%	23.10%	1.7895	54.32%	27.30%	1.6992	33.85%	25.10%	1.0284
Claude 3.7 Sonnet *(Zero)*	42.11%	18.44%	1.7789	47.21%	9.50%	1.8872	33.94%	25.93%	1.2248
Claude 3.7 Sonnet *(Few)*	42.95%	19.50%	1.6542	47.98%	9.95%	1.7318	34.55%	26.60%	1.1095
DeepSeek-R1 *(Zero)*	47.21%	7.11%	1.9987	44.65%	4.52%	1.7789	34.54%	23.15%	1.3854
DeepSeek-R1 *(Few)*	47.88%	7.65%	1.8527	45.27%	4.90%	1.6523	35.20%	23.70%	1.2432
R1-Distill-Qwen-14B *(Zero)*	49.54%	25.90%	1.9295	50.62%	29.36%	1.9448	29.39%	11.57%	0.7920
R1-Distill-Qwen-14B *(Few)*	50.29%	26.80%	1.7164	51.37%	30.67%	1.7104	30.20%	12.25%	0.7548
Qwen2.5-72B *(Zero)*	42.20%	3.80%	1.8654	47.98%	13.29%	1.9918	34.50%	21.14%	1.2180
Qwen2.5-72B *(Few)*	43.06%	4.05%	1.7348	48.66%	14.02%	1.7989	35.60%	22.05%	1.1112
Mistral-7B-v0.3 *(Zero)*	28.04%	1.10%	2.2118	23.60%	3.46%	1.9941	3.19%	0.20%	1.1479
Mistral-7B-v0.3 *(Few)*	30.47%	1.25%	2.0823	26.89%	3.70%	1.8876	10.30%	0.24%	1.0913
Meta-Llama-3.1-8B *(Zero)*	14.84%	0.17%	2.1154	20.84%	2.40%	1.7748	1.61%	0.08%	1.5004
Meta-Llama-3.1-8B *(Few)*	17.26%	0.19%	1.9451	22.46%	2.60%	1.6435	6.90%	0.12%	1.3817
Qwen2.5-7B *(Zero)*	16.39%	0.12%	1.8479	51.95%	12.51%	1.9099	25.33%	4.30%	1.2411
Qwen2.5-7B *(Few)*	18.46%	0.15%	1.7059	53.14%	13.15%	1.7441	26.85%	4.55%	1.1371
$SurveyLM_{\text{Mistral-7B-v0.3}}$	54.59%**	41.34%**	0.0671**	65.56%**	**35.35%****	0.0857**	51.06%**	35.73%**	0.2097**
$SurveyLM_{\text{Meta-Llama-3.1-8B}}$	**56.11%****	41.47%**	0.0675**	64.89%**	32.99%**	0.0845**	**51.44%****	33.84%**	0.2137**
$SurveyLM_{\text{Qwen2.5-7B}}$	55.47%**	**41.79%****	**0.0541****	**67.17%****	34.94%**	**0.0686****	51.52%**	**36.22%****	**0.1849****

Table 4: Survey Response Modeling (Task4) Results. SurveyLM achieves high accuracy and F1 scores while yielding substantially smaller Wasserstein distances. Two-sided t-tests vs. zero-shot baselines: $^{*}p < 0.05$, $^{**}p < 0.01$. ↑ indicate that higher values are better, while ↓ indicate that lower values are better.

Experimental Results & Analysis

Table 3 and Table 4 report aggregated results for Tasks 1–3 and Task 4. For Task 1, we divided accuracy into three categories which specific classification can be found in Appendix B.1. Figure 2 complements these with a radar chart overview of model performance across tasks. Detailed per-label scores and metrics are provided in Appendix B and C.

Task 1: Social Role Modeling. SurveyLM models significantly outperform all baselines in predicting fine-grained demographic attributes. While top-tier models like GPT-4o perform decently on binary traits (e.g., gender), they struggle with complex categories such as family structure and social roles. Few-shot prompting yields limited gains. SurveyLM models improve accuracy by over 10 points across attributes, with the strongest gains in familial and social context fields. Notably, a fine-tuned 7B model (Qwen) surpasses its 72B counterpart, highlighting the impact of domain alignment over model scale.

Task 2: Semi-structure Interview Modeling. General-purpose models (e.g., GPT-4o, DeepSeek-R1) reach moderate naturalness but lack stylistic consistency and fail to match the interview style. SurveyLM models, with two-stage alignment, yield more fluent, coherent, and contextually appropriate responses, demonstrating the importance of fine-tuning for qualitative generation.

Task3. Attitude Stance Modeling. General-purpose models tend to default to majority stances with generic reasoning, showing 28–45% accuracy and limited few-shot improvement. They perform poorly on group-level metrics Wasserstein Distance. SurveyLM models exceed 55% accuracy and generate higher-quality explanations (Jaccard higher), confirming the value of alignment for accurate, interpretable attitude inference in public opinion modeling.

Task4. Survey Response Modeling. In both zero- and few-shot settings, general-purpose models yield low-fidelity outputs (accuracy <50%, low F1). SurveyLM achieves high accuracy and F1 scores while yielding substantially smaller Wasserstein distances on ASE, CHIP, and GSS. The pronounced jump in F1 shows that the model no longer "plays it safe" by predicting the midpoint response for everyone. It learns to recognize and output the full spectrum of options. This balanced improvement mitigates the **central-tendency bias** that questionnaires often suffer from. These results demonstrate the strength of supervised alignment for structured response prediction.

Ablation Analysis. To assess the impact of each alignment stage, we conduct ablation tests on three backbones (Mistral-7B, LLaMA-3.1-8B, Qwen2.5-7B), removing either Stage I (foundation adaptation, *w/o Foundation*) or Stage II (task-specific tuning, *w/o Task SFT*), keeping other conditions fixed.

As shown in Table 5, removing either stage degrades performance. Omitting foundation adaptation yields moderate drops, while skipping task-specific tuning causes severe degradation. Full SurveyLM models outperform ablated variants, confirming that foundation-level pretraining and task-specific supervision are complementary. This highlights the effectiveness of our two-stage alignment strategy in adapting general-purpose LLMs to social survey tasks.

Equity-Oriented Gains. Figure 3 shows Task 3 accuracy across demographic attributes using the Qwen2.5-7B. Sur-

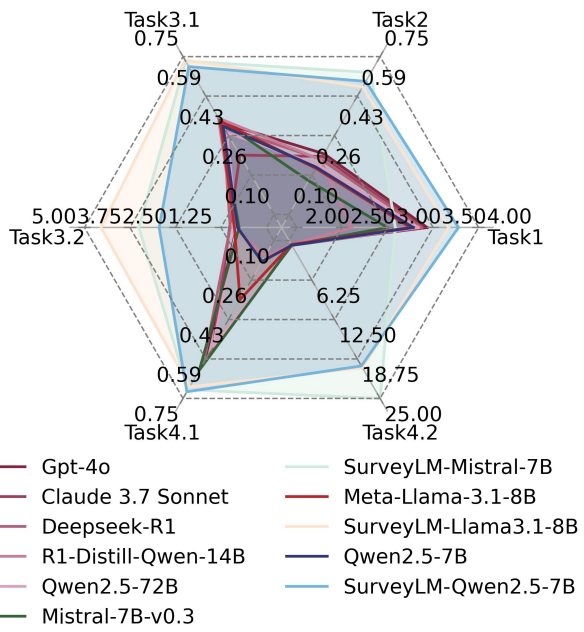

Figure 2: Radar chart of task performance. Axes are task scores (higher is better; WD inverted).

veyLM consistently shifts scores toward higher accuracy, while baselines often remain near or below midpoint. Notably, SurveyLM significantly boosts accuracy for **underrepresented groups**, including rural, elderly (76+), self-employed, and low-/middle-income individuals, reducing disparities with advantaged counterparts (urban, college-educated, high-income) and enhancing demographic equity. Full comparisons across models are in the Appendix C.2.

Social Impact and Alignment Implications

Our experiments show that socially grounded alignment enhances both performance and fairness. AlignSurvey and SurveyLM help recover signals from marginalized groups, reducing reliance on dominant narratives and supporting more

Model	T1	T2	T3		T4(ASE)	
	Acc	Avg	Acc	WD	Acc	WD
$SurveyLM_{\text{Mistral}}$	**0.47**	2.80	**0.57**	**0.297**	**0.55**	**0.067**
w/o Foundation	0.46	**3.10**	0.56	0.303	0.53	0.077
w/o Task SFT	0.38	2.96	0.44	0.313	0.36	2.182
$SurveyLM_{\text{Llama}}$	**0.45**	**3.24**	**0.55**	**0.458**	**0.56**	**0.068**
w/o Foundation	0.44	3.07	0.54	0.463	0.55	0.085
w/o Task SFT	0.39	3.01	0.36	0.484	0.34	2.098
$SurveyLM_{\text{Qwen}}$	**0.49**	**3.32**	**0.57**	**0.385**	**0.56**	**0.054**
w/o Foundation	0.47	3.12	0.56	0.389	0.55	0.889
w/o Task SFT	0.35	2.79	0.47	0.397	0.36	2.002

Table 5: Ablation results on Tasks 1–4. T1–T4 correspond to Tasks 1–4, respectively. For Tasks 1 and 2 we report accuracy and the average score; for Task 4 we report results on the ASE split only. Detailed results appear in Appendix C.1.

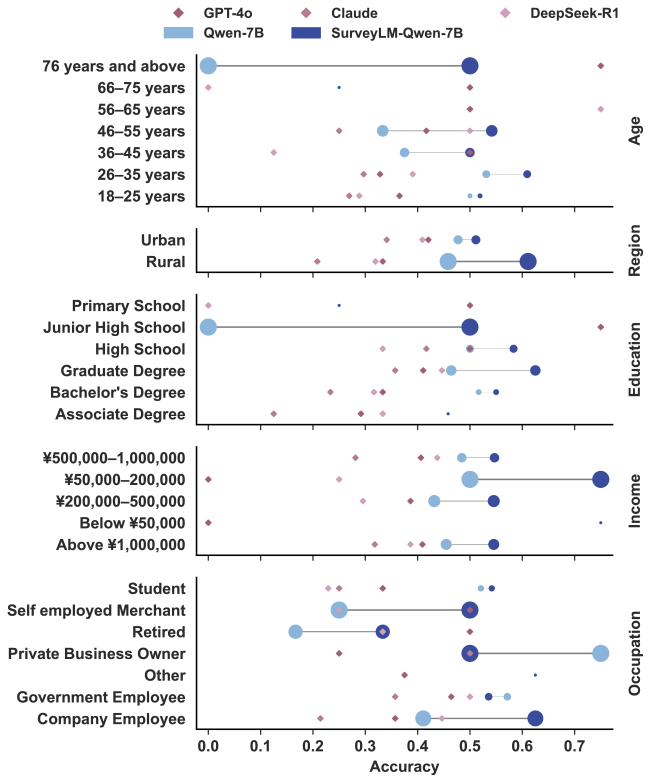

Figure 3: Multi-demographic Task 3 accuracy for Qwen2.5-7B. Circle size encodes the accuracy gain of SurveyLM over the base model. SurveyLM consistently raises accuracy, with especially large gains for underrepresented groups (rural residents, age 76+, self-employed, low- and middle-income), narrowing gaps with advantaged groups (urban, college-educated, high-income) and improving demographic equity.

inclusive policy diagnostics, welfare targeting, and equitable decision-making. More broadly, this work demonstrates how targeted alignment can bridge LLMs with real-world societal needs,especially in governance, auditing, and digital public services.

Conclusion & Future Work

This paper presents AlignSurvey, the first benchmark to systematically replicate the full pipeline of social surveys using LLMs. By integrating Social Foundation Corpus and Entire-Pipeline Survey Datasets, it enables comprehensive evaluation across demographic modeling, qualitative interaction, attitude inference, and structured response prediction. Experiments demonstrate alignment on AlignSurvey can recover signals from underrepresented groups, reducing bias and supporting more inclusive, policy-relevant modeling.

Future directions include iterative improvement through human-in-the-loop feedback and expanded coverage across diverse cultural settings for broader applicability.

Ethical Statement

We take ethics seriously. Our dataset is sourced from publicly accessible content, including video-platform APIs, research-cleared books, and licensed surveys. Personally identifiable information is removed or anonymized before training, and only privacy-preserving, non-harmful content is released. Raw audio, video, or verbatim transcripts are never shared.

We implement multiple safeguards to prevent misuse, including controlled data access, content filtering, and usage guidance tailored to responsible LLM-based survey applications. Data is used solely for non-commercial academic research, and compliance with privacy and copyright regulations is continuously monitored. See Appendix G for further details.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 72574198, 72434004), the Key Project of Humanities and Social Sciences of the Ministry of Education of China (Grant No. 2023JZDZ038), and the National Social Science Fund of China (Grant No. 23BZZ088). We thank all collaborators contributing to data collection, annotations and valuable feedback.

References

Abdurahman, S.; Atari, M.; Karimi-Malekabadi, F.; Xue, M. J.; Trager, J.; Park, P. S.; Golazizian, P.; Omrani, A.; and Dehghani, M. 2024a. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7): pgae245.

Abdurahman, S.; Atari, M.; Karimi-Malekabadi, F.; Xue, M. J.; Trager, J.; Park, P. S.; Golazizian, P.; Omrani, A.; and Dehghani, M. 2024b. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3.

Abeliuk, A.; Gaete, V.; and Bro, N. 2025. Fairness in LLM-Generated Surveys. *arXiv preprint arXiv:2501.15351*.

Ahmed, A.; Pereira, L.; and Jane, K. 2024. Mixed methods research: Combining both qualitative and quantitative approaches. *en. In: ResearchGate (Sept. 2024)*, 1–10.

Aroyo, L.; Taylor, A. S.; Díaz, M.; Homan, C. M.; Parrish, A.; Serapio-García, G.; Prabhakaran, V.; and Wang, D. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. *ArXiv*, abs/2306.11247.

Binz, M.; Akata, E.; Bethge, M.; Brändle, F.; Callaway, F.; Coda-Forno, J.; Dayan, P.; Demircan, C.; Eckstein, M. K.; Éltető, N.; et al. 2024. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Binz, M.; and Schulz, E. 2023. Turning large language models into cognitive models. *ArXiv*, abs/2306.03917.

Chu, E.; Andreas, J.; Ansolabehere, S.; and Roy, D. 2023. Language Models Trained on Media Diets Can Predict Public Opinion. *ArXiv*, abs/2303.16779.

Cui, Z.; Li, N.; and Zhou, H. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 1–8.

De Duro, E. S.; Taietta, E.; Improta, R.; and Stella, M. 2024. Phdgpt: Introducing a psychometric and linguistic dataset about how large language models perceive graduate students and professors in psychology. *arXiv preprint arXiv:2411.10473*.

Dominguez-Olmedo, R.; Hardt, M.; and Mendler-Dünner, C. 2024. Questioning the Survey Responses of Large Language Models. *Advances in Neural Information Processing Systems*.

ESOMAR. 2024. Global Market Research. https://esomar.org/publications/report-1. Accessed: 2025-11-03.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. *ArXiv*, abs/2402.01306.

Evans, J. R.; and Mathur, A. 2018. The value of online surveys: A look back and a look ahead. *Internet research*, 28(4): 854–887.

Fetters, M. D.; Curry, L. A.; and Creswell, J. W. 2013. Achieving integration in mixed methods designs—principles and practices. *Health services research*, 48(6pt2): 2134–2156.

Giorgi, T.; Cima, L.; Fagni, T.; Avvenuti, M.; and Cresci, S. 2025. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 653–670.

He, Z.; Dorn, R.; Guo, S.; Chu, M. D. H.; and Lerman, K. 2024. Community-Cross-Instruct: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities. In *Conference on Empirical Methods in Natural Language Processing*.

Heffetz, O.; and Reeves, D. B. 2019. Difficulty of reaching respondents and nonresponse Bias: Evidence from large government surveys. *Review of Economics and Statistics*, 101(1): 176–191.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Hu, T.; Kyrychenko, Y.; Rathje, S.; Collier, N.; van der Linden, S.; and Roozenbeek, J. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1): 65–75.

Hwang, E.; Majumder, B. P.; and Tandon, N. 2023. Aligning Language Models to User Opinions. *ArXiv*, abs/2305.14929.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36: 24678–24704.

Kalton, G. 2009. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2): 125–141.

Kim, J.; and Yang, Y. 2024. Few-shot Personalization of LLMs with Mis-aligned Responses. *ArXiv*, abs/2406.18678.

Kirk, H. R.; Whitefield, A.; Rottger, P.; Bean, A. M.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams,

A.; He, H.; Vidgen, B.; and Hale, S. A. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *Neural Information Processing Systems*.

Kopf, A.; Kilcher, Y.; von Rutte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Duc, N. M.; Stanley, O.; Nagyfi, R.; Shahul, E.; Suri, S.; Glushkov, D.; Dantuluri, A. V.; Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant Conversations - Democratizing Large Language Model Alignment. *ArXiv*, abs/2304.07327.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles*.

Lambert, N.; Pyatkin, V.; Morrison, J. D.; Miranda, L. J. V.; Lin, B. Y.; Chandu, K. R.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; Smith, N. A.; and Hajishirzi, H. 2024. RewardBench: Evaluating Reward Models for Language Modeling. *ArXiv*, abs/2403.13787.

Lee, S.; Peng, T.-Q.; Goldberg, M. H.; Rosenthal, S. A.; Kotcher, J. E.; Maibach, E. W.; and Leiserowitz, A. 2024. Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8): e0000429.

Li, L.; Li, J.; Chen, C.; Gui, F.; Yang, H.; Yu, C.; Wang, Z.; Cai, J.; Zhou, J. A.; Shen, B.; et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Liu, H.; Cao, Y.; Wu, X.; Qiu, C.; Gu, J.; Liu, M.; and Hershcovich, D. 2025. Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation. *Information Processing & Management*, 62(4): 104099.

Mellon, J.; Bailey, J.; Scott, R.; Breckwoldt, J.; Miori, M.; and Schmedeman, P. 2024. Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, 11(1): 20531680241231468.

Moon, S.; Abdulhai, M.; Kang, M.; Suh, J.; Soedarmadji, W.; Behar, E. K.; and Chan, D. M. 2024. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.

Moy, P.; and Murphy, J. 2016. Problems and prospects in survey research. *Journalism & Mass Communication Quarterly*, 93(1): 16–37.

Prosser, C.; and Mellon, J. 2018. The twilight of the polls? A review of trends in polling accuracy and the causes of polling misses. *Government and Opposition*, 53(4): 757–790.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023a. Whose Opinions Do Language Models Reflect? *ArXiv*, abs/2303.17548.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023b. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Simmons, G. 2022. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. *ArXiv*, abs/2209.12106.

Suh, J.; Jahanparast, E.; Moon, S.; Kang, M.; and Chang, S. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Sun, C.; Yang, K.; Reddy, R. G.; Fung, Y. R.; Chan, H. P.; Zhai, C.; and Ji, H. 2024. Persona-DB: Efficient Large Language Model Personalization for Response Prediction with Collaborative Data Refinement. *ArXiv*, abs/2402.11060.

Thapa, S.; Shiwakoti, S.; Shah, S. B.; Adhikari, S.; Veeramani, H.; Nasim, M.; and Naseem, U. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1): 1–30.

Tourangeau, R. 2004. Survey research and societal change. *Annu. Rev. Psychol.*, 55(1): 775–801.

Tsai, T.-I.; Luck, L.; Jefferies, D.; and Wilkes, L. 2025. Challenges in adapting a survey: ensuring cross-cultural equivalence. *Nurse researcher*, 33(2).

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; and Miao, Z. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–21.

Wright, J. D.; Marsden, P. V.; et al. 2010. Survey research and social science: History, current practice, and future prospects. *Handbook of survey research*, 3–26.

Zhang, X.; Lin, J.; Mou, X.; Yang, S.; Liu, X.; Sun, L.; Lyu, H.; Yang, Y.; Qi, W.; Chen, Y.; et al. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

Zhao, S.; Dang, J.; and Grover, A. 2023. Group Preference Optimization: Few-Shot Alignment of Large Language Models. *ArXiv*, abs/2310.11523.

Zhou, Q.; Zhang, J.; Wang, D.; Liu, Q.; Li, T.; Dong, J. S.; Wang, W.; and Guo, Q. 2025. Fair-PP: A Synthetic Dataset for Aligning LLM with Personalized Preferences of Social Equity. *arXiv preprint arXiv:2505.11861*.

Zhou, Z.; Li, Y.; and Yu, J. 2024. Exploring the application of LLM-based AI in UX design: an empirical case study of ChatGPT. *Human–Computer Interaction*, 1–33.

Zhu, L.; Huang, X.; and Sang, J. 2025. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, 4653–4661.