ALPHA: Action-Based Learning for Pluralistic Human Alignment in Large Language Models

Aanisha Bhattacharyya[⋆1,2,3], Susmit Agrawal[⋆1#], Yaman Kumar Singla[⋆1], Tarun Ram Menta[1], Nikitha Sr[1], Rajiv Ratn Shah[2], Changyou Chen[3], Balaji Krishnamurthy[1]

[1]Adobe Media and Data Science Research (MDSR)
[2]Indraprastha Institute of Information Technology Delhi (IIIT Delhi)
[3]State University of New York at Buffalo
behavior-in-the-wild@googlegroups.com

Abstract

Large language models are widely used, yet aligning them with societal values remains challenging. Current approaches often rely on human annotations, which are hard to scale, or synthetic data produced by models that may themselves be misaligned, making it difficult to capture genuine public opinion. This limits scalability and introduces demographic biases that reduce the representativeness and fairness of model behavior. We introduce a novel approach to pluralistic alignment through behavioral learning, grounded in the psychological principle that observed actions exhibit strong consistency with underlying opinions. Specifically, we present ALPHA50M, a dataset of over 50 million samples derived from 1.5 million real-world advertisements and incorporating rich behavioral signals inferred from demographic engagement patterns. Models trained on this data achieve state-of-the-art zero-shot performance on diverse alignment benchmarks spanning cultural reasoning, political views, and social values. We also propose two new benchmarks. OpinionQA-XL aggregates large-scale survey questions covering over 100 societal topics, while GSS evaluates models' ability to capture temporal shifts in societal opinions across decades. Our results demonstrate that learning from behavioral signals enables models to align with diverse societal values across demographic groups, capture underlying social and cultural norms, and generalize to unseen surveys, topics, and time periods beyond the training distribution. This behavioral learning paradigm offers a scalable and demographically broad alternative to existing alignment techniques.

1 Introduction

"Only in *actions* can you fully recognize the forces operative in social behavior"

–Milgram (1974)

LLM-powered assistants have gained popularity, with some reaching over 800 million users per week. If not properly aligned, these models can amplify harmful stereotypes, representational bias, or inaccuracies, especially in sensitive domains like healthcare and civic services, leading to misinformation or unequal treatment that disproportionately affects marginalized groups (Weidinger et al. 2022; Zhang et al. 2023). For instance, GPT-3.5 has been shown to recommend less guideline-consistent care for women and African-American patients with chest pain (Zhang et al. 2023). Another study found that leading chatbots advised women and minorities to request significantly lower salaries than equally qualified white men, with disparities reaching up to $120,000 in identical job scenarios (Yamshchikov et al. 2025). Aligning LLMs with the full diversity of human values remains a major challenge. Popular alignment methods such as Instruction Finetuning (IFT) (Ouyang et al. 2022), Reinforcement Learning with Human Feedback (RLHF) (Kaufmann et al. 2024), and Direct Preference Optimization (DPO) (Rafailov et al. 2023; Zhao, Dang, and Grover 2023) rely on curated feedback datasets (Touvron et al. 2023), which are expensive to scale and often reflect narrow demographic perspectives. Studies show that models trained with human feedback often internalize the values of annotators who are predominantly young, liberal, well-educated, and non-religious, resulting in skewed outputs (Santurkar et al. 2023; Durmus et al. 2023; Ryan, Held, and Yang 2024; AlKhamissi et al. 2024). For example, models frequently produce more left-aligned responses to sociopolitical questions than the broader public would (Santurkar et al. 2023). This misalignment arises because annotators and synthetic supervision data are not demographically representative. As a result, scaling alignment across diverse global populations remains a key open challenge.

This lack of demographic breadth in training data creates a gap between LLM behavior and the values of real-world users. As these models increasingly influence decision-making and public discourse, it becomes critical to ground them in a broad understanding of public values. However, building and continuously updating alignment datasets that represent a wide range of societal perspectives remains a significant challenge, especially when relying on small-scale sources like public opinion surveys, which have seen steadily declining response rates in recent years (Berinsky 2017; Kochhar 2023; Silver et al. 2024). Achieving pluralistic alignment requires large, continuously refreshed datasets to capture the diversity and evolution of human values. This makes it essential to develop scalable and adaptive alignment methods that can generalize beyond sparse, static, and demographically skewed datasets.

Project page: https://behavior-in-the-wild.github.io/align-via-actions.html

⋆Equal Contribution.

Work done while at Adobe.

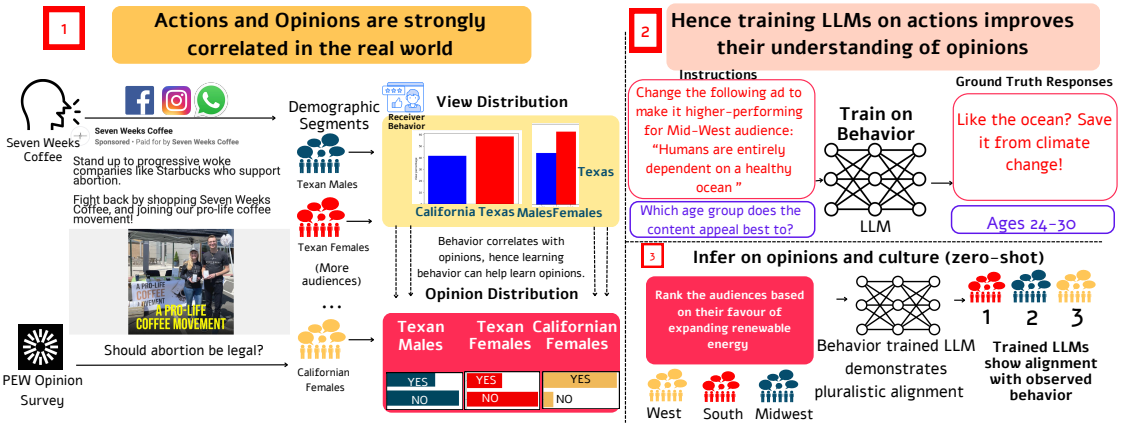

Figure 1: Behavior and Opinions are strongly correlated. The behavioral data, which contains the ad content, the audience, and the behavior that the audience showed towards the ad, helps in understanding the audience. While behavior is already being collected at scale, it is conventionally not used to train large language models. We use these sparse in-the-wild behavioral signals to train our model on transcreation, transsuasion tasks and find that this helps in aligning LLMs with opinions.

While building alignment datasets at scale remains difficult, acquiring human digital action data such as likes, comments, and shares is comparatively straightforward. These behavioral signals are routinely captured for applications like market research and advertising. Psychological research has shown that behavior both reflects and shapes attitudes (Fazio and Zanna 1981), with strong correlations observed across domains from voting (Kelley and Mirer 1974) to military performance (Stouffer et al. 1949). For instance, Kelley and Mirer (1974) showed that cues like party affiliation or past voting could predict political attitudes and choices, even when not self-reported. Someone frequently engaging with environmentally focused content likely holds pro-environmental views. This consistency implies actions can serve as reliable proxies for underlying beliefs. Yet paradoxically, during LLM pretraining, readily available engagement data (e.g., upvotes, likes) is often discarded as noise (Biderman, Bicheno, and Gao 2022; Penedo et al. 2023; Khandelwal et al. 2024), and alignment is later attempted using curated annotations (Shi et al. 2024; Huang and Yang 2023; Bai et al. 2022) or costly surveys (Hwang, Majumder, and Tandon 2023; Zhao, Dang, and Grover 2023; Li et al. 2024). Leveraging behavioral data offers a scalable alternative, enabling models to learn human attitudes, cultural norms, and values directly from action data providing implicit, demographically diverse supervision.

Building on the established correlation between attitudes and behaviors in psychological literature, we investigate the question: "*Can unsupervised training of LLMs on actions sampled from digital analytics achieve pluralistic human alignment comparable to that attained through supervised training on expert-annotated datasets?*" To explore this hypothesis, we utilize data from the Meta Ads Library (Meta Platforms 2025), containing ads displayed on Meta platforms (Facebook, Instagram, WhatsApp, and Messenger). This dataset provides comprehensive information, including ad content, publisher details, campaign duration, advertiser expenditure, and viewer demographics segregated by region, age, and gender (Fig. 2).

We begin by demonstrating a strong correlation between public opinions captured in Pew Research surveys and user behavior reflected in Meta Ads viewership data (§3.2). Building on this insight, we design instruction finetuning tasks to train LLMs to predict user behavior based on ad content (§3.3), enabling the models to learn user preferences from behavioral signals. As part of this effort, we introduce **ALPHA50M**, a large-scale instruction dataset comprising 50 million examples. LLMs fine-tuned on ALPHA50M significantly outperform both their base versions and chat models (i.e., base models fine-tuned on expert-curated IFT datasets) on benchmarks measuring alignment with societal opinions and cultural norms (Table 1). Notably, in zero-shot settings, our models also surpass those trained on carefully curated public opinion datasets across a wide range of social and cultural issues. These results highlight the potential of behavioral data as a scalable and effective signal for achieving pluralistic alignment in LLMs.

Our work makes the following contributions:

• **We propose a novel approach for pluralistic alignment** of LLMs by fine-tuning them on *human behavioral signals*, such as likes, shares, and comments, captured through web analytics. These actions offer scalable and demographically diverse proxies for social attitudes, reducing dependence on costly and narrow annotation pipelines. Grounded in psychological findings that behavior reflects underlying values, this method enables alignment at population scale. Models trained on such behavior-derived data achieve strong *zero-shot alignment* across a range of social and cultural benchmarks, including OpinionQA, GlobalOpinionQA, CultureBench, and CultureNLI, as summarized in Table 1.

•We introduce the **ALPHA50M dataset**, a comprehensive instruction training set derived from 1.5 million advertisements by over 120,000 advertisers in the Meta Ads Archive. ALPHA50M comprises 50 million instruction training samples designed to teach LLMs about human behavior, significantly surpassing existing datasets in scale Each instruction incorporates advertisement caption, advertiser information,

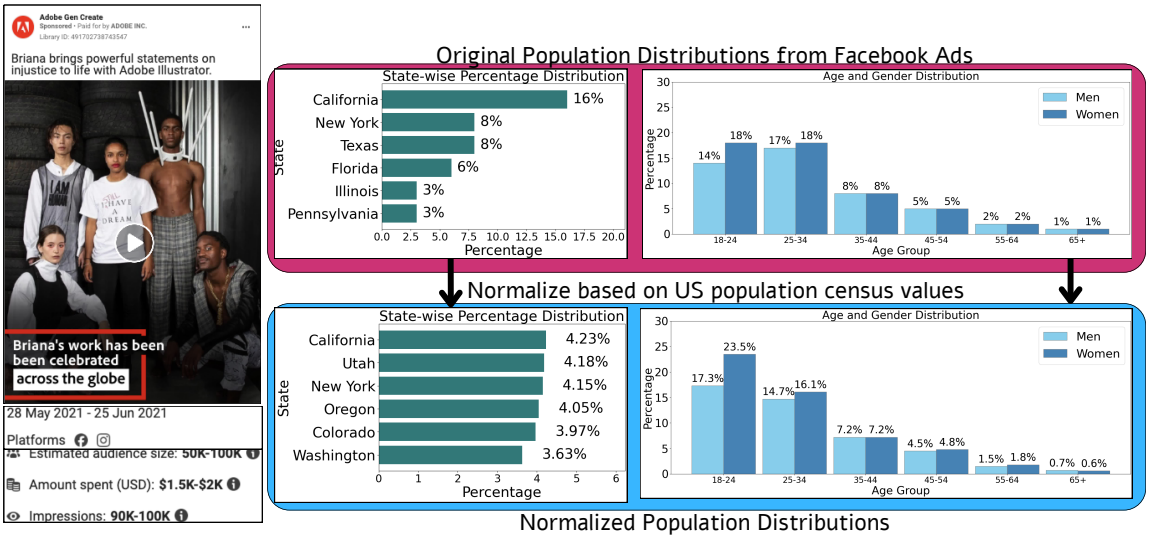

Figure 2: A sample advertisement from the Meta Ad Library.

publication date, media verbalization, and target audience. We release ALPHA50M to facilitate LLM alignment and further research.

• We introduce **two new benchmarks** to evaluate the social alignment of LLMs across time, topics, and demographic diversity First, **OpinionQA-XL** is a major expansion of OpinionQA (Santurkar et al. 2023), increasing coverage from 1,498 to over 14,000 questions drawn from 119 Pew Research Center surveys spanning 104 distinct topics, including politics, biomedicine, technology, social media, and global affairs. This expansion provides a far more comprehensive and fine-grained testbed for probing opinion alignment. Second, **GSS** leverages the General Social Survey (General Social Survey 2025), containing 5,026 repeated questions asked across different years and respondent demographics. This dataset uniquely allows us to evaluate whether LLMs can accurately model and distinguish temporally evolving societal views. OpinionQA-XL also includes survey items postdating common LLM training cutoffs (e.g., late 2023), enabling evaluation beyond memorized content.

2 Related Works

Opinion and Culture Alignment: Aligning LLMs with subjective human opinions and cultural biases presents significant challenges. Recent studies have investigated the implicit alignment of LLMs to human perspectives and cultural norms (Hartmann, Schwenzow, and Witte 2023; Simmons 2023; Cao et al. 2023; Johnson et al. 2022; Masoud et al. 2024; Naous et al. 2024; Wang et al. 2024). Proposed alignment methods include targeted prompting to emulate specific demographic groups (Jiang et al. 2022; Argyle et al. 2023) and fine-tuning approaches such as RLHF (Ouyang et al. 2022; Daniels-Koch and Freedman 2022) or instruction-based fine-tuning on opinion or cultural data (Huang and Yang 2023; Zhao, Dang, and Grover 2023; Li et al. 2024; Shi et al. 2024). However, these techniques rely on explicit human annotations, which are resource-intensive and prone to errors. In contrast, our work demonstrates that LLMs can effectively align with human opinions and cultural norms using in-the-wild behavioral signals, eliminating the need for explicit annotations.

Measuring Alignment: Recent work has evaluated the alignment of large language models (LLMs) with human opinions and cultural values. Durmus et al. (2023), for example, assess whether LLM outputs entail or agree with human responses as a proxy for belief alignment. Others, including Santurkar et al. (2023) and Shi et al. (2024), use large-scale public surveys to test how closely LLMs reflect societal views. The CultureBank dataset (Shi et al. 2024), containing 23,000 cultural descriptors from TikTok and Reddit, serves as both a training source and benchmark. These studies also introduce metrics for evaluating alignment with demographic subgroups. A consistent finding is that LLMs often mirror views of certain groups, typically US-based, left-leaning, educated, affluent, or non-religious populations. Off-the-shelf instruction-tuned models also produce generic responses lacking cultural nuance. We adopt and extend these metrics to assess pluralistic alignment and show that training on in-the-wild behavioral signals improves zero-shot alignment across models and tasks, leading to more culturally aware and demographically grounded responses.

3 Experimental Setup

In this section, we describe our process for collecting and cleaning behavioral data, analyze its correlation with social opinion trends, and outline the tasks designed to train LLMs on these signals. These tasks help LLMs learn patterns in user behavior, content, and audience preferences embedded in the data. Our core hypothesis is that such behavioral signals are reflective of underlying social and cultural opinions, and that training on them enables models to internalize the structure of human preferences. As behavior and opinion are often closely linked, these tasks are aimed at improving the pluralistic alignment of LLMs with diverse human perspectives.

3.1 Collecting Behavior Data

We collect in-the-wild behavioral data from Meta's Ad Library, including ad content, creation date, spend, impressions, language, and demographic delivery metrics. Fig.2 shows an example ad by "Adobe Gen Create." Ad delivery varies across states and demographic groups due to factors like population size, audience interests, and advertiser targeting. However, larger groups naturally receive more impressions, which can obscure true patterns of interest. To address this, we normalize ad delivery using demographic distributions from the 2020 US Census (Bureau 2024), removing the confounding effect of group size and revealing which groups are genuinely more interested. Figure2 illustrates this process. On the left is a sample ad for Adobe Illustrator, showing estimated audience size, spend, and impressions. The top charts display the original delivery: state-wise (top-left) shows California received 16% of impressions, followed by New York and Texas; age-gender (top-right) shows highest delivery among users aged 18–34 across both genders. These reflect both size and interest. After census-based normalization (bottom row), California still ranks high, but smaller states like Utah and Oregon emerge, indicating high per-capita engagement. The normalized age-gender chart shows that although women aged 18–24 and 25–34 had similar delivery, the ad was more popular per capita among the 18–24 group. This allows us to rank segments by inferred interest rather than raw exposure. We also compute impressions per dollar to estimate ROI and analyze effectiveness.

We collect 1,474,367 ads after removing duplicates and incomplete entries. URLs are standardized as [URL]'. Ads are grouped by page, treating brand subsidiaries separately (e.g., Amazon Prime' vs. 'Amazon Alexa'). The final dataset includes 122,636 unique advertisers. All ads are converted to a text-to-text format to evaluate LLMs in a unified textual setting. Media content is verbalized following Bhattacharyya et al. (2023) and merged with ad text. Multi-frame ads (405,485) are combined into a single body.

3.2 Correlation Analysis

Psychological research has long shown that behavior reflects underlying attitudes across domains (Fazio and Zanna 1981). While most studies examine this in controlled settings, we explore whether **behavioral similarity** between demographics can indicate **opinion similarity** at scale. Rather than correlating individual actions with beliefs, we assess whether demographic groups that behave similarly via ad engagement also express similar views via survey responses. If these similarity measures correlate, behavioral data from digital actions could serve as a scalable proxy for opinion estimation.

To investigate this relationship, we measure similarity between pairs of exemplar demographics, A and B, across two dimensions: (i) behavioral similarity, such as how similarly they engage with ads, and (ii) opinion similarity, such as how closely their survey responses align. A strong correlation between these similarity scores across all demographic pairs would suggest that behavioral patterns serve as reliable indicators of broader opinion alignment. Indeed, we find a high overall correlation of $\mathbf{r = 0.87}$ between behavioral and opinion similarity scores across regional pairs, indicating substantial substitutability. Applying the same approach to age-gender groups, we observe a similarly positive correlation of $\mathbf{r = 0.65}$ between their behaviors and opinions. Full methodological details are provided in Appendix.

3.3 Constructing ALPHA50M Dataset

Having established that behavior is statistically correlated with opinions, we now formulate the training methodology to teach LLMs human behavior. We propose four tasks to train LLMs on behavioral data from Meta ads:

(1) **Targeted Advertisement Generation (TAG)** This task trains LLMs to create ads for specific target audiences based on Meta ad engagement data. Meta ads have signals that depict, out of all the demographic segments (defined by age group, gender, or region), which segment engaged with the ad more compared to others. Given the segments that engaged with the ad most, the LLM is instructed to generate the ad for the given segment. Input parameters include target audience, advertising budget, ad dates, marketer's name, and ad body keywords (extracted using KeyBERT). The output is the generated ad content.

(2) Target Audience Prediction (TAP) This task trains LLMs to predict the most engaged audience given the ad, marketer name and time of the ad as input. The LLM is then required to: a) Simulate gender-based audience responses, predicting whether the ad would appeal more to male or female audiences. b) Simulate age group-specific reactions, identifying the optimal target age group for the ad. c) Simulate regional audience behaviors across U.S. states, predicting the state where the ad would have the highest appeal. TAP complements TAG by teaching LLMs to infer underlying correlations between content phrasing and audience preferences, enhancing their understanding of demographic-specific content affinities.

(3) **Transcreation (TC):** Transcreation adapts a message from a source audience to a target audience while preserving its core meaning (Khanuja et al. 2024). We define transcreation tasks by age (TC-A), gender (TC-G), and region (TC-R), converting ads that resonate with one demographic to appeal to another. We identified 33.4 million ad pairs from the same marketer with equivalent meanings, each reflecting a change in one demographic variable. For region-based TC, we retained pairs with a rank difference of at least three positions. For age- and gender-based TC, we kept pairs where the target demographic was not top-ranked in the source ad but ranked first in the target.

(4) **Transsuasion Tasks**: Transsuasion (Singh et al. 2025) aims to make content more persuasive for a given audience, sender, time, and channel. In the original setup, the task involves pairs of similar ads, where one performs poorly and the other performs well, and trains LLMs to generate more persuasive versions of the lower-performing ad by learning from the higher-performing one. We extend this framework to two specific goals: enhancing audience engagement (TS-E) and improving advertiser budget utilization (TS-B), while keeping other variables constant. We generate 50 million ad pairs ($Ad1$, $Ad2$) from the same brands targeting the same demographics, where $Ad2$ is higher impact than $Ad1$. In TS-B,

$Ad2$ has a larger budget; in TS-E, it shows higher engagement for the demographic. The LLM is trained to transform $Ad1$ into $Ad2$. Pairs are filtered to ensure both semantic and surface-level similarity: (1) SentenceBERT similarity greater than 0.7 (Reimers and Gurevych 2019), and (2) Levenshtein distance greater than 15 characters. §

Test Set Creation. For each task, we hold out portions of data as test sets to validate LLM training. These test sets are designed not only to monitor model performance but also to evaluate the model's ability to learn task-specific behavioral patterns in a robust and generalizable way. We create two types of test sets: (1) ads from previously unseen advertisers, which assess the model's capacity to generalize beyond known brand language and marketing styles, and (2) ads published after June 2023, which evaluate the model's adaptability to new social and cultural contexts. These test sets help us measure how effectively the model learns from behavioral tasks and whether it can apply those learnings to unfamiliar or evolving scenarios. More broadly, they serve as a way to evaluate the model's ability to perform task-driven generalization while moving toward the broader goal of pluralistic alignment. This includes aligning with a diverse range of social opinions, cultural expressions, and audience expectations across time and contexts.

3.4 Benchmarking Cultural and Opinion Alignment in Behavior-Trained Models

To investigate cultural and opinion alignment, we conduct zero-shot evaluations using popular culture and opinion alignment benchmarks:

(1) **OpinionQA** (Santurkar et al. 2023), derived from Pew Opinion Surveys, evaluates LLM alignment with the US public and specific demographic groups. It tests whether models show preferential alignment with viewpoints (e.g., conservative vs. liberal) across diverse topics. The dataset includes 1,498 questions from 15 PEW surveys, covering 60 demographic groups with data up to July 2021. For instance, in response to "Do you think climate change is a major threat to the well-being of the United States?", Pew data shows strong partisan divergence, with liberal majorities answering "Yes" and conservative groups less likely to agree. Santurkar et al. (2023) showed that stronger zero-shot performance implies better opinion alignment, making the benchmark well-suited for behaviorally trained LLMs. They also introduce metrics comparing model opinion distributions to the public (*Representativeness*,

(2) **OpinionQA-XL**: We substantially expanded the original OpinionQA dataset to include PEW surveys up to August 2023, resulting in *OpinionQA-XL*. This new version contains 14,554 questions drawn from 119 surveys, a 9.7x increase over the original and introduces 68 new topics, including *Climate Change*, *Space Tourism*, and the *Digital Economy*, thereby greatly broadening its topical coverage.

(3) **GlobalOpinionQA** (Durmus et al. 2023) contains 2,556 multiple-choice questions from Pew Research Center's Global Attitudes surveys (2,203 questions) and the World Values Survey (353 questions). Covering politics, technology, religion, and social values, it captures diverse opinions from 153 countries. GlobalOpinionQA evaluates how well LLMs simulate complex, subjective global questions and allows comparison of model responses to real-world opinions across cultures, revealing whether models reflect global attitudes or exhibit regional biases.

(5) **General Social Survey (GSS):** The GSS, funded by the NSF and conducted by the National Opinion Research Center, is a long-running sociological study tracking U.S. residents' attitudes, behaviors, and experiences from 1972 to 2018. It repeatedly surveys similar demographic cohorts using consistent questions, enabling the analysis of temporal trends in public opinion. Rather than framing this as a forecasting task, we assess *temporal steerability*, or whether models can simulate prevailing public sentiment when prompted as respondents from specific historical years. For instance, in response to the question, "How effective are courts in dealing with criminals?" with options ['ABOUT RIGHT', 'NOT HARSH ENOUGH', 'TOO HARSH'], survey data shows that in the post-2010 decade, increasing numbers selected 'TOO HARSH' over 'ABOUT RIGHT', indicating evolving public perceptions of the justice system. Using such questions asked across multiple years, we evaluate the model's log-probability over response options and compute steerability across ten representative years. We also compute *representativeness* to assess alignment with aggregate public opinion.

(6) **CultureBank:** We hypothesize that socially-aligned models should extrapolate knowledge to align with cultural differences. To test this, we use the CultureBank dataset (Shi et al. 2024), which includes 23,000 cultural descriptors from TikTok and Reddit. These span diverse behaviors, norms, and practices, such as "Americans in France experience culture shock with electricity bills" or "In Japan, tipping is not customary." Models are evaluated on alignment with these descriptors using GPT-4 grounded entailment scores to measure how well they handle nuanced cultural and contextual variation.

(7) **CultureNLI** (Huang and Yang 2023) contains 2,700 culture-related natural language inference samples annotated by US and Indian annotators. It provides a framework to assess LLMs' cultural awareness. Premises focus on normative behaviors, with annotators from different cultures labeling entailment relationships within their cultural context. Models are evaluated by computing the entailment scores of model responses with human annotations.

4 Experiments and Results

4.1 Training

We finetuned Llama-2-chat (7B/13B/70B) (Touvron et al. 2023) and Llama-3-chat (8B) (Grattafiori et al. 2024) on ALPHA50M for one epoch using 32 A100 80 GB GPUs. To preserve conversational capabilities, we incorporated ShareGPT data (Zheng et al. 2023) and behavioral data from the CBC dataset (Khandelwal et al. 2024), which improved task performance. We compared trained models to 5-shot inference from similar or larger models like GPT-4. To verify learning, we evaluated all training tasks and assessed generated ads on TAG, TS, and TC using Perplexity, BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and BERTScore (Zhang et al.

Tasks→	OpinionQA-XL		OpinionQA		GlobalOpinionQA		CultureBank		CultureNLI		GSS	
Model (zero-shot)	**Represent-ativeness (↑)**	**Steer-ability (↑)**	**Represent-ativeness (↑)**	**Steer-ability (↑)**	**Avg Sim (↑)**	**Skew (↓)**	**Redd-it (↑)**	**Tik -Tok (↑)**	**US (↑)**	**IN (↑)**	**Represent-ativeness (↑)**	**Steer-ability (↑)**
Llama-2-7B-chat	83.61	79.09	86.18	79.18	83.6	2.2	85.93	92.08	39.2	39.5	85.46	74.23
Vicuna-7B-v1.5	72.26	77.55	77.63	77.68	84.94	1.92	64.88	55.02	**55.72**	**91.45**	85.34	74.22
Llama-2-7B-SFT -CultureBank	82.70	78.46	84.94	78.55	85.4	1.5	85.93	92.08	39.2	39.6	85.00	73.68
ALPHA-7B (Ours)	**85.15**	**81.95**	**88.43**	**81.98**	**86.69**	**1.43**	**92.39**	**95.87**	47.14	43.92	**87.94**	**75.01**
Meta-Llama-3-8B	83.71	-	77.80	-	86.82	1.52	92.39	95.87	39.18	39.58	88.64	76.01
Meta-Llama-3-8B-chat	83.93	-	79.92	-	87.01	1.59	85.38	92.89	39.18	39.54	91.99	77.79
ALPHA-8B (Ours)	**85.16**	-	**81.41**	-	**88.30**	1.89	**85.38**	**92.89**	**42.07**	**43.44**	**92.09**	**78.60**
LLama-2-13B-base	80.45	79.03	83.03	79.14	83.13	**1.45**	73.19	89.02	53.34	49.48	83.84	73.08
Llama-2-13B-chat	81.18	81.11	84.29	81.35	84.03	1.96	86.17	**92.34**	60.08	61.73	83.54	73.73
Vicuna-13B	79.06	78.73	83.44	78.85	86.99	1.91	85.93	92.08	52.07	40.23	83.88	73.21
ALPHA-13B (Ours)	**85.76**	**83.54**	**89.44**	**83.53**	**87.31**	1.49	**86.28**	92.25	**62.26**	**66.44**	**85.58**	**76.92**
Mixtral-8x7B-Instruct	84.96	82.31	88.39	82.25	79.5	2.7	87.35	88.59	59.90	60.80	90.47	79.12
Mixtral-8X7B-SFT -CultureBank	84.40	79.66	78.69	79.67	81.80	2.80	86.19	92.08	61.50	61.30	91.06	78.37
Mixtral-8x7B-DPO -CultureBank	82.70	80.22	78.79	80.90	80.50	2.60	86.19	91.74	56.30	55.40	90.68	79.72
GPT-4o	83.27	-	68.28	70.78	84.94	1.78	-	-	**80.00**	72.00	85.39	76.30
Llama-2-70B-chat	85.08	82.40	88.83	82.28	83.6	2.2	87.17	92.76	69.70	68.90	89.90	77.18
ALPHA-70B (Ours)	**86.65**	**83.23**	**89.95**	**83.31**	**86.31**	**1.67**	**88.48**	**92.65**	73.87	**73.67**	**93.88**	**81.30**

Table 1: Comparison of all models across Opinion and Culture tasks shows that **our behavior-trained models**, despite being zero-shot, **outperform or match baseline models** of similar or larger sizes across multiple benchmarks. **ALPHA-7B shows marked gains across 5 of 7 tasks**, while **ALPHA-8B, 13B, and 70B consistently improve across all benchmarks.** These results highlight that **unsupervised in-the-wild behavior signals enable strong zero-shot alignment** with human opinions and cultural values across demographics, topics, and time. **Best scores are shown in bold.** We use Llama-2 (Touvron et al. 2023) for 7B/13B/70B and Llama-3 (Grattafiori et al. 2024) for the 8B variant.

2020). Lower Perplexity scores on ground truth ads post-training indicate better alignment with target demographics. TAP was evaluated by accuracy on a balanced test set. Our models outperform strong baselines, showing that current LLMs often lack behavioral signals. Results on training tasks and ablations are in extended Appendix.

4.2 Zero-Shot Pluralistic Alignment

We evaluate the behavior-trained models in zero-shot settings across a range of cultural and opinion benchmarks. Their performance is compared against similarly sized models (7B, 8B, 13B, 70B) as well as larger models such as LLaMA-70B, Mixtral-8x7B, and GPT-4o. Results are summarized in Table 1. We find that models trained on in-the-wild behavioral signals from the ALPHA50M dataset outperform both comparable and larger models on opinion and cultural alignment tasks in zero-shot evaluations, even when the baselines are trained on clean, annotated data. These results suggest that behaviorally supervised models are not only better at modeling social nuance but also demonstrate stronger pluralistic alignment. That is, they more accurately reflect the diversity of viewpoints present in human populations, rather than collapsing toward a single consensus or majority view. This has important implications for building models that can engage fairly with a wide range of social and cultural perspectives.

Cultural Alignment without Supervised Data: As shown in Table 1, our behaviorally trained models (ALPHA-7B to 70B) outperform models explicitly trained on curated cultural tasks across both CultureBank and CultureNLI. For instance, ALPHA-7B achieves higher alignment scores on CultureBank (92.39 and 95.87) than Llama-2-7B-CultureBank (85.93 and 92.08), despite using no task-specific cultural supervision. On CultureNLI, ALPHA-13B and ALPHA-70B show the strongest alignment with both US and IN cultural norms. These results demonstrate that clean, annotated cultural datasets are less effective than our in-the-wild, annotation-free training method for achieving cultural alignment in zero-shot settings.

Zero-shot Alignment to Social Opinions: As shown in Table 1, behavior-tuned models (ALPHA-7B, 8B, 13B, 70B) consistently outperform both similarly sized and larger baselines on OpinionQA and OpinionQA-XL across representativeness and steerability metrics. These baselines, including chat-optimized models like Llama-2-chat, Vicuna, and Mixtral-Instruct, are finetuned using standard instruction-following or supervised datasets. For example, ALPHA-7B exceeds Llama-2-7B-chat by +2.25 points on OpinionQA-XL representativeness (85.15 vs. 83.61) and +2.86 on steerability (81.95 vs. 79.09). Even compared to models like Mixtral-8x7B-Instruct and GPT-4o, ALPHA-70B achieves the highest scores on OpinionQA-XL (86.65, 83.23) and OpinionQA (89.95, 83.31). These gains reflect stronger zero-shot alignment with U.S. population-level views (via representativeness) and improved simulation of subgroup-specific opinions (via steerability), despite no explicit demographic supervision. Although ALPHA50M contains only limited cues (such as region and age-gender), models generalize well across nine demographic axes including income, education, and

Model	Knowledge Cut-off	Opinion Alignment
Llama-2-7B-chat	Sept, 2022	82.97
Vicuna-7B-v1.5	Sept, 2022	79.29
ALPHA-7B (Ours)	Sept, 2022	**84.42**
Meta-Llama-3-8B	March 2023	83.24
Meta-Llama-3-8B-Instruct	March 2023	83.65
DeepSeek-R1-Distill-Llama-8B	March 2023	82.76
ALPHA-8B (Ours)	March 2023	**84.89**
Llama-2-13B	Sept, 2022	81.38
Llama-2-13B-chat	Sept, 2022	82.97
ALPHA-13B (Ours)	Sept, 2022	**84.72**
Llama-2-70B-chat	Sept, 2022	84.92
ALPHA-70B (Ours)	Sept, 2022	**85.16**

Table 2: Evaluation based on surveys from **March to August 2023**, after the pretraining cutoff of base models, shows that ALPHA models (finetuned LLaMA 2 & 3) exhibit significantly higher alignment with public opinion. This indicates that behavioral training enables models to generalize better to social views beyond their original knowledge window.

political affiliation. These results show that in-the-wild behavioral signals are a scalable and effective alternative to curated supervision for social alignment.

Generalization across Topics: Despite limited topical overlap between ALPHA50M's training data (Meta Ads) and OpinionQA-XL (only 28.25% overlap), ALPHA-13B demonstrates strong alignment across a wide range of unseen domains, including Religion, Technology, Misinformation, and Online Dating. ALPHA-13B consistently outperforms models across eight diverse categories, achieving over 91 on all topics (e.g., 93.10 in Journalism, 92.77 in Religion, 93.93 in Ethnicity). This suggests that behavior-based learning enables robust generalization beyond the observed training distribution and captures stable patterns of social alignment even in novel thematic contexts.

Knowing One Culture Helps Learn Others: Although ALPHA is trained only on behavioral signals from US-targeted Meta Ads, it generalizes effectively to global and culturally distinct benchmarks. ALPHA-13B outperforms similarly sized chat-tuned models on GlobalOpinionQA (85.61 vs. 83.75 for Llama-2-13B-chat) and CultureNLI (91.88 vs. 89.14), reflecting improved alignment with global and Indian cultural judgments. This generalization likely arises from the cosmopolitan composition of US online audiences, whose behavioral signals encode a wide range of cultural, ethnic, and ideological patterns. These results suggest that behavior data from one region, when diverse and representative, can support cross-cultural opinion modeling and alignment.

Temporal Alignment: We evaluate models' ability to simulate shifts in public opinion over time using the General Social Survey (GSS), a long-standing benchmark of US societal attitudes. By prompting models to respond as individuals from specific historical years, we compute temporal steerability as the alignment between predicted and actual survey responses for each year. ALPHA models, trained on temporally grounded behavioral data, exhibit stronger steerability than instruction-tuned counterparts, capturing longitudinal opinion changes on issues such as gender roles, race, and political affiliation (e.g., GSS steerability for ALPHA-70B: 81.30 vs. 77.18 for LLaMA-2-70B-chat). Additionally, we test model generalization on surveys conducted between March and August 2023, after the knowledge cutoffs of both LLaMA-2 (September 2022) and LLaMA-3 (March 2023). Because these surveys fall beyond the pretraining data window, they serve as uncontaminated evaluation benchmarks. This isolates the effects of social learning and prevents leakage from memorized text. ALPHA models outperform all size-matched baselines across these temporally out-of-distribution surveys (e.g., ALPHA-13B: 84.72 vs. 82.97 for LLaMA-2-13B-chat; ALPHA-8B: 84.89 vs. 83.65 for LLaMA-3-8B-Instruct). The ability to extrapolate to unseen, forward-looking opinions is essential for real-world deployment, where models must stay aligned with public sentiment as it evolves. Robustness to temporal shift is thus a key criterion for holistic and pluralistic alignment.

Behavior Task Contribution to Pluralistic Alignment: We perform an ablation analysis to isolate the contribution of individual behavior tasks to alignment performance. Training ALPHA-13B on 50K samples per task, we find that Transsuasion and TAG drive the largest improvements in opinion representation (OpinionQA-XL: 84.82 and 84.57), while TAP yields the highest gains in cultural generalization (CNLI-IN: 68.39). GSS scores remain consistent across tasks, suggesting temporal robustness is less sensitive to task type. Importantly, training on the full ALPHA50M dataset across all tasks delivers the strongest results across all benchmarks, including OpinionQA-rep (86.65) and GSS (85.58), underscoring the complementary nature of the tasks and the benefits of learning from diverse behavioral signals. This ablation highlights that while individual tasks contribute targeted gains, broad pluralistic alignment emerges only through large-scale, multi-task behavioral training.

5 Conclusion

This work demonstrates that by observing actions, we can infer opinions leveraging models trained on sparse behavioral signals to achieve pluralistic alignment with human culture and opinions. This approach offers a scalable, dynamic alternative to traditional culture-specific data annotation, overcoming the limitations of expert dependency, high costs, static datasets, and cultural biases. We curate a new dataset for alignment and show through zero-shot evaluation on benchmarks like OpinionQA and GlobalOpinionQA that training on such data yields state-of-the-art human and cultural alignment, highlighting the potential of LLMs to model behavior from sparse signals and advance the understanding of opinion dynamics.

Acknowledgments

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, the Center of Design and New Media, and the Center of Excellence in Healthcare at IIIT Delhi. This work is partially supported by NSF AI Institute-2229873, NSF RI-

2223292, an Amazon research award, and an Adobe gift fund. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

AlKhamissi, B.; ElNokrashy, M.; AlKhamissi, M.; and Diab, M. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Berinsky, A. J. 2017. Measuring public opinion with surveys. *Annual review of political science*, 20(1): 309–329.

Bhattacharyya, A.; Singla, Y. K.; Krishnamurthy, B.; Shah, R. R.; and Chen, C. 2023. A Video Is Worth 4096 Tokens: Verbalize Videos To Understand Them In Zero Shot. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9822–9839. Singapore: Association for Computational Linguistics.

Biderman, S.; Bicheno, K.; and Gao, L. 2022. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

Bureau, U. S. C. 2024. 2020 Census Results.

Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; and Hershcovich, D. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67. United States: Association for Computational Linguistics (ACL). Publisher Copyright: © 2023 Association for Computational Linguistics.; 1st Workshop on Cross-Cultural Considerations in NLP, C3NLP 2023 ; Conference date: 05-05-2023.

Daniels-Koch, O.; and Freedman, R. 2022. The Expertise Problem: Learning from Specialized Feedback. In *NeurIPS ML Safety Workshop*.

Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Fazio, R. H.; and Zanna, M. P. 1981. Direct experience and attitude-behavior consistency. In *Advances in experimental social psychology*, volume 14, 161–202. Elsevier.

General Social Survey. 2025. General Social Survey. Accessed: 2025-02-15.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Wyatt, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Guzmán, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Thattai, G.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I.; Misra, I.; Evtimov, I.; Zhang, J.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Prasad, K.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; El-Arini, K.; Iyer, K.; Malik, K.; Chiu, K.; Bhalla, K.; Lakhotia, K.; Rantala-Yeary, L.; van der Maaten, L.; Chen, L.; Tan, L.; Jenkins, L.; Martin, L.; Madaan, L.; Malo, L.; Blecher, L.; Landzaat, L.; de Oliveira, L.; Muzzi, M.; Pasupuleti, M.; Singh, M.; Paluri, M.; Kardas, M.; Tsimpoukelli, M.; Oldham, M.; Rita, M.; Pavlova, M.; Kambadur, M.; Lewis, M.; Si, M.; Singh, M. K.; Hassan, M.; Goyal, N.; Torabi, N.; Bashlykov, N.; Bogoychev, N.; Chatterji, N.; Zhang, N.; Duchenne, O.; Çelebi, O.; Alrassy, P.; Zhang, P.; Li, P.; Vasic, P.; Weng, P.; Bhargava, P.; Dubal, P.; Krishnan, P.; Koura, P. S.; Xu, P.; He, Q.; Dong, Q.; Srinivasan, R.; Ganapathy, R.; Calderer, R.; Cabral, R. S.; Stojnic, R.; Raileanu, R.; Maheswari, R.; Girdhar, R.; Patel, R.; Sauvestre, R.; Polidoro, R.; Sumbaly, R.; Taylor, R.; Silva, R.; Hou, R.; Wang, R.; Hosseini, S.; Chennabasappa, S.; Singh, S.; Bell, S.; Kim, S. S.; Edunov, S.; Nie, S.; Narang, S.; Raparthy, S.; Shen, S.; Wan, S.; Bhosale, S.; Zhang, S.; Vandenhende, S.; Batra, S.; Whitman, S.; Sootla, S.; Collot, S.; Gururangan, S.; Borodinsky, S.; Herman, T.; Fowler, T.; Sheasha, T.; Georgiou, T.; Scialom, T.; Speckbacher, T.; Mihaylov, T.; Xiao, T.; Karn, U.; Goswami, V.; Gupta, V.; Ramanathan, V.; Kerkez, V.; Gonguet, V.; Do, V.; Vogeti, V.; Albiero, V.; Petrovic, V.; Chu, W.; Xiong, W.; Fu, W.; Meers, W.; Martinet, X.; Wang, X.; Wang, X.; Tan, X. E.; Xia, X.; Xie, X.; Jia, X.; Wang, X.; Goldschlag, Y.; Gaur, Y.; Babaei, Y.; Wen, Y.; Song, Y.; Zhang, Y.; Li, Y.; Mao, Y.; Coudert, Z. D.; Yan, Z.; Chen, Z.; Papakipos, Z.; Singh, A.; Srivastava, A.; Jain, A.; Kelsey, A.; Shajnfeld, A.; Gangidi, A.; Victoria, A.; Goldstand, A.; Menon, A.; Sharma, A.; Boesenberg, A.; Baevski, A.; Feinstein, A.; Kallet, A.; Sangani, A.; Teo, A.; Yunus, A.; Lupu, A.; Alvarado, A.; Caples, A.; Gu, A.; Ho, A.; Poulton, A.; Ryan, A.; Ramchandani, A.; Dong, A.; Franco, A.; Goyal, A.; Saraf, A.; Chowdhury, A.; Gabriel, A.; Bharambe, A.; Eisenman, A.; Yazdan, A.; James, B.; Maurer, B.; Leonhardi, B.; Huang, B.; Loyd, B.; Paola, B. D.; Paranjape, B.; Liu, B.; Wu, B.; Ni, B.; Hancock, B.; Wasti, B.; Spence, B.; Stojkovic, B.; Gamido, B.; Montalvo, B.; Parker, C.; Burton, C.; Mejia, C.; Liu, C.; Wang, C.; Kim, C.; Zhou, C.; Hu, C.; Chu, C.-H.; Cai, C.; Tindal, C.; Feichtenhofer, C.; Gao, C.; Civin, D.; Beaty, D.; Kreymer, D.; Li, D.; Adkins,

D.; Xu, D.; Testuggine, D.; David, D.; Parikh, D.; Liskovich, D.; Foss, D.; Wang, D.; Le, D.; Holland, D.; Dowling, E.; Jamil, E.; Montgomery, E.; Presani, E.; Hahn, E.; Wood, E.; Le, E.-T.; Brinkman, E.; Arcaute, E.; Dunbar, E.; Smothers, E.; Sun, F.; Kreuk, F.; Tian, F.; Kokkinos, F.; Ozgenel, F.; Caggioni, F.; Kanayet, F.; Seide, F.; Florez, G. M.; Schwarz, G.; Badeer, G.; Swee, G.; Halpern, G.; Herman, G.; Sizov, G.; Guangyi; Zhang; Lakshminarayanan, G.; Inan, H.; Shojanazeri, H.; Zou, H.; Wang, H.; Zha, H.; Habeeb, H.; Rudolph, H.; Suk, H.; Aspegren, H.; Goldman, H.; Zhan, H.; Damlaj, I.; Molybog, I.; Tufanov, I.; Leontiadis, I.; Veliche, I.-E.; Gat, I.; Weissman, J.; Geboski, J.; Kohli, J.; Lam, J.; Asher, J.; Gaya, J.-B.; Marcus, J.; Tang, J.; Chan, J.; Zhen, J.; Reizenstein, J.; Teboul, J.; Zhong, J.; Jin, J.; Yang, J.; Cummings, J.; Carvill, J.; Shepard, J.; McPhie, J.; Torres, J.; Ginsburg, J.; Wang, J.; Wu, K.; U, K. H.; Saxena, K.; Khandelwal, K.; Zand, K.; Matosich, K.; Veeraraghavan, K.; Michelena, K.; Li, K.; Jagadeesh, K.; Huang, K.; Chawla, K.; Huang, K.; Chen, L.; Garg, L.; A, L.; Silva, L.; Bell, L.; Zhang, L.; Guo, L.; Yu, L.; Moshkovich, L.; Wehrstedt, L.; Khabsa, M.; Avalani, M.; Bhatt, M.; Mankus, M.; Hasson, M.; Lennie, M.; Reso, M.; Groshev, M.; Naumov, M.; Lathi, M.; Keneally, M.; Liu, M.; Seltzer, M. L.; Valko, M.; Restrepo, M.; Patel, M.; Vyatskov, M.; Samvelyan, M.; Clark, M.; Macey, M.; Wang, M.; Hermoso, M. J.; Metanat, M.; Rastegari, M.; Bansal, M.; Santhanam, N.; Parks, N.; White, N.; Bawa, N.; Singhal, N.; Egebo, N.; Usunier, N.; Mehta, N.; Laptev, N. P.; Dong, N.; Cheng, N.; Chernoguz, O.; Hart, O.; Salpekar, O.; Kalinli, O.; Kent, P.; Parekh, P.; Saab, P.; Balaji, P.; Rittner, P.; Bontrager, P.; Roux, P.; Dollar, P.; Zvyagina, P.; Ratanchandani, P.; Yuvraj, P.; Liang, Q.; Alao, R.; Rodriguez, R.; Ayub, R.; Murthy, R.; Nayani, R.; Mitra, R.; Parthasarathy, R.; Li, R.; Hogan, R.; Battey, R.; Wang, R.; Howes, R.; Rinott, R.; Mehta, S.; Siby, S.; Bondu, S. J.; Datta, S.; Chugh, S.; Hunt, S.; Dhillon, S.; Sidorov, S.; Pan, S.; Mahajan, S.; Verma, S.; Yamamoto, S.; Ramaswamy, S.; Lindsay, S.; Lindsay, S.; Feng, S.; Lin, S.; Zha, S. C.; Patil, S.; Shankar, S.; Zhang, S.; Zhang, S.; Wang, S.; Agarwal, S.; Sajuyigbe, S.; Chintala, S.; Max, S.; Chen, S.; Kehoe, S.; Satterfield, S.; Govindaprasad, S.; Gupta, S.; Deng, S.; Cho, S.; Virk, S.; Subramanian, S.; Choudhury, S.; Goldman, S.; Remez, T.; Glaser, T.; Best, T.; Koehler, T.; Robinson, T.; Li, T.; Zhang, T.; Matthews, T.; Chou, T.; Shaked, T.; Vontimitta, V.; Ajayi, V.; Montanez, V.; Mohan, V.; Kumar, V. S.; Mangla, V.; Ionescu, V.; Poenaru, V.; Mihailescu, V. T.; Ivanov, V.; Li, W.; Wang, W.; Jiang, W.; Bouaziz, W.; Constable, W.; Tang, X.; Wu, X.; Wang, X.; Wu, X.; Gao, X.; Kleinman, Y.; Chen, Y.; Hu, Y.; Jia, Y.; Qi, Y.; Li, Y.; Zhang, Y.; Zhang, Y.; Adi, Y.; Nam, Y.; Yu; Wang; Zhao, Y.; Hao, Y.; Qian, Y.; Li, Y.; He, Y.; Rait, Z.; DeVito, Z.; Rosnbrick, Z.; Wen, Z.; Yang, Z.; Zhao, Z.; and Ma, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv:2301.01768.

Huang, J.; and Yang, D. 2023. Culturally Aware Natural Language Inference. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hwang, E.; Majumder, B.; and Tandon, N. 2023. Aligning Language Models to User Opinions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5906–5919. Singapore: Association for Computational Linguistics.

Jiang, H.; Beeferman, D.; Roy, B.; and Roy, D. 2022. CommunityLM: Probing Partisan Worldviews from Language Models. arXiv:2209.07065.

Johnson, R. L.; Pistilli, G.; Menédez-González, N.; Duran, L. D. D.; Panai, E.; Kalpokiene, J.; and Bertulfo, D. J. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv:2203.07785.

Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2024. A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925.

Kelley, S.; and Mirer, T. W. 1974. The simple act of voting. *American Political Science Review*, 68(2): 572–591.

Khandelwal, A.; Agrawal, A.; Bhattacharyya, A.; Kumar, Y.; Singh, S.; Bhattacharya, U.; Dasgupta, I.; Petrangeli, S.; Shah, R. R.; Chen, C.; and Krishnamurthy, B. 2024. Large Content And Behavior Models To Understand, Simulate, And Optimize Content And Behavior. In *The Twelfth International Conference on Learning Representations*.

Khanuja, S.; Ramamoorthy, S.; Song, Y.; and Neubig, G. 2024. An image speaks a thousand words, but can everyone listen? On translating images for cultural relevance. arXiv:2404.01247.

Kochhar, R. 2023. Survey Methodology: The American Trends Panel survey methodology.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Masoud, R. I.; Liu, Z.; Ferianc, M.; Treleaven, P.; and Rodrigues, M. 2024. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. arXiv:2309.12342.

Meta Platforms, I. 2025. Facebook Ads Library. https://www.facebook.com/ads/library/. Accessed: March 2024.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. arXiv:2305.14456.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Ryan, M. J.; Held, W.; and Yang, D. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

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

Shi, W.; Li, R.; Zhang, Y.; Ziems, C.; Horesh, R.; de Paula, R. A.; Yang, D.; et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Silver, L.; Huang, C.; Clancy, L.; and Prozorovsky, A. 2024. Methodology: About Pew Research Center's Spring 2024 Global Attitudes Survey.

Simmons, G. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. arXiv:2209.12106.

Singh, S. K.; Singla, Y. K.; I, H. S.; and Krishnamurthy, B. 2025. Measuring And Improving Persuasiveness Of Generative Models. In *The Thirteenth International Conference on Learning Representations*.

Stouffer, S. A.; Lumsdaine, A. A.; Lumsdaine, M. H.; Williams Jr, R. M.; Smith, M. B.; Janis, I. L.; Star, S. A.; and Cottrell Jr, L. S. 1949. The American soldier: Combat and its aftermath.(Studies in social psychology in World War II), Vol. 2.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, W.; Jiao, W.; Huang, J.; Dai, R.; tse Huang, J.; Tu, Z.; and Lyu, M. R. 2024. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. arXiv:2310.12481.

Weidinger, L.; et al. 2022. Ethical and Social Risks of Harm from Language Models. *arXiv preprint arXiv:2204.05862*.

Yamshchikov, I. P.; et al. 2025. AI chatbots systematically advise women and minorities to negotiate lower salaries. Computer World / Technical University of Applied Sciences Würzburg-Schweinfurt study.

Zhang, A.; Yuksekgonul, M.; Guild, J.; Zou, J.; and Wu, J. C. 2023. ChatGPT Exhibits Gender and Racial Biases in Acute Coronary Syndrome Management. *arXiv preprint arXiv:2311.14703*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zhao, S.; Dang, J.; and Grover, A. 2023. Group Preference Optimization: Few-Shot Alignment of Large Language Models. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.