Align Via Actions : Learning Behavior Aligns LLMs With Human Opinions in Zero-Shot

Aanisha Bhattacharyya*, **Susmit Agrawal***, **Yaman K Singla***, **Tarun Menta**, **Nikitha SR**, **Balaji Krishnamurthy**

Adobe Media and Data Science Research (MDSR)

Abstract

Large language models (LLMs) have become ubiquitous in various applications, but aligning them with societal expectations remains challenging. To align LLMs with humans, current alignment methods rely heavily on human-annotated datasets, which are expensive, difficult to scale, and often biased toward specific demographic subgroups. We introduce a novel approach for LLM alignment by training on behavioral data. Our approach is based on the maxim in psychology that actions (behavior) have a strong consistency with opinions. Leveraging this insight, we developed AlignViaActions (AVA50M) comprising over 50 million samples derived from 1.5 million advertisements, including content and demographic viewing behaviors. We train LLMs on AVA50M, demonstrating significant improvements over existing alignment techniques across multiple societal and cultural alignment benchmarks, including GlobalOpinionQA, OpinionQA, CultureNLI, and CultureBank. Through this, we demonstrate that by observing and learning from behavior, LLMs can infer the underlying opinions and cultural norms. This approach addresses key limitations of current methods, offering improved scalability, demographic representation, and adaptability to evolving societal views. Our results suggest the potential for behavioral data to replace or complement traditional expert-annotation-based alignment techniques. Our datasets and code are available at https://behavior-in-the-wild.github.io/align-via-actions.

1 Introduction

"Only in *actions* can you fully recognize the forces operative in social behavior" - Milgram (1974)

LLM-powered chat assistants have exploded in popularity, with the popular ones being used by more than 100 million active users per week (OpenAI, 2023). Their usage is now ubiquitous across many open-ended applications, including writing agents and marketing chatbots. Conventionally, the procedure to train LLMs behind these chatbots consists of pre-training on a large dataset with the task of "predicting the next token". However, successful memorization of human knowledge does not assure a model's propensity to perform as per societal expectations. To align these models with societal expectations, practitioners next use algorithms of Instruction Finetuning (IFT) (Ouyang et al., 2022) and Reinforcement Learning with Human Feedback (RLHF) (Kaufmann et al., 2024) or Direct Preference Optimization (DPO) (Rafailov et al., 2023; Zhao et al., 2023) to optimize models for attributes such as helpfulness and harmlessness and give them their chat assistant persona.

Compared to the unsupervised pertaining datasets, the datasets employed in the final stages of IFT and RLHF are notably more curated and of superior quality compared to those used in earlier stages. Therefore, these datasets require substantial human annotation, rendering them costly to produce and consequently smaller in scale. The reliance on annotators also presents challenges in scaling beyond a limited set of demographic groups. Studies like Santurkar et al. (2023a); Durmus et al. (2023); Ryan et al. (2024); AlKhamissi et al. (2024) investigating IFT and RLHF demonstrate that because of utilizing a restricted annotator pool (e.g., as reported in OpenAI's InstructGPT paper (Ouyang et al., 2022)), the resultant models align heavily with specific subgroups, most common being young, liberal, high-income, well-

*Equal Contribution. Get in touch with us at behavior-in-the-wild@googlegroups.com.

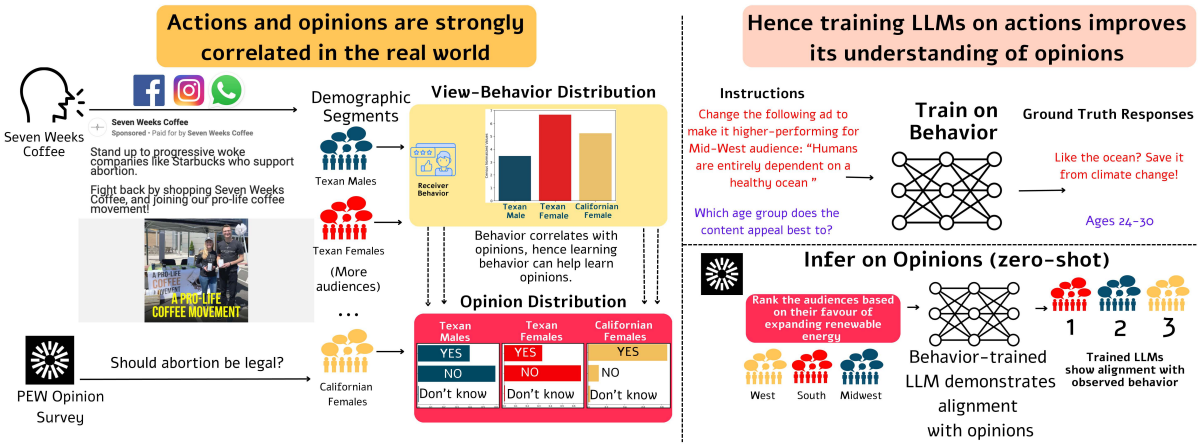

Figure 1: Behavior and Opinions are strongly correlated. The behavioral data, which contains the ad content, the audience, and the behavior that the audience showed towards the ad, helps in understanding the audience. While behavior is already being collected at scale, it is conventionally not used to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

educated, and atheistic groups. Moreover, the dependence on human annotators impedes the ability to rapidly update models in response to evolving societal perspectives on contentious issues such as same-sex marriage and abortion, or to comprehensively address the full spectrum of socially relevant topics (Bai et al., 2022; Durmus et al., 2023). Consequently, there is a pressing need for a scalable method that can align LLMs with diverse subgroups while maintaining adaptability to changing social norms.

Moreover, building alignment datasets is not a one-time effort and scaling datasets to capture diverse opinions or even maintaining their recency presents significant challenges. The Pew Research Center's opinion surveys, among the most extensive and frequently cited, exemplify these difficulties. Despite decades of expertise, Pew's response rates have plummeted from 43% in the 1990s to below 4% in 2024, complicating survey execution (Berinsky, 2017; Kochhar, 2023; Silver et al., 2024). While building alignment datasets at scale remains problematic, behavioral data acquisition is comparatively straightforward. Digital analytics routinely capture user behaviors such as likes, shares, comments, and subscriptions for various purposes.

Psychological research, while done on smaller scales, has demonstrated that behavior can serve as both an outcome and an indicator of attitude (Fazio and Zanna, 1981). In psychological research, strong attitude-behavior consistency has been observed across diverse domains, from voting patterns (Kelley and Mirer, 1974) to military combat performance (Stouffer et al., 1949). Paradoxically, during LLM pretraining, behavioral data from digital analytics sources (*e.g.*, upvotes, likes) is often discarded as noise (Biderman et al., 2022; Penedo et al., 2023; Khandelwal et al., 2023). Subsequently, to align LLMs, behavioral data is again neglected and instead expert annotations (Shi et al., 2024; Huang and Yang, 2023; Bai et al., 2022) or opinion surveys (Hwang et al., 2023; Zhao et al., 2023; Li et al., 2024) are used.

Building on the established correlation between attitudes and behaviors in psychological literature, we investigate the question: "*Can LLMs learn opinions at scale through observation of group behaviors sampled from digital analytics sources?*" To explore this hypothesis, we utilize data from the Meta Ads Library (Facebook, 2024), which contains advertisements 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 initially demonstrate a strong correlation between opinions captured in established surveys from Pew Research and user behaviors as reflected in Meta Ads viewership statistics (§3.1). Subsequently, building on this positive result, we develop instruction fine-tuning tasks to train LLMs in predicting user behavior based on ad content (§3.3). Results indicate that the fine-tuned LLMs significantly outper-

form their base and chat versions on tasks assessing alignment with societal opinions and cultural norms, providing evidence that behavioral learning can enhance LLM alignment with human opinions (Table 2). Moreover, in zero-shot evaluations, our behavior-trained models surpass those trained on opinion datasets. As a contribution to the field, we release the instruction training set comprising of 50 million instructions as the AVA50M (AlignViaActions 50 Million) dataset. AVA50M is designed to facilitate the large-scale alignment of LLMs through behavioral instruction tuning.

With this work, we make the following contributions:

- We propose a novel approach of aligning LLMs with group opinions by fine-tuning them using *behavioral signals* derived from web analytics. This method is grounded in psychological research demonstrating that behavior strongly correlates with attitudes. Our approach addresses several limitations of existing LLM alignment techniques that rely on expert annotations. It offers improved scalability across diverse demographic groups and topics, adaptability to evolving opinions over time, reduced dependence on a limited pool of annotators to represent group perspectives, and mitigation of operational challenges such as high costs and quality control issues.
- We demonstrate that models trained on behavioral data, even with sparse opinion signals, outperform those trained on expert annotations or opinion surveys in zero-shot evaluations. This superiority is evidenced across four diverse datasets: OpinionQA, GlobalOpinionQA, CultureBench, and CultureNLI. Our findings suggest that behavioral data can be effectively utilized for LLM alignment.
- We introduce the **AVA50M (AlignViaActions 50Million) dataset**, a comprehensive instruction training set derived from 1.5 million advertisements by over 120,000 advertisers in the Meta Ads Archive. AVA50M comprises 50 million instruction training samples designed to teach LLMs about human behavior, significantly surpassing existing datasets in scale (Table 1). Each instruction incorporates advertisement caption, advertiser information, publication date, media verbalization, and target audience. We release AVA50M to facilitate large-scale LLM alignment and for further research in this domain.
- We present **OpinionsQA-XL**, a substantial expansion of the OpinionsQA dataset (Santurkar et al., 2023b), used to evaluate human-LLM opinion alignment based on PEW survey results. We have expanded the dataset from 1,498 questions covering 15 surveys to over 14,000 questions, encompassing the complete set of 117 surveys. This substantial expansion provides a more comprehensive and robust tool for assessing LLM alignment with human opinions across a broader range of topics and time periods.

Dataset	#Samples	In-the-wild?
OpinionQA	1498	✗
Global-OpinionQA	2556	✗
CultureBank	23K	✗
OpinionQA-XL (Ours)	14554	✗
AVA50M (AlignViaActions) (Ours)	50M	✓

Table 1: Comparison of the Opinion and Culture Alignment Datasets present in literature with our datasets. See Table A3 for a task breakdown of AVA50M.

2 Related Works

Opinion and Culture Alignment of LLMs: 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 et al., 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; OpenAI, 2023; Daniels-Koch and Freedman, 2022) or instruction-based fine-tuning on opinion or cultural data (Huang and Yang, 2023; Zhao et al., 2023; Li et al., 2024; Shi et al., 2024). However, these techniques often rely on explicit human annotations, which are resource-intensive and prone to errors. In contrast, our work demonstrates that LLMs can be effectively aligned with human opinions and cultural norms using in-the-wild behavioral signals, eliminating the need for explicit annotations.

Measuring Opinion Alignment: A few recent works have also attempted to measure alignment to human opinions and cultures: Durmus et al. (2023) propose checking

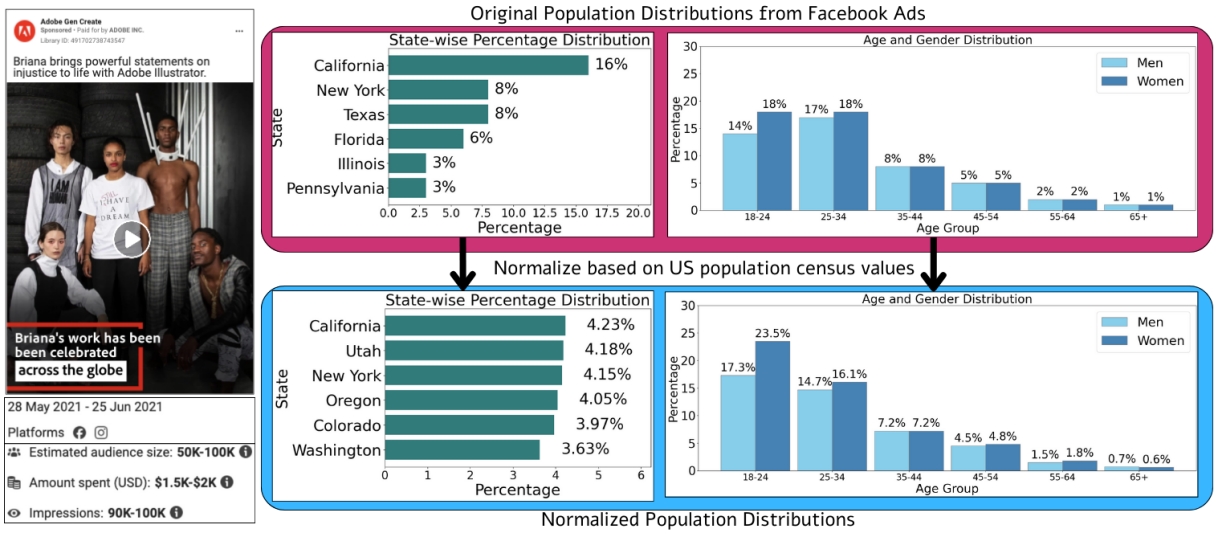

Figure 2: A sample advertisement from the Meta Ad Library.

the entailment of LLM responses against human responses as a proxy of human opinions. Works like Santurkar et al. (2023b); Shi et al. (2024) propose using public opinion surveys to estimate alignment. They propose metrics to quantify the alignment of LLMs with specific demographic groups. All of these works show that LLMs tend to be more similar to the opinions of certain populations (USA, Europe, South America) and certain groups (left-leaning, highly educated, rich, and atheists). We adopt the metrics proposed in these works to measure the alignment of our models to human culture and opinions.

3 Setup

In this section, we describe the process to collect and clean the behavioral data, the correlation of behavior with opinions, and the tasks designed to train LLMs on behavioural signals.

3.1 Collecting Behavior Data

We collect behavioral data from Meta's Ad Library, encompassing advertisement content, creation date, expenditure, impressions, language, and demographic delivery metrics. Fig. 2 illustrates an exemplar ad by "Adobe Gen Create". Ad delivery across regions is influenced by both state population and potential audience interest. To isolate the latter factor, we normalize regional delivery by state population using 2020 census data (Bureau, 2024). This normalization yields a ranked list of states for each ad, with higher rankings indicating greater potential interest. Additionally, we process marketer behavior signals in the form of 'spend' per unit 'impressions' to estimate the return on investment for each advertisement, calculated as impressions per unit spent.

The collected ads comprise 1,494,781 advertisements, reduced to 1,474,367 after removing duplicates, ads with empty bodies, missing page links, dates, or location information. URLs within ads are standardized to '[URL]'. Ads are grouped by publishing page, treating brand subsidiaries as distinct entities (e.g., 'Amazon Prime' and 'Amazon Alexa'). The final dataset encompasses 122,636 unique advertisers. Given our focus on LLMs, we convert all ad content to text-to-text format. Media elements (images or videos) are verbalized using schemes described in Bhattacharyya et al. (2023) and concatenated with the ad text. For multi-frame ads (405,485 instances), verbalized contents are combined into a single text body.

3.2 Correlation Analysis

Quantifying behavioral similarity between states: We employ Spearman's rank correlation to assess the similarity of ad-related behaviors between pairs of states $(S1, S2)$. Let (X_i^a) and (X_j^a) denote the ranks of states (i) and (j) for ad (a), respectively. The Spearman's rank correlation coefficient $(\rho_{i,j})$ between states (i) and (j) is calculated as:

$$\rho_{i,j} = 1 - \frac{6\sum_a (X_i^a - X_j^a)^2}{n(n^2-1)} \quad (1)$$

where n represents the total number of ads analyzed. Fig. A5 shows the behavioral correlation heatmap for all states.

Quantifying opinion alignment between states: Following the methodology of Santurkar et al. (2023b), we employ Wasserstein

distance (WD) to compute similarity scores between opinion distributions of all state pairs. For each state pair $(S1, S2)$ with corresponding opinion distributions $(D1, D2)$, we calculate the similarity score as:

$$\text{similarity (S1,S2)} = \sum(1 - WD(D1, D2)) \quad (2)$$

Fig. A6 shows the opinion correlation between US regions as per OpinionQA-XL.

Correlating opinion-based and behavior-based similarities: We analyze the alignment between opinions expressed in PEW surveys and behaviors observed in Meta ads through correlation analysis. Our results reveal a strong positive correlation (r = 0.83) between regional opinion data from PEW surveys and behavioral data from Meta, indicating substantial congruence between these distinct data sources. Extending this methodology, we examine correlations between age-gender-specific opinions from PEW and corresponding behaviors from the Meta Ads dataset, yielding a positive correlation of r = 0.65. These findings demonstrate substantial correlations between behaviors and opinions across two independent data sources, suggesting a high degree of consistency in public attitudes and actions.

3.3 Constructing AVA50M Dataset

We propose four key tasks to train LLMs on behavioral data derived from Meta ads: (1) A content simulation task generates targeted advertisements for specific audiences (TAG), teaching LLMs audience-preferred phraseology across regions, age groups, and genders. (2) An audience prediction task (TAP) that trains LLMs to infer target audiences from given content, fostering an implicit understanding of audience-content relationships. (3) Transcreation tasks (TC) instructing LLMs to adapt content semantically from one audience segment to another, maintaining core meaning while adjusting presentation. (4) Transsuasion tasks teaching LLMs to generate more persuasive language as judged by higher advertiser budget allocation (TSB) or higher engagement from audiences (TSE). Table A3 contains the distribution of the data for each type of task. We detail the dataset curation process for each task below. A diagrammatic representation of the data curation process is given in Fig. A3.

(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 viewed the ad more compared to others. Given the demographic segments that preferred the ad most, the LLM is instructed to perform the task of generating the ad for the given demographics. 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. Section A.1 provides the instruction format for the task.

(2) Target Audience Prediction (TAP) This task trains LLMs to predict the most receptive audience for a given advertisement. The model receives an ad as input, along with parameters from the TAG task, excluding the target audience. 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 advertisement. 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. The instruction format for this task is provided in Section A.1.

(3) **Transsuasion Tasks**: Transsuasion, as defined by Singh et al. (2024), involves transforming non-persuasive content into persuasive content while preserving communication factors such as audience, sender, time, and channel. We extend this concept to enhance audience engagement (TS-E) and advertiser budget (TS-B) while controlling for other variables. TS-B tasks require the LLM to generate advertisements potentially receiving higher marketing budgets, indicating an increase in the marketer's confidence. TS-E tasks focus on generating ads with higher engagement for specific demographic segments (age, gender, or location), while maintaining consistent budgets to avoid budget as a confounding variable.

TS-E is divided into transsuasion across Age (TS-A), Gender (TS-G), and Region (TS-R) transsuasion.

To create samples for various TS tasks, we first create 50 million advertisement pairs ($Ad1$, $Ad2$) from common brands targeting identical demographics. Next, we ensure semantic and phraseological similarity in the ad pair using two filters: (1) SentenceBERT similarity > 0.7 (Reimers and Gurevych, 2019), and (2) Levenshtein Distance > 15 words. Next, for TS-B, pairs are arranged with $Ad1$ having a lower budget than $Ad2$. Similarly, in TS-E, the pair is arranged such that $Ad1$ exhibits lower engagement for a specific demographic segment than $Ad2$. The LLM is tasked with transforming $Ad1$ into $Ad2$. Section A.1 details the instruction format for these tasks.

(4) **Transcreation (TC)** Transcreation involves adapting a message that resonates with a source audience to align with a target audience while preserving the original semantics (Singh et al., 2024; Khanuja et al., 2024). We formulate transcreation tasks based on age (TC-A), gender (TC-G), and regions (TC-R), focusing on converting ads appealing to specific demographic groups to appeal to different target audiences. To generate transcreation samples, we identified 33.4 million pairs of ads from the same marketer with equivalent meanings. For region-based transcreation, we retained those pairs where a region's rank differed by at least three positions between the original and target ads. Similarly, age-based and gender-based transcreation tasks retained pairs where the target demographic was not top-ranked in the source ad but ranked first in the target ad.

Test Set Creation. For each of the above tasks, we hold out a portion of the created data to serve as the test set to validate LLM's training on behavioral data. We create two types of test sets: (1) holding out ads from a set of advertisers, and (2) holding out ads published after a certain date (June 2023). The first set evaluates the LLM's ability to generalize to use cases of unseen advertisers. The second assesses the LLM's capacity to adapt to new contexts emerging after the training period. These sets primarily serve to monitor the learning process.

3.4 Evaluating Cultural and Opinion Alignment in Behavior-Trained Models

To investigate the cultural and opinion alignment of models trained using the AVA50M dataset, we conducted zero-shot evaluations using several literature cultural and opinion alignment benchmarks covered next.

(1) **OpinionQA** (Santurkar et al., 2023b) derived from Pew Opinion Surveys, quantifies the alignment of LLMs with the general US population and specific demographic groups. It assesses whether LLMs exhibit preferential alignment with particular viewpoints (e.g., conservative vs. liberal) across diverse topics. The dataset comprises 1,498 survey questions from 15 distinct PEW opinion surveys, spanning 60 demographic groups, with data up to July 2021. Santurkar et al. (2023b) demonstrated that LLMs exhibiting superior zero-shot performance on this dataset indicate stronger opinion alignment, making it an ideal probe for assessing the opinion alignment of behaviorally trained LLMs. The study introduces an alignment metric to compare opinion distributions of LLM against general public opinions (*Representativeness*, Eq.3) and also with target demographics (*Steerability*, Eq. 4).

(2) **OpinionQA-XL**: We significantly expanded OpinionQA to cover PEW surveys up to the last survey of November 2022, resulting in OpinionQA-XL. This dataset contains 14,554 questions from 119 surveys, a 9.7-fold increase over OpinionQA. We extracted questions from survey PDFs using optical character recognition, with GPT-4-turbo for error correction, and finally, verified and corrected the questions extracted manually. OpinionQA-XL adds 68 new topics over OpinionQA, including topics like Climate Change, Space Tourism, and Digital Economy, significantly broadening the dataset's scope.

(3) **GlobalOpinionQA** (Durmus et al., 2023) comprises 2,556 multiple-choice questions from Pew Research Center's Global Attitudes surveys (2,203 questions) and the World Values Survey (353 questions), designed to capture diverse opinions on global issues across countries.

(4) **CultureBank:** We hypothesize that opinion-aligned models should be able to extrapolate their knowledge to align with cultures

Model (zero-shot)	OpinionQA-XL		OpinionQA		GlobalOpinionQA		CultureBank		CultureNLI	
	Represent-ativeness (↑)	Steer-ability (↑)	Represent-ativeness (↑)	Steer-ability (↑)	Avg Sim (↑)	Skew (↓)	Redd-it (↑)	Tik -Tok (↑)	US (↑)	IN (↑)
Llama-2-7B-chat	83.61	79.09	86.18	79.18	83.6	2.2	85.93	92.08	39.2	39.5
Mistral-7B-Instruct	82.56	80.10	84.69	80.37	79.3	3.2	70.02	67.23	42.5	43.8
Vicuna-7B-v1.5	72.26	77.55	77.63	77.68	84.94	1.92	64.88	55.02	**55.72**	**56.15**
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
Behavior Finetuned LLama-2-7B-chat	**85.15**	**81.95**	**88.43**	**81.98**	**86.69**	**1.43**	**92.39**	**95.87**	47.14	43.92
LLama-2-13B-base	80.45	79.03	83.03	79.14	83.13	**1.45**	73.19	89.02	53.34	49.48
Llama-2-13B-chat	81.18	81.11	84.29	81.35	84.03	1.96	86.17	**92.34**	60.08	61.73
Vicuna-13B	79.06	78.73	83.44	78.85	86.99	1.91	85.93	92.08	52.07	40.23
Behavior Finetuned LLama-2-13B-chat	**85.76**	**83.54**	**89.44**	**83.53**	**87.31**	1.49	**86.28**	92.25	**62.26**	**66.44**
Mixtral-8x7B-Instruct	84.96	82.31	88.39	82.25	79.5	2.7	87.35	88.59	59.90	60.80
Mixtral-8X7B-SFT -CultureBank	84.40	79.66	78.69	79.67	81.80	2.80	86.19	92.08	61.50	61.30
Mixtral-8x7B-DPO -CultureBank	82.70	80.22	78.79	80.90	80.50	2.60	86.19	91.74	56.30	55.40
Llama-2-70B-chat	85.08	82.40	88.83	82.28	83.6	2.2	87.17	92.76	69.70	68.90
Behavior Finetuned LLama-2-70B-chat	**86.65**	**83.23**	**89.95**	**83.31**	**86.31**	**1.67**	**88.48**	**92.65**	**73.87**	**73.67**

Table 2: Comparison of all the models across Opinion and Culture tasks shows that our models trained on sparse in-the-wild behaviour signals, despite being zero-shot, outperforms models in opinion alignment and comes close to cultural alignment tasks. Furthermore, the model shows strong results beating even larger models trained on clean annotated data. We train variants of Llama-2 (Touvron et al., 2023).

as well. To test this, we use the CultureBank dataset (Shi et al., 2024), containing 23,000 cultural descriptors from TikTok and Reddit. Models are evaluated based on the grounded entailment scores using GPT-4 as a judge.

(5) **CultureNLI** (Huang and Yang, 2023) contains 2,700 culture-related natural language inference samples annotated by U.S. 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. The models are evaluated using by computing the entailment scores of model responses with human annotations.

(6) **Birds of Feather:** Homophily, or "love of same" has been widely explored in the social sciences (McPherson et al., 2001). This principle posits that people belonging to groups based on common features such as age or region share similar opinions. We propose the Birds of Feather benchmark to evaluate how closely models trained on the behavior of certain groups can also predict the opinions of other groups. To this end, we take the four US regions - Northeast, Midwest, West, and South, and train the model on combinations of three regions at a time, with the validation sets containing data from each of the four regions. We also train a model on data from all of US, and validate on the same validation set. Each training set contains 7.5M samples to isolate the impact of regional information from the number of training samples. Through this benchmark, we aim to show that homophilic groups can be captured using simple behavioral signals, and thus can be used to predict opinions of unseen groups with just opinions of their corresponding homophilic groups.

4 Experiments and Results

4.1 Training

We fine-tuned Llama-2-chat variants (7B/13B/70B) (Touvron et al., 2023) on AVA50M for one epoch using 32 A100 80 GB GPUs. To maintain conversational capabilities, we incorporated data from ShareGPT (Zheng et al., 2023). Additionally, we included behavioral data from the CBC dataset (Khandelwal et al., 2023), which empirically enhanced performance on our training tasks. We compare the performance of trained models against the 5-shot inference of similar-sized models and also much larger models like GPT-3.5 and GPT-4.

To verify task learning, we quantitatively assessed model performance across all training tasks. We evaluated the generated advertisements on TAG, TS, and TC tasks using NLP metrics of Perplexity, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020). For Perplexity evaluation, we compared average LLM perplexity on generating ground truth advertisements for specific demographics before and after training. Lower

post-training perplexity over better ads indicates improved alignment with the target demographic. We evaluated the TAP task using accuracy on a balanced test set.

The results for all models are provided in (1) TS tasks: Tables A12-A11, (2) TC tasks: Tables A8-A5, (3) TAG tasks: Tables A13-A15, (4) TAP tasks: Tables A16-A18. We see significantly higher scores on our metrics across the board. In general, the models show higher BLEU, ROUGE, and BERTScore scores for the generation tasks. Further, perplexity change shows that the trained models have learnt demographic preferences better than the untrained models on the time and advertiser stratified test sets.

4.2 Evaluation on Opinion Alignment in Zero-shot

We test the behavior-trained models in zero-shot settings on various cultural and opinion benchmarks. We compare the performance of these models against similar-sized (7B, 13B, 70B) models as well as bigger models such as Mixture of Experts Mixtral-8x7B and GPT-3.5 and GPT-4. The results are given in Table 2. We find that models trained on sparse in-the-wild behavior signals of the AVA50M dataset, despite being zero-shot, outperform models of similar and larger sizes in opinion and cultural alignment tasks, even when the baseline models are trained on clean annotated data.

We observe a few general trends: (1) there is not much difference in opinion alignment scores between models of various sizes, indicating that opinion alignment might not be a scale property. This insight could be useful in developing smaller opinion-aligned LLMs for use cases such as regional personal assistants. (2) Our models in 0-shot outperform baselines explicitly trained on cultural data. This empirically proves that behavioral signals are sufficient to learn about cultural opinions - carefully curated cultural datasets are not required to learn about cultural preferences. (3) Both representatives and steerability see an increase after training on behavior. Increasing representativeness denotes a higher alignment of default LLM opinion distribution with the general US population. Higher steerability denotes that the LLMs are able to better simulate the opinion distribution of a demographic group when prompted to do so.

Through our experiments, we also discovered some surprising patterns: (1) training on US-based ad data leads to an increase in similarity with Global opinions as well as cultural alignment with Indian culture over the base instruction fine-tuned model, as indicated by our experiments on GlobalOpinionQA and CultureNLI. This indicates latent correlations between the opinions of various demographic groups and further corroborates our hypothesis that knowledge about the opinions and behavior of one group can inform about the opinions and behaviors of correlated groups. (2) The Vicuna model, which was pretrained on data distilled from ChatGPT, achieves poor performance as compared to the instruction fine-tuned Chat models. We hypothesize that this occurs because ChatGPT is highly biased towards certain groups (Shi et al., 2024; Ouyang et al., 2022), which is reflected in Vicuna.

Our results on birds of feather setup are shown in Table A4. We see that despite having the same number of samples in all training sets, there are large differences in the representativeness of opinions of the US population. Specifically, we see that removing the regional information of West US causes a drop in overall representativeness. Further, the Northeast region has the lowest impact on the overall opinion alignment for each region. We believe this is due to the reason that Northeast shows lowest opinion alignment with any other region as evidenced in Fig. A6. As expected, the best alignment occurs when all regions are present in the training data.

5 Conclusion

This work demonstrates that by observing behaviors, we can effectively infer opinions, leveraging artificial agents trained on sparse behavioral signals to align with human culture and opinions. This approach offers a scalable, dynamic alternative to traditional culture-specific data annotation, circumventing the limitations of expert dependency, high costs, static datasets, and potential biases. We curate a new dataset for behavior alignment, and show through zero-shot evaluation using datasets like OpinionQA and GlobalOpinionQA that by training on this data, we achieve state-of-the-

art opinion alignment, showcasing the potential of large language models in modeling behavior from sparse signals and advancing the understanding of opinion dynamics.

6 Ethics And Societal Considerations

6.1 Meta Ads: A Novel Data Source for LLM Alignment

This study introduces a novel approach to aligning Large Language Models (LLMs) using behavioral data from Meta's advertising platforms. Traditionally, LLM alignment has relied on expert-annotated data or opinion surveys. Our method leverages advertisements created for Meta platforms (Facebook, Instagram, Messenger, and WhatsApp), which significantly enhances model performance compared to base models.

Meta's ad transparency initiative (Facebook, 2024) provides public access to advertisements related to social issues. Advertisers specify target audience parameters including age range, gender, location, preferences, and device type (Meta Platforms, Inc., a). Meta employs both automated technologies and manual verification to classify ads as political or issue-related (Meta Platforms, Inc., b,c).

6.2 Limitations and Future Directions

While our approach shows promising results, several limitations warrant further investigation:

1. Data source diversity: Our study primarily utilizes Meta's behavioral data for opinion alignment. Although Meta's platforms offer a broad demographic representation, incorporating data from other sources (e.g., Google Ads, Snapchat) could provide complementary insights. Platforms like Reddit might capture niche interests, while Google Ads could offer intent-based user data. However, API accessibility varies across platforms, presenting a challenge for comprehensive data integration.

2. Language and cultural expansion: The current work focuses on English-language ads, which effectively represent a wide range of issues and cultural contexts. This is evidenced by improvements in both representativeness and steerability when training on the AVA50M dataset. To enhance LLM utility across diverse cultures and languages, similar work must be conducted in non-English contexts. Our future plans include releasing AVA500M, an expanded dataset comprising 500 million ads from over 50 countries in 20 languages.

6.3 AVA50M: License and Terms of Use

We obtain all our data from the Facebook Ads Library, which does not contain any Personally Identifiable Information (PII), such as names, addresses, or contacts of individuals. We follow the Terms and Conditions put forward by Meta Inc. in the use of their data for research purposes.

To promote responsible use of our research and datasets, we will release an Acceptable Use Policy that explicitly prohibits the use of our dataset for applications where content generated could be harmful. This includes banning its use for abusive and fraudulent activities (e.g., spam generation and distribution), deceptive and misleading content (e.g., coordinated inauthentic behavior or presenting model-generated outputs as human-written), and sensitive use cases such as political campaigning and lobbying. We will actively monitor and enforce this policy to the best of our abilities. Additionally, we encourage other researchers and developers to adopt similar ethical guidelines when working with persuasive language models.

We shall release the datasets created using the process proposed in this paper in multiple phases, upon request. All requests will be recorded so as to prevent unauthorized and adverse use of the data, for example for negative marketing or brand defamation. We shall also build a community platform where other use cases of our data can be presented and discussed freely and publicly.

References

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

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

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

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

Aanisha Bhattacharyya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9822–9839, Singapore. Association for Computational Linguistics.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

United States Census Bureau. 2024. 2020 census results.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 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)*, pages 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.

Oliver Daniels-Koch and Rachel Freedman. 2022. The expertise problem: Learning from specialized feedback. In *NeurIPS ML Safety Workshop*.

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

Facebook. 2024. Facebook ads library.

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

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

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. *Preprint*, arXiv:2209.07065.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *Preprint*, arXiv:2203.07785.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

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

Ashmit Khandelwal, Aditya Agrawal, Aanisha Bhattacharyya, Yaman Kumar, Somesh Singh, Uttaran Bhattacharya, Ishita Dasgupta, Stefano Petrangeli, Rajiv Ratn Shah, Changyou Chen, et al. 2023. Large content and behavior models to understand, simulate, and optimize content and behavior. In *The Twelfth International Conference on Learning Representations*.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *Preprint*, arXiv:2404.01247.

Rakesh Kochhar. 2023. Survey methodology: The american trends panel survey methodology.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona,

Spain. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(Volume 27, 2001):415–444.

Meta Platforms, Inc. a. Facebook ads - about. https://www.facebook.com/ads/about/?entry_product=ad_library.

Meta Platforms, Inc. b. How are ads about social issues, elections and politics identified on facebook? https://www.facebook.com/business/help/214754279118974?id=288762101909005.

Meta Platforms, Inc. c. How are ads about social issues, elections and politics identified on facebook? https://www.facebook.com/help/180607332665293/.

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

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

OpenAI. 2023. Openai devday: Opening keynote.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023a. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

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

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Laura Silver, Christine Huang, Laura Clancy, and Andrew Prozorovsky. 2024. Methodology: About pew research center's spring 2024 global attitudes survey.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *Preprint*, arXiv:2209.12106.

Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. 2024. Measuring and improving persuasive abilities of generative models.

Samuel A Stouffer, Arthur A Lumsdaine, Marion Harper Lumsdaine, Robin M Williams Jr, M Brewster Smith, Irving L Janis, Shirley A Star, and Leonard S Cottrell Jr. 1949. The american soldier: Combat and its aftermath.(studies in social psychology in world war ii), vol. 2.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin

Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *Preprint*, arXiv:2310.12481.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Siyan Zhao, John Dang, and Aditya Grover. 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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

7 Appendix

Task Type	Constants	Variable	#Samples
TAG-G	Region, Budget, Age	Gender	120,468
TAG-A	Region, Budget, Gender	Age	120,531
TAG-R	Age, Budget, Gender	Region	381,247
TAG-B	Region, Age, Gender	Budget	2,667,937
TAG-I	Region, Age, Gender, Budget	Impressions	777,584
TAP-G	Region, Budget, Age	Gender	429,103
TAP-A	Region, Budget, Gender	Age	432,015
TAP-R	Age, Budget, Gender	Region	389,932
TAP-B	Region, Age, Gender	Budget	2,928,056
TAP-I	Region, Age, Gender, Budget	Impressions	787,245
TS-B	Region, Gender, Age	Budget	1,194,623
TS-E	Region, Gender, Age, Budget	Impressions	1,157,250
TS-A	Region, Budget, Gender	Age	9,307,920
TS-R	Age, Budget, Gender	Region	10,585,531
TS-G	Region, Budget, Age	Gender	7,001,126
TC-A	Region, Budget, Gender	Age	10,190,663
TC-G	Age, Budget, Gender	Region	2,310,127
TC-R	Region, Budget, Age	Gender	7,733,129
Total			58,514,486

Table A3: Distribution of instructions in the AVA50M dataset.

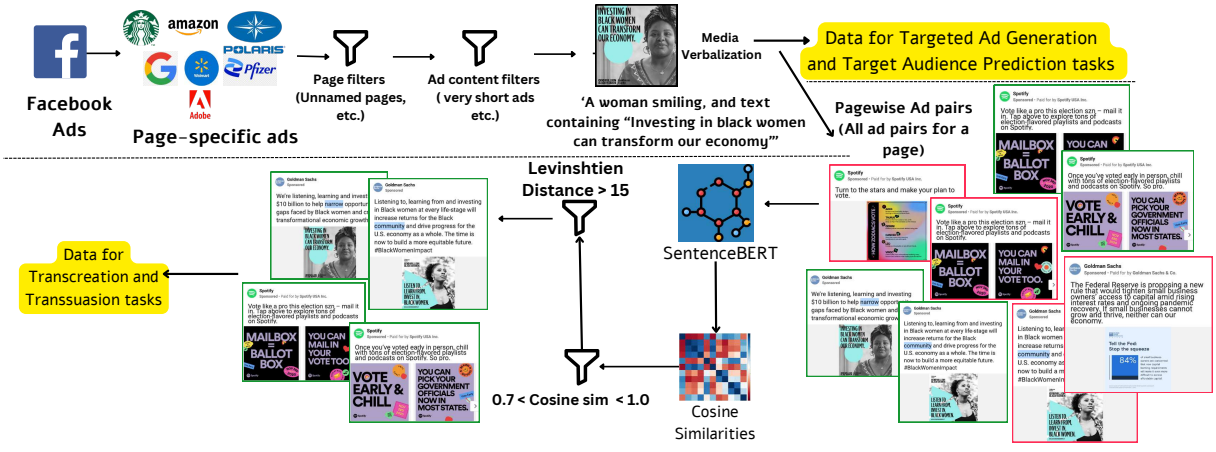

Figure A3: Overview of the data curation pipeline to create the AVA50M dataset.

Figure A4: Division of US territories into regions, used for our Birds Of Feather analysis. Image taken from https://www.burningcompass.com/countries/united-states/regions/us-regions-map.html.

Dropped Region (# training samples)	Representativeness of region on OpinionQA-XL				
	Northeast	Midwest	West	South	Average
West (7.5M)	0.8095	0.8069	0.8081	0.8038	0.8071
Midwest (7.5M)	0.8253	0.8228	0.8240	0.8204	0.8231
South (7.5M)	0.8370	0.8351	0.8360	0.8333	0.8354
Northeast (7.5M)	0.8476	0.8462	0.8467	0.8453	0.8465
None (7.5M)	0.8541	0.8520	0.8539	0.8528	0.8532

Table A4: Representativeness of Birds-of-feather analysis where we leave one region out while training the model. Each region is uniformly represented in the experiment, through this experiment we see the effect of regions on overall opinion alignment of the model. It shows that the region 'West' has maximum impact on overall opinion alignment of the model.

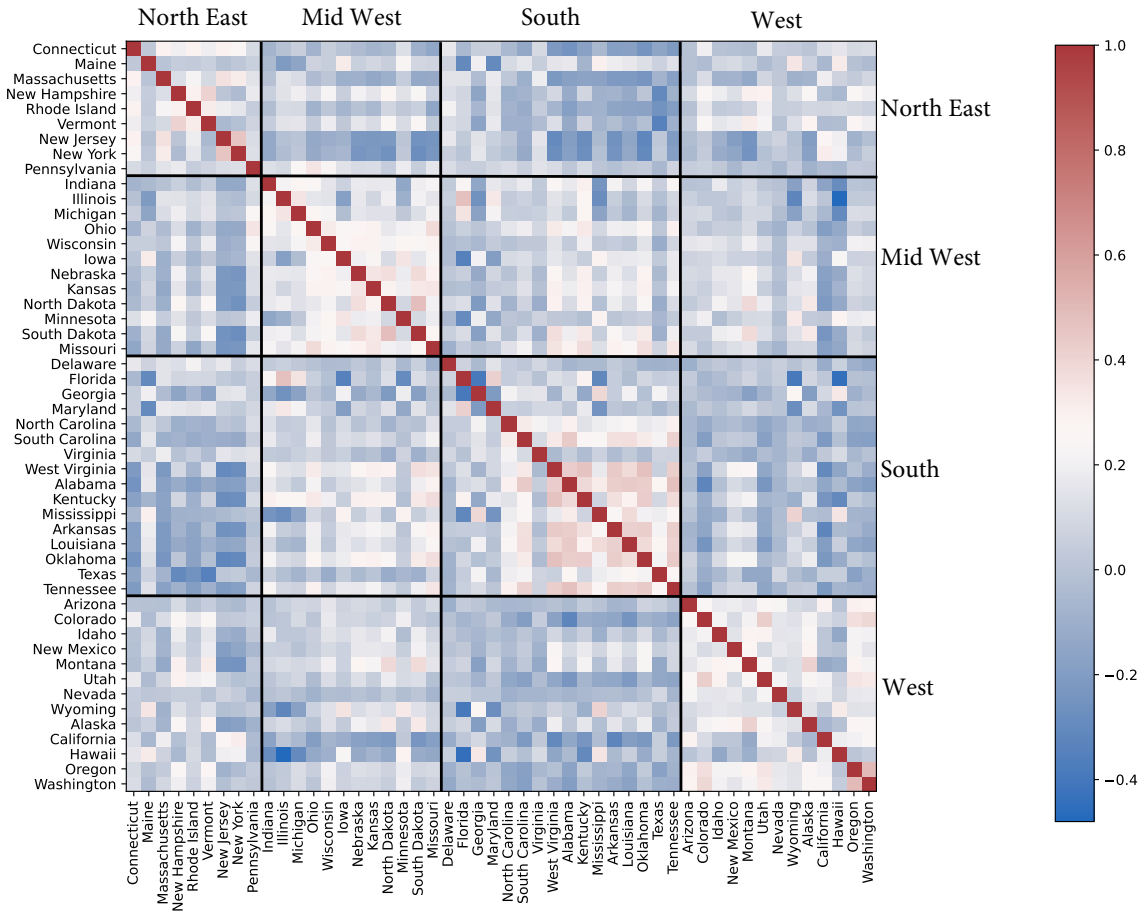

Figure A5: Correlation between behavior of people from different states in the US using the AVA50M dataset. The procedure to calculate correlation is explained in §3.2

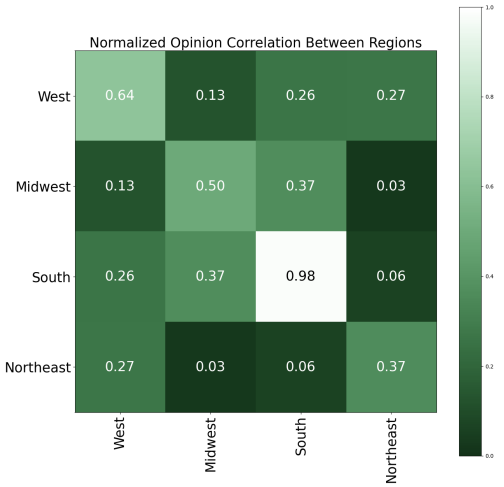

Figure A6: Correlation of opinions between regions as per PEW surveys. Such correlations indicate that it may be possible to model opinions of a given region based on opinion information from other regions.

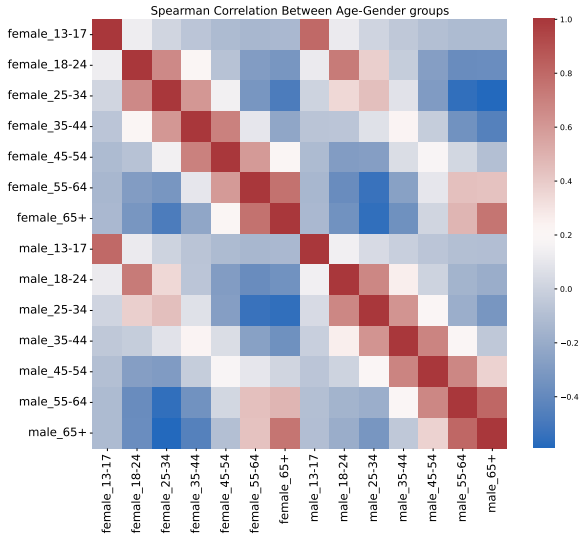

Figure A7: Correlation between behavior of people across different age-gender groups using the AVA50M dataset.

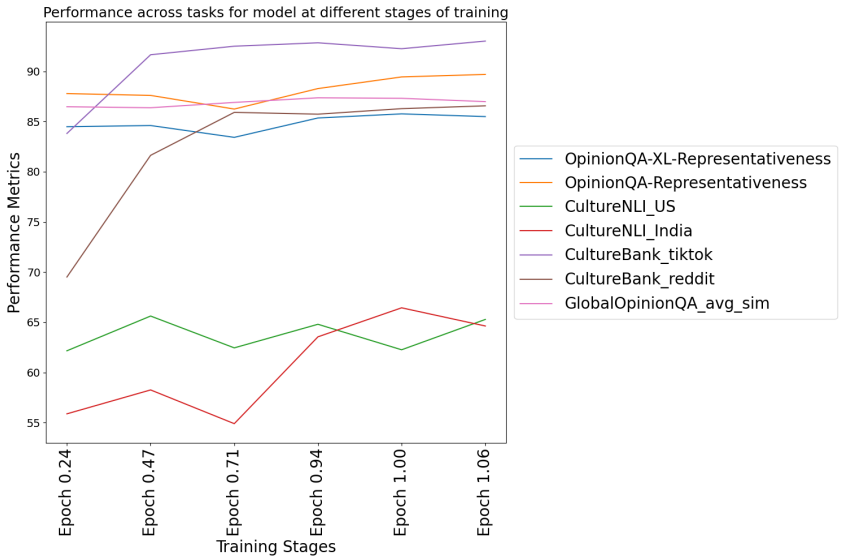

Figure A8: Performance on different benchmark datasets over training epochs. We see that the LLM's performance saturates near the 1-epoch mark. Beyond this point, training further gives negligible performance improvements.

A.1 Task Prompts

Here we provide the prompt templates for our training tasks. The values within {} are filled using values obtained from the Facebook Ad API.

Sample prompt for TAG

Given page name and keywords for an ad advertised in {regions}, generate the content for the advertisement suitable for the age group {age group}. The Advertisement for {brand} is titled {Title}. The advertisement was published on {platforms}. The campaign ran between {start_date and end_date}. The campaign was marketed to {USA/Global} audience. The amount of money that was spent on this campaign is between {lower_bound to upper_bound} USD. Keywords: {list of keywords}. Based on the preferences of the target {age group/region/gender}, generate the content. Try to use the given keywords while generating the content. Answer:

Sample prompt for TAP

Given an advertisement advertised in {states}, for which {age group, gender, states} does the advertisement perform best? The Advertisement for {Marketer Name} is titled {Title}. The advertisement was published on {platforms}. The campaign ran between {start_date and end_date}. The campaign was marketed to {USA/Global} audience. The amount of money that was spent on this campaign is between {lower_bound to upper_bound} USD. Content of the advertisement is {Ad body}. Based on the content, predict the {age_group/states/gender} for which the advertisement performs best. Answer:

Sample prompt for TS-X tasks

Given an advertisement created for {Marketer} targeted at {Given Audience}, the current advertisement does not perform well for the target audience. Create a new advertisement with a similar theme, but change it to better align with the preferences and cultural nuances of the audience {Target Audience}. Ensure the new advertisement is more appealing to the target {Target Audience}.
Current Advertisement: The advertisement titled {Title} with content : {Given Ad Content}. Answer:

Sample prompt for TC-X tasks

Given an advertisement for {Marketer} that performs well for {Given Audience}, create a similar advertisement tailored for {Target Audience}. Maintain the original theme and content while adapting it to resonate with the preferences and cultural nuances of {Target Audience}. Consider the unique characteristics and interests of the {Target Audience} to enhance the advertisement's effectiveness. Below is the advertisement for {Given Audience}: The advertisement titled {Title} with content : {Given Ad Content}. Answer:

A Task Samples

Sample prompt for TS-A tasks

Prompt: Given an advertisement created for DoorDash targeted at age group 25-34, the current advertisement does not perform well for the target audience. Create a new advertisement with a similar theme, but change it to better align with the preferences and cultural nuances of the audience in the age group 25-34. Ensure the new advertisement is more appealing to the target age groups: 25-34.
Current Advertisement: The advertisement titled Get Sushi Maki Delivered with content : All June, join us to fight hunger in the Miami community. When you order from Sushi Maki, we'll donate a meal* to Feeding South Florida. Terms: [URL]
Answer:
Ground Truth Response: The advertisement titled Get Saucy Asian Delivered with content : All June, join us to fight hunger in the San Francisco community. When you order from Saucy Asian, we'll donate a meal* to the SF-Marin Food Bank. Terms: [URL]

Sample prompt for TS-R tasks

Prompt: Given an advertisement created for Polaris targeted at Oregon, the current advertisement does not perform well for the target audience. Create a new advertisement with a similar theme, but change it to better align with the preferences and cultural nuances of the audience in Oregon. Ensure the new advertisement is more appealing to the target region: Oregon.
Current Advertisement: The advertisement titled Subscribe Now Don't Miss A Single Update! Join Now To Make A Difference. with content : Join now to make a difference in the lives of human trafficking victims and survivors. Join Polaris's digital community today and see how you can make a difference! We cannot end this one person, one survivor at a time. With your help, we can target the systems that make trafficking possible. Join Polaris's digital community today and see how you can make a difference!
Answer:
Ground Truth Response: The advertisement titled Subscribe Now with content : We cannot end this one person, one survivor at a time. With your help, we can target the systems that make trafficking possible. Join Polaris's digital community today and see how you can make a difference!

Sample prompt for TS-B tasks

Prompt: Given an advertisement created for Polaris targeted at Oregon, the current advertisement does not perform well for the target audience. Create a new advertisement with a similar theme, but change it to better align with the preferences and cultural nuances of the audience in Oregon. Ensure the new advertisement is more appealing to the target region: Oregon.
Current Advertisement: The advertisement titled Subscribe Now Don't Miss A Single Update! Join Now To Make A Difference. with content : Join now to make a difference in the lives of human trafficking victims and survivors. Join Polaris's digital community today and see how you can make a difference! We cannot end this one person, one survivor at a time. With your help, we can target the systems that make trafficking possible. Join Polaris's digital community today and see how you can make a difference!
Answer:
Ground Truth Response: The advertisement titled Subscribe Now with content : We cannot end this one person, one survivor at a time. With your help, we can target the systems that make trafficking possible. Join Polaris's digital community today and see how you can make a difference!

Sample prompt for TS-E tasks

Prompt: Given an advertisement for Nissan marketed in common states: ['all of US'] that received fewer views due to its lack of effectiveness, create a new advertisement with a similar theme. Ensure the new advertisement aligns with the brand and caters to audience preferences to achieve higher views on platforms like Facebook and Instagram.
Current Advertisement: The advertisement titled Double your donation to National Parks with content : Double your donation to National Parks. Help preserve and protect America's national parks. Now through February 8, Nissan TITAN will match any gift to the National Park Foundation 2-for-1, up to $200,000. #WeAreParks #CallingAllTITANS
Answer:
Ground Truth Response: The advertisement with content : Join Nissan TITAN in helping the National Park Foundation preserve and protect America's national parks. #WeAreParks #CallingAllTITANS

Sample prompt for TS-G tasks

Prompt: Given an advertisement for Goldman Sachs, create a similar advertisement better tailored for a female audience. Maintain the original theme and content while adapting it to resonate with the preferences and cultural nuances of the female audience. Consider the unique characteristics and interests of the female population to enhance the advertisement's effectiveness. Below is the original advertisement:
Advertisement: The advertisement titled The Daily Check-In with content : How does India's downgraded growth outlook compare to previous downturns and policy responses? Prachi Mishra, Goldman Sachs Research's Chief India Economist, explains. Now, create the advertisement better suited for a female audience:
Answer:
Ground Truth Response: The advertisement titled The Daily Check-In with content : Since [early March] we have dramatically downgraded our economic growth forecasts for India. Our focus was 5.8% before, now it's at 1.6%. This is a gigantic 420-basis-point downgrade. Goldman Sachs Research's Prachi Mishra on India's growth outlook.

Sample prompt for TC-R tasks

Prompt: Given an advertisement for The Stem Well that performs well in Oregon, create a similar advertisement tailored for Louisiana. Maintain the original theme and content while adapting it to resonate with the preferences and cultural nuances of the audience in Louisiana. Consider the unique characteristics and interests of the people in Louisiana to enhance the advertisement's effectiveness. Below is the advertisement for Hawaii:
Advertisement for Hawaii: The advertisement with content : Breaking News out of Princeton, NJ: Isabella Green from Cleveland, Ohio is a 2023 Princeton Prize in Race Relations winner for advancing equity in pediatric cancer prevention and care.
To learn more about The Princeton Prize in Race Relations, visit: [URL] learn more about The Stem Well, please follow our journey and click link in bio.
#genZleader #DEIJ #makingadifference #princetonprizeinracerelations #princeton #princetonuniversity #pediatriccancer #pediatriccancerawareness #childhoodcancer #childhoodcancerawareness #stemeducation #diversity #equity #inclusion Now, create the advertisement for Louisiana: Answer:
Ground Truth Response: The advertisement with content : The 2023 Princeton Prize in Race Relations (PPRR) Symposium was a vibe. Missing Princeton and the beautiful souls advancing racial equity in communities across the nation in meaningful ways that I had the pleasure of meeting and sharing space with. I (((love))) my PPRR family.
To learn more about The Princeton Prize in Race Relations visit, [URL] learn more about The Stem Well and our programs, click link in bio.
#PPRR #princeton #diversity #equity #inclusion #inclusionmatters #diversity #diversitymatters #diversityandinclusion #changemaker #changemakers #princetonprizeinracerelations

Sample prompt for TC-A tasks

Prompt: Given an advertisement for Nissan that performs well for an audience aged 35-44, create a similar advertisement tailored for an audience in the age group 55-64. Maintain the original theme and content while adapting it to resonate with the preferences and cultural nuances of the 55-64 audience. Consider the unique characteristics and interests of the 55-64 years old population to enhance the advertisement's effectiveness. Below is the advertisement for 35-44:
Advertisement for 35-44: The advertisement titled Double your donation to National Parks with content : Double your donation to National Parks. Help preserve and protect America's national parks. Now through February 8, Nissan TITAN will match any gift to the National Park Foundation 2-for-1, up to $200,000. #WeAreParks #CallingAllTITANS
Now, create the advertisement for 55-64:
Answer:
Ground Truth Response: The advertisement with content : Join Nissan TITAN in helping the National Park Foundation preserve and protect America's national parks. #WeAreParks #CallingAllTITANS

Sample prompt for TC-G tasks

Prompt: Given an advertisement for Johnson & Johnson that performs well for a female audience, create a similar advertisement tailored for a male audience. Maintain the original theme and content while adapting it to resonate with the preferences and cultural nuances of the male audience. Consider the unique characteristics and interests of the male population to enhance the advertisement's effectiveness. Below is the advertisement for female:
Advertisement for female: The advertisement titled Our 2022 Impact with content : We're committed to creating a healthier, more equitable world for all. Our 2022 Health for Humanity Report spotlights cutting-edge innovations and inspiring stories of progress as we reimagine the future of healthcare.
Now, create the advertisement for males:
Answer:
Ground Truth Response: The advertisement titled Our 2022 Impact with content : Our 2022 Health for Humanity Report is now live. Learn how we continue to innovate for the future of our communities, our employees and our planet.

Sample prompt for TAG-R tasks

Prompt: Given page name, keywords and top 3 states of USA where the advertisement performs best, generate the content for the advertisement. The Advertisement for Goldman Sachs is titled A More Equitable Future. The advertisement was published on facebook, instagram. The campaign ran between 22th March 2021 and 26th March 2021. The campaign was marketed to USA audience , where estimated audience size is between 100001 to 500000 are present. The amount of money that was spent on this campaign is between 600 to 699 USD. Top 3 states are Maryland, Louisiana, Alabama. Keywords: blackwomenimpact, women, investing, economy, black. Based on the preferences of the states of USA where the advertisement performs best, generate the content. Try to use the given keywords while generating the content. Answer:
Ground Truth Response: Content of the advertisement is Listening to, learning from and investing in Black women at every life-stage will increase returns for the Black community and drive progress for the U.S. economy as a whole. The time is now to build a more equitable future. #BlackWomenImpact.

Sample prompt for TAG-A tasks

Prompt: Given page name and keywords for an ad advertised in Illinois, Iowa, Florida, generate the content for the advertisement suitable for the age group 25-34. The Advertisement for Uber is titled Because care begins with getting there.. The advertisement was published on facebook, instagram. The campaign ran between 18th January 2022 and 13th February 2022. The campaign was marketed to USA audience , where estimated audience size is between 100001 and 500000. The amount of money that was spent on this campaign is between 7000 to 7999 USD. Keywords: uber, healthcare, appointments, aidsfoundationchicago, transportation. Based on the preferences of the target age group, generate the content. Try to use the given keywords while generating the content.
Answer:
Ground Truth Response: Content of the advertisement is Every year an estimated 3.6 million Americans miss their appointments due to a lack of reliable transportation. Uber Health offers a solution to address the transportations needs of thousands of patients and caregivers, helping Illinois healthcare organizations like @aidsfoundationchicago move health forward.

Sample prompt for TAG-G tasks

Prompt: Given page name and keywords for an ad advertised in Hawaii, Nebraska, Colorado, generate the content for the advertisement suitable for the female gender. The Advertisement for BuzzFeed is titled 8 Reasons You Should Be Supporting Small Businesses. The advertisement was published on facebook. The campaign ran between 01th May 2019 and 02th May 2019. The campaign was marketed to USA audience . The amount of money that was spent on this campaign is between 0 to 99 USD. Keywords: vistaprint, shopping, businesses, small, business. Based on the preferences of the target gender, generate the content. Try to use the given keywords while generating the content.
Answer:
Ground Truth Response: Content of the advertisement is Shopping small can make a big difference in your community. Support local establishments this Small Business Week May 5–11 and join Vistaprint in their efforts to support small businesses. Shopping small can make a big difference in your community. Support local establishments this Small Business Week May 5–11 and join Vistaprint in their efforts to support small businesses.

Sample prompt for TAP-A tasks

Prompt: Given an advertisement advertised in Mississippi, Alabama, Kentucky, for which age group does the advertisement performs best? The Advertisement for Goldman Sachs is titled Talks at GS. The advertisement was published on facebook. The campaign ran between 02th April 2020 and 10th April 2020. The campaign was marketed to USA audience . The amount of money that was spent on this campaign is between 0 to 99 USD Content of the advertisement is In a new Talks at GS podcast, William Gale of The Brookings Institution discusses how the recently passed federal relief package impacts the future of the US deficit. . Based on the content, predict the age group for which the advertisement performs best.
Answer:
Ground Truth Response: Best target age group for this ad is the audience aged 55-64.

Sample prompt for TAP-R tasks

Prompt: Given an advertisement, find top 3 states of USA where the advertisement performs best. The Advertisement for BuzzFeed is titled How Much Do You Know About Public Transportation? The advertisement was published on facebook. The campaign ran between 13th May 2019 and 16th May 2019. The campaign was marketed to USA audience. The amount of money that was spent on this campaign is between 1000 to 1499 USD. Content of the advertisement is Investing in public transportation leads to economic growth, and Infrastructure Week is the perfect time to support your local public transit! Investing in public transportation leads to economic growth, and Infrastructure Week is the perfect time to support your local public transit! Learn more at [URL]. Based on the content, predict the top 3 states of USA where the advertisement performs best.
Answer:
Ground Truth Response: Top 3 states where the ad will perform best are Massachusetts, New York, Vermont.

Sample prompt for TAP-G tasks

Prompt: Given an advertisement advertised in Maryland, Delaware, Pennsylvania, for which gender does the advertisement perform best? The Advertisement for Nissan is titled 2021 Nissan Sentra. The advertisement was published on facebook. The campaign ran between 09th January 2020 and 20th March 2020. The campaign was marketed to USA audience . The amount of money that was spent on this campaign is between 7000 to 7999 USD Content of the advertisement is Starting at $19,460 MSRP
Excl. taxes, title and license $139 per month lease, 36 months, $4,639 initial payment
As shown 2021 Sentra $259 per month lease, 36 months, $2,929 initial payment
For well-qualified lessees. Excl. taxes, title and license 0.0% APR financing for up to 36 months for well-qualified buyers Excl. taxes, title and license The official website of Nissan specializing in deals, offers, incentives, & rebates. Compare models to lease or buy from your local Nissan dealer. Based on the content, predict the gender for which the advertisement performs best.
Answer:
Ground Truth Response: Best target audience for this ad is Male.

A Results Table

A.1 Measuring Opinion Alignment of LLMs

Santurkar et al. (2023b) propose the following metrics to evaluate LLM alignment with human opinions using OpinionQA. The *alignment* between two distributions over answer choices D_1 and D_2 for a question q taken from a set of questions Q is given by:

$$\mathcal{A}(D_1, D_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} 1 - \frac{\mathcal{WD}(D_1(q), D_2(q))}{N-1}$$

OpinionQA is then evaluated using two metrics - *Representativeness*, computed as

$$\mathcal{R}_m^O(Q) = \mathcal{A}(D_m, D_O, Q). \tag{3}$$

which measures the Wasserstien distance between the distribution of opinions of ATP survey takers and the distribution of model outputs; and *Steerability*, computed as

$$\mathcal{S}_m^G(Q) = \frac{1}{|Q|} \sum_{q \in Q} \max_{c_G \in [\text{QA,B,P}]} \mathcal{A}(D_m(q; c_G), D_G(q))$$

which measures the Wasserstien distance between the opinion distributions of a specific demographic group and the model's outputs when prompted to represent that group. QA, B, P refer to different prompting strategies, as described in Santurkar et al. (2023b).

The prompt we used to evaluate Opinion-QA and Opinion-QA-XL:

Representativeness Prompt:
Question: {question}
Possible answers:
A. {option-A}
B. {option-B}
C. {option-C}
D. {option-D}
Answer only with A, B, C, or D and nothing else.
Example:
Question: Please think about what things will be like in 2050, about 30 years from now. Thinking about the future of the United States, would you say you are.
Possible answers:
A. Very optimistic
B. Somewhat optimistic
C. Somewhat pessimistic
D. Very pessimistic
Answer only with A, B, C, or D and nothing else.

Steerability Prompt:
Answer the following question considering yourself as a person {factor} Question: {question}
Possible answers:
A. {option-A}
B. {option-B}
C. {option-C}
D. {option-D}
Answer only with A, B, C, or D and nothing else.
Example:
Answer the following question considering yourself as a person whose level of education is College graduate/some postgrad. Question: How important is it to you, if at all, that professional athletes who you support share your political views?
Possible answers:
A. Do not follow professional sports
B. Not at all important
C. Not too important
D. Somewhat important
E. Very important
Answer only with A,B,C,D or E and nothing else.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	21.23	5.75	3.86	3.05	17.79	0.81
Llama-7B-chat	27.69	7.90	4.20	2.44	23.38	0.84
Mistral-7B	26.31	9.71	6.77	5.17	21.19	0.80
Llama-13b	24.42	7.66	4.99	3.76	20.21	0.82
Llama-13b-chat	27.87	8.47	4.67	2.79	23.31	0.85
Vicuna-1.5-13B	35.08	15.84	10.36	6.91	28.92	0.86
Llama-70B	24.36	7.38	4.69	3.49	20.24	0.82
Llama-70B-chat	27.52	8.33	4.64	2.84	23.49	0.85
GPT4-Turbo	27.94	9.76	4.37	0.84	26.58	0.87
Behavior Finetuned LLama-2-7B-chat	39.26	18.70	12.23	7.23	29.65	0.87
Behavior Finetuned LLama-2-13B-chat	43.14	21.47	14.36	9.04	31.41	0.87
Behavior Finetuned LLama-2-70B-chat	41.76	21.71	15.17	10.25	31.96	0.87

Table A5: Comparison of Behavior Finetuned Models against baseline LLMs on the TC-R task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore	Δ PPL
Llama-7B	17.36	3.10	1.91	1.36	13.68	0.80	-2.36
Llama-7B-chat	23.94	4.37	1.87	0.71	20.09	0.84	-8.95
Mistral-7B	14.82	2.71	1.37	0.87	0.12	0.59	-
Llama-13b	18.25	3.00	1.56	0.98	14.43	0.80	-5.10
Llama-13b-chat	23.61	4.23	1.76	0.73	19.08	0.84	-10.59
Vicuna-1.5-13B	26.68	6.21	2.94	1.44	20.87	0.84	-5.21
Llama-70B	21.95	4.53	2.28	1.29	17.44	0.82	-5.95
Llama-70B-chat	23.12	4.13	1.69	0.68	19.55	0.84	-12.69
GPT4-turbo	25.47	9.14	4.43	1.57	27.42	0.86	-
Behavior Finetuned LLama-2-7B-chat	43.64	24.48	18.61	14.41	38.18	0.88	7.37
Behavior Finetuned LLama-2-13B-chat	44.63	26.72	21.12	17.05	38.88	0.89	4.57
Behavior Finetuned LLama-2-70B-chat	47.22	28.54	22.68	18.54	41.41	0.89	4.52

Table A6: Comparison of Behavior Finetuned Models against baseline LLMs on the TS-E task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore	Δ PPL
Llama-7B	16.29	2.36	1.27	0.89	12.84	0.78	2.47
Llama-7B-chat	22.03	3.29	1.19	0.51	18.70	0.82	0.77
Mistral-7B	19.98	4.64	2.63	1.70	0.15	0.79	-
Llama-13b	17.24	2.72	1.45	0.98	13.26	0.78	0.79
Llama-13b-chat	22.69	3.88	1.48	0.63	18.70	0.81	1.22
Vicuna-1.5-13B	25.14	5.68	2.45	1.08	19.71	0.82	6.87
Llama-70B	18.64	3.02	1.41	0.85	13.91	0.79	0.54
Llama-70B-chat	21.42	3.26	1.16	0.51	18.33	0.82	0.59
GPT4-Turbo	24.16	8.03	3.57	0.68	24.05	0.86	-
Behavior Finetuned LLama-2-7B-chat	35.92	16.12	10.04	5.61	28.34	0.84	5.03
Behavior Finetuned LLama-2-13B-chat	43.31	24.17	17.79	13.17	32.97	0.86	4.94
Behavior Finetuned LLama-2-70B-chat	37.67	18.07	11.93	7.60	30.01	0.85	2.44

Table A7: Comparison of Behavior Finetuned Models against baseline LLMs on the TS-B task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	20.74	5.37	3.53	2.77	15.98	0.80
Llama-7B-chat	25.84	6.60	3.34	1.90	21.25	0.85
Mistral-7B	25.18	8.19	5.32	3.94	19.14	0.80
Llama-13b	23.17	6.45	3.98	2.96	17.99	0.82
Llama-13b-chat	25.85	6.36	3.07	1.71	20.89	0.84
Vicuna-1.5-13B	31.52	11.53	6.66	4.06	25.15	0.85
Llama-70B	23.95	6.26	3.42	2.33	18.55	0.83
Llama-70B-chat	25.08	5.88	2.76	1.49	20.81	0.85
GPT4-Turbo	34.72	13.20	6.76	3.02	29.62	0.87
Behavior Finetuned LLama-2-7B-chat	35.93	15.70	9.63	5.58	27.82	0.86
Behavior Finetuned LLama-2-13B-chat	38.09	17.21	10.78	6.46	28.53	0.87
Behavior Finetuned LLama-2-70B-chat	38.09	17.21	10.78	6.46	29.00	0.87

Table A8: Comparison of Behavior Finetuned Models against baseline LLMs on the TC-A task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	22.02	6.43	4.48	3.61	17.18	0.81
Llama-7B-chat	26.22	6.73	3.52	2.12	21.89	0.84
Mistral-7B	25.90	8.92	6.03	4.64	17.99	0.76
Llama-13b	25.40	8.41	5.59	4.27	19.95	0.82
Llama-13b-chat	26.42	6.73	3.37	1.88	21.37	0.84
Vicuna-1.5-13B	31.94	12.46	7.49	4.73	25.51	0.86
Llama-70B	24.21	7.09	4.38	3.29	18.96	0.82
Llama-70B-chat	26.26	6.93	3.54	2.12	21.51	0.85
GPT4-Turbo	35.38	14.53	7.90	3.82	31.24	0.88
Behavior Finetuned LLama-2-7B-chat	39.00	18.79	12.29	7.91	30.34	0.87
Behavior Finetuned LLama-2-13B-chat	41.40	21.11	14.57	10.16	32.04	0.87
Behavior Finetuned LLama-2-70B-chat	40.85	20.80	14.24	9.83	32.09	0.87

Table A9: Comparison of Behavior Finetuned Models against baseline LLMs on the TC-G task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore	Δ PPL
Llama-7B	18.41	3.26	1.91	1.44	13.55	0.80	1.10
Llama-7B-chat	21.77	2.93	1.04	0.53	17.08	0.83	0.48
Mistral-7B	20.54	3.92	1.97	1.33	14.47	0.77	-1.09
Llama-13b	18.75	3.25	1.76	1.30	13.19	0.79	0.92
Llama-13b-chat	22.95	3.76	1.49	0.76	17.92	0.83	0.72
Vicuna-1.5-13B	39.29	18.92	12.80	8.88	30.66	0.86	-0.36
Llama-70B	19.77	3.04	1.40	0.89	14.10	0.80	0.80
Llama-70B-chat	21.89	3.29	1.19	0.58	17.18	0.83	0.06
GPT4-Turbo	27.49	9.58	4.54	1.64	26.62	0.86	-
Behavior Finetuned LLama-2-7B-chat	39.60	19.39	13.05	8.74	30.76	0.87	1.40
Behavior Finetuned LLama-2-13B-chat	41.83	21.13	14.63	10.16	31.47	0.87	1.94
Behavior Finetuned LLama-2-70B-chat	41.45	20.97	14.40	10.03	31.87	0.87	1.15

Table A10: Comparison of Behavior Finetuned Models against baseline LLMs on the TS-A task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore	Δ PPL
Llama-7B	20.44	4.39	2.61	1.93	15.46	0.81	-3.12
Llama-7B-chat	26.05	6.17	2.95	1.75	21.12	0.84	-3.03
Mistral-7B	27.74	10.54	7.27	5.69	21.48	0.80	-1.33
Llama-13b	24.77	7.63	4.94	3.74	19.17	0.83	0.16
Llama-13b-chat	26.56	6.73	3.28	1.82	20.70	0.84	-0.46
Vicuna-1.5-13B	31.56	11.30	6.42	3.89	24.46	0.86	-1.13
Llama-70B	25.06	6.93	3.87	2.58	19.41	0.83	0.14
Llama-70B-chat	26.66	6.86	3.36	1.85	20.93	0.85	-1.05
GPT4-Turbo	30.86	12.76	6.80	3.33	30.02	0.87	-
Behavior Finetuned LLama-2-7B-chat	39.14	19.38	13.09	8.83	29.93	0.87	3.55
Behavior Finetuned LLama-2-13B-chat	43.07	23.18	16.71	12.33	33.20	0.88	5.67
Behavior Finetuned LLama-2-70B-chat	42.38	22.05	15.39	10.94	32.21	0.88	2.48

Table A11: Comparison of Behavior Finetuned Models against baseline LLMs on the TS-G task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore	Δ PPL
Llama-7B	18.40	3.26	1.90	1.44	14.11	0.80	1.04
Llama-7B-chat	21.77	2.92	1.04	0.53	18.20	0.83	-0.50
Mistral-7B	19.94	3.92	2.33	1.78	14.93	0.78	-1.43
Llama-13b	18.75	3.25	1.76	1.30	14.49	0.79	0.69
Llama-13b-chat	22.95	3.76	1.49	0.76	18.40	0.83	-0.47
Vicuna-1.5-13B	25.16	5.28	2.20	1.18	19.40	0.82	2.63
Llama-70B	20.71	3.90	2.18	1.54	15.19	0.82	0.28
Llama-70B-chat	21.26	2.72	0.95	0.43	18.30	0.83	-0.86
GPT4-Turbo	16.43	4.24	1.73	0.43	20.36	84.77	-
Behavior Finetuned LLama-2-7B-chat	40.06	20.14	14.07	9.73	33.03	0.88	2.35
Behavior Finetuned LLama-2-13B-chat	41.83	21.13	14.63	10.16	34.08	0.88	3.66
Behavior Finetuned LLama-2-70B-chat	43.72	24.16	17.90	13.49	36.60	0.88	1.79

Table A12: Comparison of Behavior Finetuned Models against baseline LLMs on the TS-R task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	15.17	1.49	0.29	0.12	10.60	0.80
Llama-7B-chat	17.92	2.45	0.73	0.33	14.60	0.83
Mistral-7B	16.74	1.74	0.40	0.18	9.20	0.70
Llama-13b	17.84	1.84	0.40	0.19	11.00	0.78
Llama-13b-chat	20.25	2.72	0.70	0.27	13.96	0.82
Vicuna-1.5-13B	19.06	2.79	0.86	0.35	14.85	0.82
Llama-70B	20.17	2.59	0.71	0.32	13.21	0.80
Llama-70B-chat	21.44	3.07	0.88	0.36	14.78	0.83
GPT4-Turbo	10.54	2.13	0.92	0.46	15.47	0.81
Behavior Finetuned LLama-2-7B-chat	28.59	10.31	6.80	4.64	22.25	0.84
Behavior Finetuned LLama-2-13B-chat	28.34	10.17	6.69	4.55	22.26	0.84
Behavior Finetuned LLama-2-70B-chat	28.52	10.30	6.75	4.59	22.33	0.84

Table A13: Comparison of Behavior Finetuned Models against baseline LLMs on the TAG-A task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	17.39	1.88	0.45	0.20	10.75	0.79
Llama-7B-chat	20.80	3.23	1.04	0.51	14.94	0.83
Mistral-7B	18.68	2.08	0.48	0.20	8.98	0.72
Llama-13b	17.64	1.85	0.39	0.18	10.20	0.75
Llama-13b-chat	20.98	3.04	0.86	0.35	14.61	0.82
Vicuna-1.5-13B	22.40	3.42	1.11	0.50	14.81	0.80
Llama-70B	19.49	2.36	0.59	0.26	12.87	0.81
Llama-70B-chat	21.84	3.38	1.12	0.53	15.24	0.83
GPT4-Turbo	26.79	8.29	4.93	3.04	22.89	0.84
Behavior Finetuned LLama-2-7B-chat	28.39	10.25	6.69	4.45	21.67	0.84
Behavior Finetuned LLama-2-13B-chat	27.53	9.75	6.40	4.27	21.90	0.84
Behavior Finetuned LLama-2-70B-chat	28.33	10.16	6.57	4.34	22.12	0.84

Table A14: Comparison of Behavior Finetuned Models against baseline LLMs on the TAG-R task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BERTScore
Llama-7B	16.82	1.61	0.36	0.20	10.45	0.79
Llama-7B-chat	20.24	2.72	0.82	0.37	14.73	0.83
Mistral-7B	18.96	2.03	0.47	0.21	9.05	0.70
Llama-13b	17.63	1.76	0.37	0.19	10.76	0.77
Llama-13b-chat	20.23	2.70	0.74	0.29	14.15	0.82
Vicuna-1.5-13B	21.82	3.13	0.94	0.39	14.26	0.81
Llama-70B	17.62	1.84	0.41	0.20	11.14	0.80
Llama-70B-chat	20.34	2.77	0.79	0.32	14.24	0.82
GPT4-Turbo	19.64	5.37	2.92	1.71	21.43	0.84
Behavior Finetuned LLama-2-7B-chat	28.67	10.22	6.62	4.40	21.93	0.84
Behavior Finetuned LLama-2-13B-chat	28.12	10.09	6.69	4.54	22.36	0.84
Behavior Finetuned LLama-2-70B-chat	28.52	10.19	6.51	4.31	22.32	0.84

Table A15: Comparison of Behavior Finetuned Models against baseline LLMs on the TAG-G task.

Model	Accuracy
Random Accuracy	50.00
Llama-7b	51.10
Llama-7b-chat	47.94
Mistral-7b	50.08
Llama-13b	49.43
Llama-13b-chat	44.98
Vicuna-1.5-13B	38.73
Llama-70b	55.40
Llama-70b-chat	56.73
GPT4-Turbo	44.26
Behavior Finetuned LLama-2-7B-chat	58.89
Behavior Finetuned LLama-2-13B-chat	63.40
Behavior Finetuned LLama-2-70B-chat	61.96

Table A16: Comparison of Behavior Finetuned Models against baseline LLMs on the TAP-G task.

Model	Accuracy
Random Accuracy	2.00
Llama-7b	4.41
Llama-7b-chat	11.01
Mistral-7b	4.33
Llama-13b	3.53
Llama-13b-chat	8.81
Vicuna-1.5-13B	7.44
Llama-70b	3.77
Llama-70b-chat	10.54
GPT4-Turbo	5.61
Behavior Finetuned LLama-2-7B-chat	5.89
Behavior Finetuned LLama-2-13B-chat	8.20
Behavior Finetuned LLama-2-70B-chat	5.89

Table A17: Comparison of Behavior Finetuned Models against baseline LLMs on the TAP-R task.

Model	Accuracy
Random Accuracy	14.29
Llama-7b	18.27
Llama-7b-chat	22.46
Mistral-7b	17.92
Llama-13b	17.90
Llama-13b-chat	20.85
Vicuna-1.5-13B	20.69
Llama-70b	17.85
Llama-70b-chat	19.07
GPT4-Turbo	25.54
Behavior Finetuned LLama-2-7B-chat	23.75
Behavior Finetuned LLama-2-13B-chat	24.05
Behavior Finetuned LLama-2-70B-chat	22.98

Table A18: Comparison of Behavior Finetuned Models against baseline LLMs on the TAP-A task.