The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have sparked wide interest in validating and comprehending the human-like cognitive-behavioral traits LLMs may have. These cognitive-behavioral traits include typically *Attitudes, Opinions, Values* (AOV). However, measuring AOV embedded within LLMs remains opaque, and different evaluation methods may yield different results. This has led to a lack of clarity on how different studies are related to each other and how they can be interpreted. This paper aims to bridge this gap by providing an overview of recent works on the evaluation of AOV in LLMs. Moreover, we survey related approaches in different stages of the evaluation pipeline in these works. By doing so, we address the potential and challenges with respect to understanding the model, human-AI alignment, and downstream application in social sciences. Finally, we provide practical insights into evaluation methods, model enhancement, and interdisciplinary collaboration, thereby contributing to the evolving landscape of evaluating AOV in LLMs.

1 Introduction

Recent years have witnessed a remarkable improvement in the development and deployment of Large Language Models (LLMs), holding the promise of boosting various domains, from computer sciences to social sciences and beyond. Amid the excitement surrounding their capabilities lies an important question: How well do these LLMs capture and convey human cognitive-behavioral traits?

By drawing upon theories from social sciences (such as Katz, 1960; Rokeach, 1968; Ajzen, 1988; Bergman, 1998), we consider human cognitive-behavioral traits, in our case primarily **Attitudes, Opinions, Values (AOV)**, the fundamental component of human cognition, shaping our perceptions, decisions, and interactions. By studying whether and how the LLM outputs reflect AOV and how these AOV compare to human AOV, we can better understand their potential to act as autonomous agents that could mirror human AOV. The AOV in LLMs also impact users in downstream applications, such as writing assistants (Jakesch et al., 2023), and affect decision-making processes and perceptions (Eigner and Händler, 2024).

In recent studies, survey questionnaires that were originally used to estimate public opinions in the social sciences are now being popularly utilized to evaluate the opinions of LLMs and subsequently to study the alignment with human opinions (Santurkar et al., 2023; Hwang et al., 2023; Kim and Lee, 2024). At the same time, the wide range of evaluation methods used to assess LLM responses has led to inconsistent outcomes, complicating reliable assessment of the models (Dominguez-Olmedo et al., 2023; Wang et al., 2024a,b, *inter alia*). However, this variability in evaluation methods has largely been overlooked—posing risks of missing subtleties in LLM performance, yielding incomplete or biased assessments. This oversight raises significant questions about the model's true capabilities and its alignment with human opinions.

Motivated by the rising interest in studying the human-like traits of LLMs, in this paper, we present the first survey on the evaluation of AOV in LLMs. Before moving into the details, we first position our survey in the context of other relevant surveys and then show the framework of our survey.

Related Survey Papers. While there are no survey papers specifically on AOV in LLMs, some existing works have covered related questions. Simmons and Hare (2023) review works and provide a framework for using LLMs as comparative models for subpopulations to measure public opinions. Jansen et al. (2023) offer insights on employing LLMs in public opinion survey research, concluding that LLMs can enhance survey research. Vida et al. (2023) address the research gap in the ethical aspects of NLP surveying the literature on moral

NLP, calling for a more rigorous discussion on the moral concept for NLP research. Hershcovich et al. (2022) provide a survey on NLP in cross-cultural contexts from a linguistic diversity angle, suggesting the need to preserve cultural values in models. There are also recent surveys on understanding "culture" in NLP (Liu et al., 2024) and measuring "cultures" in LLMs (Adilazuarda et al., 2024), both highlighting the future of culturally aware and adapted NLP techniques. These works mainly explored topics like studying the cultural and moral aspects in NLP or improving public opinion research with LLMs. However, there has been a lack of dedicated studies focusing on AOV and especially on evaluating the AOV within LLMs.

Our Survey Paper. Since LLMs are pretrained on vast amounts of human data, it is reasonable to hypothesize that LLMs can reflect the AOV embedded in the data. But, for that to scale, we will need definitions of the terms AOV (**WHAT is it?** §2), then to summarize what has been explored on the AOV in LLMs so far (**WHAT so far?** §3), and know the pipeline used so far in research on how LLMs are queried for the AOV embedded within (**HOW?** §4). We then discuss the research directions (**WHERE?** §5) by highlighting the potential and challenges identified from existing works and the evaluation pipeline. In the end, we provide a call for action on what to do to make these approaches possible and reliable in the future (**WHAT to do?** §6).

2 Definitions

Next, we provide definitions for the three main concepts used in this paper: *attitude*, *opinion*, and *value* (**WHAT is it?**). According to Katz (1960), an *attitude* is a durable orientation toward some object, while an *opinion* is more of a visible expression of an attitude. For this paper, we examine the two concepts simultaneously following Bergman (1998), who considers the *attitude* and *opinion* as synonymous:

> **Citation 1.** "**Attitudes (and opinions)** are always attitudes about something. This implies three necessary elements: **first**, there is the object of thought, which is both constructed and evaluated. **Second**, there are acts of construction and evaluation. **Third**, there is the agent, who is doing the constructing and evaluating. We can therefore suggest that, at its most general, an attitude is the cognitive construction and affective evaluation of an attitude object by an agent." (Bergman, 1998)

We apply the above definition to the study of LLM attitudes and opinions. These three elements are formed as follows: **first**, there is the topic under consideration as the object of thought; **second**, there is the internal mechanisms and processes within the LLM that perform the construction and evaluation of this topic; and **third**, there is the LLM itself as the agent.

On *value*, Bergman (1998)'s definition reads:

> **Citation 2.** "A **value** may be understood as the cognitive and affective evaluation of an array of objects by a group of agents." (Bergman, 1998)

This definition suggests that values extend beyond individual attitudes and opinions, denoting grouped thoughts and evaluations of an array of objects.

LLMs were trained on a great amount of textual data from billions of humans. This means that when prompted, LLMs might sometimes generate responses that *incorporate these varied perspectives rather than a single viewpoint* (Jiang et al., 2023; Cheng et al., 2023a; Jiang et al., 2024; Shu et al., 2024; Choi and Li, 2024). LLMs could be understood "as a superposition of perspectives" (Kovač et al., 2023) and have both dimensions. Thus, in our paper, we suggest to consider the terms *attitudes*, *opinions*, and *values* together and to study them as a cohesive set. We propose a two-dimensional view for it: *attitudes* and *opinions* encompass the attitudes and opinions prevalent in societal contexts, often captured through timely surveys and polls; *values* look deeper into the ethical and cultural beliefs that guide individual and collective behavior, usually more stable over time.

3 An Overview of Related Works for AOV in LLMs

In this section, we present related recent works on the evaluation of AOV in LLMs (**WHAT so far?**). We categorize the works into two main groups: *attitudes/opinions* and *values*, reflecting the two dimensions of AOV we proposed. In addition, we include works with various topics that could also shine light on AOV in LLMs. A summary of the surveyed papers, along with details on the paper selection process and an analysis of the model distribution can be found in the Appendix (§A.1, §A.2).

3.1 Attitudes/Opinions

US-Centric Public Opinion Polls. The majority of recent work on evaluating opinions in LLMs is based on US-centric public opinion surveys. Argyle et al. (2023), Bisbee et al. (2023) and Sun et al. (2024) query the model with a prompt that encompasses the socio-demographics of real human participants using the American National Election Studies (ANES) surveys. Santurkar et al. (2023) use the American Trends Panel (ATP) survey from the Pew Research Center and create the dataset OpinionQA. The OpinionQA data set has also been used by Hwang et al. (2023) and Wang et al. (2024b). Similarly, Tjuatja et al. (2024) also use ATP data to study whether LLMs exhibit human-like response biases. There are various additional US-based surveys used to study LLMs' AOV (Dominguez-Olmedo et al., 2023; Kim and Lee, 2024; Sanders et al., 2023; Lee et al., 2024a). Most of the papers found misalignment between LLM and human opinions and several observed left-leaning political bias in their comparisons.

Non-US-Centric Public Opinion Polls. Although most work relies on the US context, a few studies focus on non-US countries or cross-national comparisons. von der Heyde et al. (2023) use data from German Longitudinal Election Study (GLES, 2019) and notice strong bias also in their use case (German election prediction). Kalinin (2023) uses the Survey of Russian Elites from 1993–2020 (Zimmerman et al., 2023) and leverages LLMs to generate opinions like Russian elite individuals. Durmus et al. (2023) introduce the dataset GlobalOpinionQA based on questions and answers from cross-national surveys on diverse opinions on global issues across different countries and discover cultural and social biases of LLMs' outputs.

Non-Public-Opinion Polls. Apart from public opinion surveys, other contents are also used for studying the LLMs' sensitivity to public opinions. Jiang et al. (2022a) present a CommunityLM by fine-tuning GPT2 models (Radford et al., 2019) on partisan Twitter data finding that the fine-tuned models align well with ANES survey data. Wu et al. (2023) and Rosenbusch et al. (2023) focus on LLMs' attitudes towards US politicians. Chalkidis and Brandl (2024) fine-tune the Llama Chat model (Touvron et al., 2023) on debates in the European Parliament and discover that the adapted party-specific models can align towards respective positions. There is a web tool, OpinionGPT (Haller et al., 2023), which shows that biases of the input data influence the answers a model produces. Rozado (2023), Rozado (2024), Feng et al. (2023) and Röttger et al. (2024) use political orientation tests or political compass tests to evaluate opinions in LLMs. The varied political worldview in LLMs was further found in recent works (Ceron et al., 2024; Bang et al., 2024).

3.2 Values

Value Orientation of LLMs. For research on values, social science studies use surveys such as the World Values Survey (WVS) (Haerpfer et al., 2022) and the Hofstede Cultural Survey (Hofstede, 2005). These surveys have also been applied in recent studies to evaluate the values in LLMs. Benkler et al. (2023) find that LLMs struggle to accurately capture the moral perspectives of non-Western demographics when responding to WVS questions. Arora et al. (2023) employ the WVS and the Hofstede Cultural Survey into cloze-style questions and study the cultural expression of multilingual LMs by inducing perspectives of speakers of different languages. Cao et al. (2023) probe ChatGPT with the Hofstede Cultural Survey and Johnson et al. (2022) experiment on WVS, both showing that the model aligns mostly with American culture. In addition, Tanmay et al. (2023) measure the moral reasoning ability of LLMs using the Defining Issues Test (Rest, 1979).

Moral Foundations Theory[1] (Graham et al., 2018) has been applied in a few studies to assess the models' moral values. Simmons (2023) investigates moral biases in LLMs using Moral Foundations Theory and demonstrates that these models exhibit moral biases when prompted with a certain political identity. Haemmerl et al. (2023) probe multilingual LLMs based on their moral foundations. There are inconsistent findings regarding the evaluation of values in LLMs based on moral foundations. While Talat et al. (2022) claim that the models exhibit fluctuating ethical values, Fraser et al. (2022) find that the models' ethical values align consistently with their training data.

Curated Datasets and Frameworks. There are a few curated evaluation datasets for values in LLMs, such as ETHICS (Hendrycks et al., 2023), MoralChoice (Scherrer et al., 2024), MoralExceptQA (Jin et al., 2022), ValuePrism (Sorensen et al., 2024). A few frameworks have been established to assess the ethical reasoning capability of LLMs, such as SocialChemistry101 (Forbes et al., 2020), Delphi (Jiang et al., 2022a), the Framework for 'in-

[1]The Moral Foundations Theory (Graham et al., 2011) identifies five foundations (Care, Fairness, Loyalty, Authority, Purity) to explain shared moral themes across populations (Abdulhai et al., 2023). The Moral Foundations Questionnaire (Graham et al., 2011) scores these five foundations.

context' Ethical Policies (Rao et al., 2023), Moral Graph Elicitation (Klingefjord et al., 2024), as well as moral dilemmas and value statements (Rao et al., 2023; Agarwal et al., 2024). Ren et al. (2024) provide an evaluation pipeline ValueBench to probe value orientations encompassing 453 value dimensions. These resources and frameworks collectively enhance our ability to evaluate and understand the values embedded in LLMs.

3.3 Other Related Topics

In addition to the two main categories, several studies investigate related topics that indirectly also reveal the AOV reflected in LLMs. These include: i) trustfulness, which is closely related to AOV as it reflects the model's alignment to human values on truth and honesty (Lin et al., 2022; Joshi et al., 2024), ii) theory-of-mind, which explores the ability of LLMs to understand and predict human thoughts and emotions (Sap et al., 2022; Li et al., 2023b; Kosinski, 2024), iii) persona and personality, of which findings highlight the models' ability to reflect human-like attitudes and values through their generated personas (Miotto et al., 2022; Kovač et al., 2023; Caron and Srivastava, 2023; Cheng et al., 2023a,b; Jiang et al., 2024; Shu et al., 2024), iv) sentiment (Deshpande et al., 2023; Beck et al., 2024b; Hu and Collier, 2024), and v) mixed topics spanning politics, philosophy and personality (e.g. Perez et al., 2023).

4 How LLMs Are Queried for AOV

After defining the core concepts and discussing related works, we now provide details of the pipeline on how LLMs were queried for AOV so far **(HOW?)** to motivate our later discussion on gaps. Based on the surveyed works, we categorize the evaluation process in a taxonomy into four main stages: i) input, ii) model, iii) output, and iv) evaluation, as illustrated in Figure 1.

4.1 Input

In this section, we show methods for formatting input data before feeding them into the model. Several examples of the task design for the input can be found in the Appendix §A.3.

Persona-Based Input. In this approach, personas, i.e. the demographic profiles of a human sample, are included into the input prompt to simulate the opinions of specific sub-populations, allowing for the comparisons of LLM outputs with human responses. This method has been widely explored, for example in Santurkar et al. (2023); Hwang et al. (2023); Durmus et al. (2023); Kim and Lee (2024).

Input Perturbations. To test the robustness and consistency of the model's outputs, perturbations have been applied to the input to test the human-like response biases of the model. The most common way is to perturbate the order of the choices in close-ended questions (Lu et al., 2022; Kovač et al., 2023; Dominguez-Olmedo et al., 2023; Tjuatja et al., 2024; Wang et al., 2024b; Shu et al., 2024). Tjuatja et al. (2024) propose response bias modifications (e.g. order swapping) and non-bias perturbations (e.g. letter swapping and typos), which are also employed in Wang et al. (2024a). In addition, modifying prompt wording is another perturbation approach. Cao et al. (2023) change questions from the second to the third person, while Kovač et al. (2023) and Ceron et al. (2024) prepend a system message in the second person to the question. Hwang et al. (2023) add a Chain-of-Thought (CoT, Wei et al., 2023) style prompt wording to the original questions.

4.2 Model

In this section, we explore various inference methods used with the models after preparing the input.

Zero-Shot Inference. The zero-shot inference is the most common way to probe the LLMs by asking the model with input prompts without examples and is employed in most of the works, for example in Argyle et al. (2023); Santurkar et al. (2023); Hwang et al. (2023); Durmus et al. (2023); Sanders et al. (2023); von der Heyde et al. (2023).

Few-Shot Inference. The few-shot inference includes one or a few examples in the prompt to familiarize the model with the expected response format. For example, Santurkar et al. (2023) experimented with one-shot examples in the prompt for multiple choice survey response generation. Hendrycks et al. (2023), Sap et al. (2022), Perez et al. (2023) and Joshi et al. (2024) include a few examples in the prompt as additional ablation experimentations.

Fine-Tuning and Inference. Some studies utilize the fine-tuning approach to align LLMs with specific viewpoints by training them on data containing those opinions (e.g. partisan Twitter data, parliamentary debates), and during the inference period then evaluate these fine-tuned models on test sets (e.g. questionnaires for human public opinion polls) (Jiang et al., 2022b,a; Joshi et al., 2024; Chalkidis and Brandl, 2024; Kim and Lee, 2024). These works showed that the fine-tuned models can

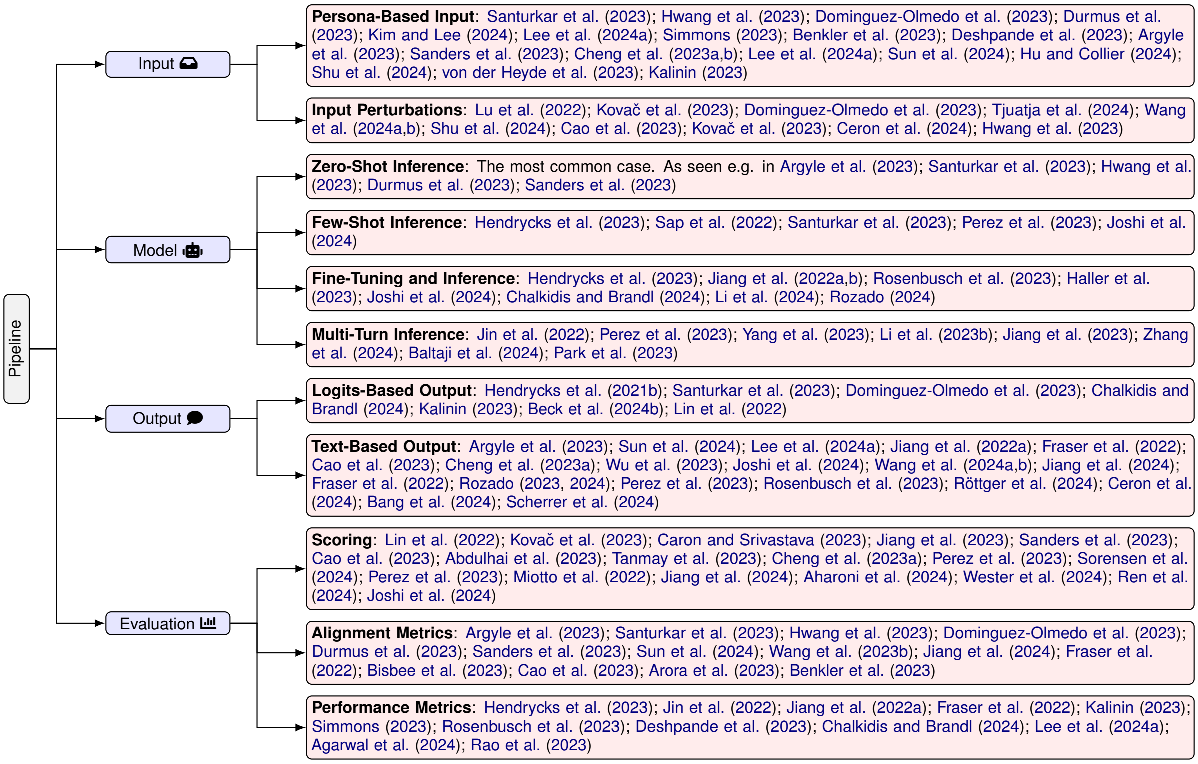

Figure 1: A taxonomy of evaluation pipeline across input → model → output → evaluation.

represent represent the opinions behind the training data.

Multi-Turn Inference. In multi-turn inference, the process is usually chain-wise or conducted by multiple agents. Perez et al. (2023) instruct LLMs to write yes/no questions with multiple stages of LM-based generation and filtering. Several works (Jin et al., 2022; Jiang et al., 2023; Yang et al., 2023) incorporate CoT processes to complete questionnaires in a multi-turn dialogue manner, while Baltaji et al. (2024) use multi-agent LLM systems for inter-cultural collaboration and debate, analyzing opinion diversity before and after discussions, based on previous research on social behaviors in LLM agents (Li et al., 2023b; Zhang et al., 2024).

4.3 Output

After defining the inputs and models and feeding the input into the model, we can now address the output side. There are two main ways for output extraction: logits-based output and text-based output.

Logits-Based Output. The first token logits of LLM outputs have been commonly used in multiple-choice question settings to transform the open-ended nature of LLM outputs into expected options, as in Hendrycks et al. (2021b); Santurkar et al. (2023); Dominguez-Olmedo et al. (2023); Chalkidis and Brandl (2024); Kalinin (2023); Beck et al. (2024b); Lin et al. (2022). This method involves calculating the log probabilities for answer options (e.g. 'A', 'B', 'C'). The option with the highest log probability is then selected as answer.

Text-Based Output. The text-based way spans different approaches that look at the textual output from the model. Argyle et al. (2023) extract texts from models' output using string matching `RegEx`. Lee et al. (2024a) employ string matching with manual modifications on incorrect matching instances. Jiang et al. (2022a) only examine the first line in the response and remove the remaining tokens. Joshi et al. (2024) train a linear probing classifier to predict the truthfulness of an answer. Wang et al. (2024a,b) annotate a subset of the outputs and fine-tune a model on the annotated subset to train a classifier for output classification. Rozado (2024), Bang et al. (2024) and Röttger et al. (2024) directly take the LLM outputs and use other LLMs to classify the stance of the target LLM outputs.

4.4 Evaluation

After extracting the LLM output, different evaluation metric approaches are applied to validate the model behavior.

Scoring. There are various approaches to scoring model-generated responses for evaluation. Some methods rely on direct rating from humans on the model-generated responses (Lin et al., 2022; Caron and Srivastava, 2023; Perez et al., 2023; Jiang et al., 2024; Sorensen et al., 2024; Aharoni et al., 2024; Wester et al., 2024; Joshi et al., 2024), while some also use model-based scoring (Kovač et al., 2023; Jiang et al., 2023; Caron and Srivastava, 2023; Sanders et al., 2023; Jiang et al., 2024; Joshi et al., 2024), or predefined scoring frameworks (Cao et al., 2023; Abdulhai et al., 2023; Tanmay et al., 2023; Cheng et al., 2023a). Usually, a rating scale is given to score the acceptability of the response. In addition, some outputs can be directly evaluated because they come in score form (e.g. when prompted with questions and options with scaled scores), such as in Miotto et al. (2022).

Alignment Metrics. By drawing upon well-known measures of inter-annotator agreement and similarity measures, alignment metrics have been employed to measure the alignment of human and LLM responses. These measures include Cohen's Kappa (Argyle et al., 2023; Hwang et al., 2023), 1-Wasserstein distance (WD) (Santurkar et al., 2023; Hwang et al., 2023; Sanders et al., 2023), Kullback–Leibler (KL) divergence (Dominguez-Olmedo et al., 2023; Sun et al., 2024), the Euclidean distance between the model's responses and the standard scores of humans (Wang et al., 2023b), Jensen-Shannon Distance for model and country alignment (Durmus et al., 2023), as well as correlation and statistical analysis (Kalinin, 2023; Sun et al., 2024; Jiang et al., 2024). Moreover, metrics have been applied to measure the alignment between variables, such as regression models for measuring the correlations between single features of different personas (Bisbee et al., 2023) and between different nations (Benkler et al., 2023).

Performance Metrics. Performance metrics (e.g. Acc., F1., Loss) have been applied to measure the quality of LLM outputs against target datasets, as in Hendrycks et al. (2023); Jin et al. (2022); Kalinin (2023); Chalkidis and Brandl (2024); Lee et al. (2024a); Agarwal et al. (2024). In Simmons (2023), performance is assessed by comparing response content with "moral foundations dictionaries". Meanwhile, Rosenbusch et al. (2023) establish a baseline accuracy by having human experts match politicians with their ideologies, against which LLM predictions are evaluated.

5 Opportunities and Challenges in Evaluating AOV in LLMs

Drawing from findings summarized in §3 and §4 from recent advances, we now focus on the methodological and practical perspectives regarding opportunities and challenges of evaluating AOV in LLMs (**WHERE?**). The next section addresses several key issues starting with the need to understand the models themselves, followed by the necessity for human-AI alignment, and finally, the implications for downstream applications in social sciences.

5.1 Understanding the Model

The essential discussion on the impact of evaluating AOV in LLMs should start with the models themselves – the agents creating output. Our understanding of these models is limited (much like our understanding of ourselves as humans) (Hassija et al., 2024). As studying how people respond to questions and express opinions helps us understand human behavior, examining how models do the same can enhance our knowledge of these models.

Evaluating AOV Helps Understand Model Behavior. By effectively evaluating AOV in LLMs, we could potentially better explain their behavior in those subjective contexts, which could reveal why models produce certain opinions and values, helping us to better interpret their outputs. Apart from the textual output, tracking model internal behavior is also of interest, for example, to examine whether there exist skill neurons (Wang et al., 2022; Voita et al., 2023). Investigating the internal working mechanisms of models enhances their interpretability, helping to make their operations more transparent and understandable. Currently, there is a lack of work linking AOV evaluations to model interpretability. Addressing this gap would significantly contribute to the understanding and reliability of LLM outputs, especially in subjective contexts.

Evaluating AOV Helps Understand Model Biases. Since LLMs are trained on large datasets that contain human-generated content, they inevitably learn and reproduce the biases present in this data (Anwar et al., 2024). For example, models often reflect Western cultural perspectives because much of the training data comes from Western sources (Johnson et al., 2022; Cao et al., 2023; Adilazuarda et al., 2024). This can lead to skewed outputs not representing diverse global perspectives. Also, in most LLMs English-centric biases exist, i.e., models show significant value bias when we move to languages other than English (Agarwal et al., 2024). To address these issues, techniques were proposed, such as bias detection (Cheng et al., 2024), adversarial training (Casper et al., 2024), and diversification of training data (Chalkidis and Brandl, 2024).

Evaluation Methods Are Not Robust. One challenge in evaluating the output of LLMs is that the methods used can themselves be brittle. For example, in multiple-choice survey question settings, several studies rely on the first token logits (probabilities) of model output to map the options with the highest logits (such as Santurkar et al., 2023; Dominguez-Olmedo et al., 2023). However, Wang et al. (2024a,b) observe that the first token logits do not always match the textual outputs and sometimes the mismatch rate can be over 50% in Llama2-7B (Touvron et al., 2023) and Gemma-7B (Team et al., 2024). A few works have also highlighted models being sensitive to option ordering (Binz and Schulz, 2023; Pezeshkpour and Hruschka, 2023; Zheng et al., 2024; Shu et al., 2024; Wei et al., 2024), making evaluation unstable. Therefore, any evaluation for AOV in LLMs should be accompanied by extensive robustness tests (Röttger et al., 2024). Wang et al. (2024a,b) propose to look at the text by training classifiers on the annotated LLM outputs, which typically requires a lot of human efforts and may not be generalizable. Developing context-aware evaluation metrics to capture human-like nuances in LLM outputs is an ongoing research focus for model interpretability.

5.2 Human-AI Alignment

After understanding the model, aligning LLMs with human AOV and ensuring that they perform safely and effectively is the next crucial phase.

Improvement in the Diversity of Alignment. Alignment methods, such as Reinforcement Learning from Human Feedback (RLHF, Ouyang et al., 2022), focus on the problem of *aligning LLMs to human values*, which requires transferring the human values into *alignment target* for training and evaluating the models (Klingefjord et al., 2024). However, current evaluations have often been coarse, highlighting the need for more fine-grained benchmarks to assess alignment effectively (Lee et al., 2024b). One fundamental challenge RLHF faces is the problem of misspecification (Casper et al., 2023). The diversity of human values cannot be easily represented by a single reward function. Current alignment evaluation benchmarks and reward model training rely on individual preference but lack consideration of the nature of diversity in human opinion. A more fine-grained evaluation of AOV with respect to *social choice* (Conitzer et al., 2024) or *social awareness* (Yang et al., 2024) will help us better understand the alignment process and design a better socially-aware alignment algorithms (Conitzer et al., 2024).

Personalization Raises Risks of Anthropomorphism. Anthropomorphizing AI models — attributing human characteristics to them — can lead to unrealistic expectations and misunderstandings about their capabilities and limitations (Weidinger et al., 2022; Kirk et al., 2024a). While aligning models with human values is important, it is equally crucial to maintain a clear distinction between human and AI capabilities. Most recent works add persona-based prompts (e.g. Santurkar et al., 2023), which include demographics of real survey participants and might lead to privacy risks in encouraging the share of intimate information (Burkett, 2020; Zehnder et al., 2021; Kirk et al., 2024a). Besides, over-personalization might raise the risk of microtargeting and malicious persuasion. Properly handling the nature and limitations of LLMs could reduce the risks associated with anthropomorphism.

5.3 Implications from and for Social Science Applications

Considering the potential and challenges from the model perspective, we will now explore the feasibility of deploying AOV in LLMs in downstream social science applications. LLMs, with their ability to process vast amounts of text data, could provide valuable insights into human values and behaviors. Again, caution must be exercised to address inherent biases and alignment issues that may arise.

Problems of Alignment with Human Survey Participants. Currently, we have no means of aligning LLMs to accurately represent the diversity of human opinions necessary for reliable public opinion polling and similar tasks. The existing literature highlights numerous challenges, particularly in replicating non-US values (Benkler et al., 2023; Arora et al., 2023; Simmons, 2023; Rao et al., 2023). While some argue that LLM surveys might provide insights into hard-to-reach populations, the risk remains significant that these groups are difficult to model accurately by LLMs (von der Heyde et al., 2023; Namikoshi et al., 2024).

Human AOV Help Evaluate AOV in LLMs.

While there is a great gap between Human AOV and those in LLMs, human-centered applications can enhance our understanding and validation of AOV in LLMs. In survey methodology, responding to a survey question involves several cognitive steps, mainly including comprehension, retrieval, judgment, and reporting (Tourangeau et al., 2000; Groves et al., 2004; Tourangeau, 2018). Figure 2 illustrates a basic model of the human survey response process. Despite fundamental differences, the behavioral study of machines can benefit from that of animals (Rahwan et al., 2019), as well as of humans (Greasley and Owen, 2016). By integrating these human-centered cognitive processes into the examination of how LLMs respond to survey questions, we are able to gain valuable insights into the models and then modify the models to better align with human processes. Still, while concepts from human AOV are certainly helpful in studying LLMs, we should also keep in mind at all times that they are after all not humans and should caution against the anthropomorphism we discussed in the previous section.

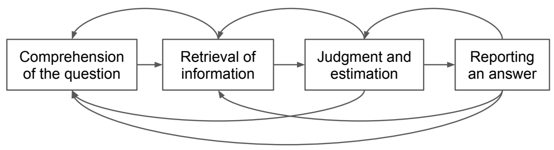

Figure 2: A simple model of the survey response process (Groves et al., 2004)

LLMs Can Generate Test Data for Survey Applications. In survey applications, LLMs can significantly improve testing pipelines by generating plausible test data (Simmons and Hare, 2023; Hämäläinen et al., 2023; Wang et al., 2023a). By simulating a variety of respondent behaviors and answers, LLMs allow the identification of weaknesses and biases in survey instruments. However, in this case, too, it is important to note the potential mismatch between model-generated data and actual human responses (Bisbee et al., 2023; von der Heyde et al., 2023; Hämäläinen et al., 2023).

6 Towards a Future of Evaluating AOV in LLMs

As we have discussed, evaluating AOV in LLMs offers opportunities alongside notable challenges (§5). To harness these opportunities while addressing the challenges, we show below key areas where focused action may lead to substantial improvements (**WHAT to do?**).

Develop A More Fine-Grained and Human-Centered Evaluation Pipeline. The current methods for evaluating AOV in LLMs within the pipeline sometimes lack the necessary rigor for robust and reliable evaluations, especially due to the unstable results from the current evaluation methods. We call for the development of a more robust and fine-grained evaluation pipeline that can better capture the nuances of human-like expressions in LLM outputs. Besides, there is a great gap in the evaluation benchmarks. The current existing benchmarks for evaluating the opinions in LLMs such as OpinionQA (Santurkar et al., 2023) and MMLU (Hendrycks et al., 2021b) are static. Interactive benchmarks such as AlpacaEval (Li et al., 2023c) and MT-Bench (Zheng et al., 2023) focus more on general preferences. Therefore, more human-centered and fine-grained benchmarks from cognitive and social sciences should also be explored and extended to validate the "human" factors within the models in real-world scenarios.

Incorporate Diverse Human Opinions and Preferences to Better Align the Model. Incorporating diverse human opinions and preferences from public sources (Huang et al., 2024) into model values helps to better align the model. For example, preference tuning like RLHF has the potential to align LLMs more closely with human values, but it requires a nuanced understanding of human preferences, at best interactively (Shen et al., 2024). Collecting fine-grained data that accurately reflects diverse human opinions and values is crucial to align the model. We need to ensure that the preference data used in aligning the model are representative and ethically sound. Best practices from survey methodology should be considered to ensure the data collection is both diverse and comprehensive (O'Hare et al., 2015; Kern et al., 2023; Beck et al., 2024a; Eckman et al., 2024; Kirk et al., 2024b).

Foster Interdisciplinary Collaboration. Understanding and improving the evaluation of AOV in LLMs requires insights from multiple disciplines. Interdisciplinary collaboration can provide a deeper understanding of both human cognitive processes and model behaviors. It is crucial to involve experts from different fields, e.g. survey methodology, psychology and sociology, to guide how we design and analyze the evaluations (Dwivedi et al., 2023; Eckman et al., 2024). Research driven by interdisciplinary hypotheses can enhance our understanding of how well LLMs capture human-like AOV from a broader perspective.

7 Limitations

In this work, we present a survey and commentary on the progress and challenges of evaluating AOV in LLMs. There are several key limitations that should be acknowledged:

Inclusivity of Related Work. This survey predominantly focuses on works with subjective context related to opinions and values. As a result, other relevant areas such as emotion detection, e.g. (Wang et al., 2023b; Li et al., 2023a), which might implicitly contain value expressions, have not been included here. Future research could explore a broader range of related works beyond AOV.

Perspective on the Evaluation Pipeline. The discussion on the evaluation pipeline in this work may be limited in scope, mainly focusing on the four evaluation stages, however decisions in each step have potential for profound impact on results. While we provide an overview of the evaluation pipeline with diverse approaches in each evaluation stage, there may be additional aspects or single features of the evaluation pipeline that were not thoroughly examined or highlighted, such as detailed pre-processing and data augmentation methods, intermediate representation analysis and error analysis methods. Future studies could delve deeper into these aspects to contribute in providing an even more comprehensive understanding of the evaluation process of AOV in LLMs.

Exploration of Use Cases. This work primarily focuses on the evaluation aspect of LLMs and does not extensively explore their potential use cases in social science and society. While evaluating AOV in LLMs is undoubtedly important, it is equally crucial to consider how these models can be applied in various domains to address real-world challenges. Future research could explore the broader implications of LLMs in social science research, policymaking, education, and other societal applications to provide a more holistic perspective on their utility and impact.

8 Ethical Considerations

Within the surveyed papers and approaches, there might exist contents that could potentially raise ethical considerations, due to the nature of the subjectivity in these topics. We report these in two key aspects:

Ethical Considerations Regarding the Data Used. In future studies involving the collection of new survey and questionnaire data, researchers must exercise caution and be mindful of ethical concerns, especially with regard to sensitive topics. It is crucial to design questions in a way that avoids causing direct or indirect harm to participants. Ensuring ethical sensitivity in the data collection process is vital to maintaining the integrity and safety of the research (Hammer, 2017). Alignment studies also often require comparing LLM responses with those from real human participants. Researchers should ensure that these human participants provide informed consent and that their privacy is protected.

Ethical Considerations in LLM Applications. As discussed in §5.2, overpersonalizing and anthropomorphizing AI models might raise privacy risks and ethical concerns. The use of LLMs in social science research brings up important ethical questions regarding privacy, consent, and the potential for harm. Most LLMs are instruction-tuned with safety mechanisms to avoid sensitive topics and conflicts (Grigis and De Angeli, 2024). Despite this, researchers must exercise extreme caution due to the potential mismatch between LLM outputs and actual human opinions, which can also lead to harmful consequences due to misleading conclusions. To prevent these issues, it is crucial to continuously monitor and address cultural and value biases in LLM outputs, ensuring that AI usage does not perpetuate stereotypes or lead to unfair treatment of any group. Additionally, opinionated LLMs can influence users' views and decision-making, necessitating careful monitoring and engineering (Jakesch et al., 2023; Sharma et al., 2024). Researchers must remain vigilant and transparent about the limitations and ethical complexities of employing LLMs in their studies.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *Preprint*, arXiv:2310.15337.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6330–6340, Torino, Italia. ELRA and ICCL.

Eyal Aharoni, Sharlene Fernandes, Daniel J. Brady, Caelan Alexander, Michael Criner, Kara Queen, Javier Rando, Eddy Nahmias, and Victor Crespo. 2024. Attributions toward artificial agents in a modified moral turing test. *Scientific Reports*, 14(1):8458.

Icek Ajzen. 1988. *Attitudes, Personality, and Behavior*, 1st edition. Open University Press, Milton Keynes.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *Preprint*, arXiv:2404.09932.

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.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Razan Baltaji, Babak Hemmatian, and Lav R Varshney. 2024. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent llm collaboration. *arXiv preprint arXiv:2405.03862*.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *Preprint*, arXiv:2403.18932.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024a. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024b. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Noam Benkler, Scott Friedman, Sonja Schmer-Galunder, Drisana Mosaphir, Vasanth Sarathy, Pavan Kantharaju, Matthew D. McLure, and Robert P. Goldman. 2022. Cultural value resonance in folktales: A transformer-based analysis with the world value corpus. In *Social, Cultural, and Behavioral Modeling*, pages 209–218, Cham. Springer International Publishing.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *Preprint*, arXiv:2312.10075.

Manfred Max Bergman. 1998. A theoretical note on the differences between attitudes, opinions, and values. *Swiss Political Science Review*, 4(2):81–93.

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

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. Synthetic replacements for human survey data? the perils of large language models.

Christopher B. Burkett. 2020. "I Call Alexa to the Stand": The Privacy Implications of Anthropomorphizing Virtual Assistants Accompanying Smart-Home Technology. *Vanderbilt Journal of Entertainment and Technology Law*, 20(4):1181–1210.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. 2024. Defending against unforeseen failure modes with latent adversarial training. *Preprint*, arXiv:2403.05030.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Preprint*, arXiv:2402.17649.

Ilias Chalkidis and Stephanie Brandl. 2024. Llama meets eu: Investigating the european political spectrum through the lens of llms. *Preprint*, arXiv:2403.13592.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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

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

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, and Tianyu Shi. 2024. Rlrf:reinforcement learning from reflection through debates as feedback for bias mitigation in llms. *Preprint*, arXiv:2404.10160.

Hyeong Kyu Choi and Yixuan Li. 2024. Beyond helpfulness and harmlessness: Eliciting diverse behaviors from large language models with persona in-context learning. *Preprint*, arXiv:2405.02501.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social choice for ai alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, Hanaa Albanna, Mousa Ahmad Albashrawi, Adil S. Al-Busaidi, Janarthanan Balakrishnan, Yves Barlette, Sriparna Basu, Indranil Bose, Laurence Brooks, Dimitrios Buhalis, Lemuria Carter, Soumyadeb Chowdhury, Tom Crick, Scott W. Cunningham, Gareth H. Davies, Robert M. Davison, Rahul Dé, Denis Dennehy, Yanqing Duan, Rameshwar Dubey, Rohita Dwivedi, John S. Edwards, Carlos Flavián, Robin Gauld, Varun Grover, Mei Chih Hu, Marijn Janssen, Paul Jones, Iris Junglas, Sangeeta Khorana, Sascha Kraus, Kai R. Larsen, Paul Latreille, Sven Laumer, F. Tegwen Malik, Abbas Mardani, Marcello Mariani, Sunil Mithas, Emmanuel Mogaji, Jeretta Horn Nord, Siobhan O'Connor, Fevzi Okumus, Margherita Pagani, Neeraj Pandey, Savvas Papagiannidis, Ilias O. Pappas, Nishith Pathak, Jan Pries-Heje, Ramakrishnan Raman, Nripendra P. Rana, Sven Volker Rehm, Samuel Ribeiro-Navarrete, Alexander Richter, Frantz Rowe, Suprateek Sarker, Bernd Carsten Stahl, Manoj Kumar Tiwari, Wil van der Aalst, Viswanath Venkatesh, Giampaolo Viglia, Michael Wade, Paul Walton, Jochen Wirtz, and Ryan Wright. 2023. Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from survey methodology can improve training data. In *Forty-first International Conference on Machine Learning*.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *Preprint*, arXiv:2402.17385.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A. Association for Computational Linguistics.

GLES. 2019. Post-election cross section (gles 2017). GESIS Data Archive, Cologne. ZA6801 Data file Version 4.0.1, https://doi.org/10.4232/1.13235.

Jesse Graham, Jonathan Haidt, Matt Motyl, Peter Meindl, Carol Iskiwitch, and Marlon Mooijman. 2018. Moral foundations theory: On the advantages of moral pluralism over moral monism. In *Atlas of moral psychology*, pages 211–222. The Guilford Press, New York, NY, US. PsycInfo Database Record (c) 2023 APA, all rights reserved.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385. R01 MH068447/MH/NIMH NIH HHS/United States.

Andrew Greasley and Chris Owen. 2016. Behavior in models: A framework for representing human behavior.

Paolo Grigis and Antonella De Angeli. 2024. Playwriting with large language models: Perceived features, interaction strategies and outcomes. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, AVI '24, New York, NY, USA. Association for Computing Machinery.

Robert M. Groves, Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Wiley–Interscience, Hoboken, NJ.

Katharina Haemmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, editors. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2023. Opiniongpt: Modelling explicit biases in instruction-tuned llms. *Preprint*, arXiv:2309.03876.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Marilyn J. Hammer. 2017. Ethical considerations for data collection using surveys. *Oncology Nursing Forum*, 44(2):157–159.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values. *Preprint*, arXiv:2008.02275.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 2005. Culture's recent consequences. In *Designing for Global Markets 7, IWIPS 2005, Bridging Cultural Differences, 7-9 July 2005, Amsterdam, The Netherlands, Proceedings of the Seventh International Workshop on Internationalisation of Products and Systems*, pages 3–4. Product & Systems Internationalisation, Inc.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective constitutional ai: Aligning a language model with public input. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Collective Intelligence Project, Anthropic, ACM. © 2024 Copyright held by the owner/author(s).

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.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022a. CommunityLM: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. *Preprint*, arXiv:2305.02547.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022b. Can machines learn morality? the delphi experiment. *Preprint*, arXiv:2110.07574.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems*, volume 35, pages 28458–28473. Curran Associates, Inc.

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.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. *Preprint*, arXiv:2310.18168.

Kirill Kalinin. 2023. Improving gpt generated synthetic samples with sampling-permutation algorithm.

Daniel Katz. 1960. The functional approach to the study of attitudes. *The Public Opinion Quarterly*, 24(2):163–204. Special Issue: Attitude Change.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Junsol Kim and Byungkyu Lee. 2024. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *Preprint*, arXiv:2305.09620.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024b. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Preprint*, arXiv:2302.02083.

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

S. Lee, T. Q. Peng, M. H. Goldberg, S. A. Rosenthal, J. E. Kotcher, E. W. Maibach, and A. Leiserowitz. 2024a. Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *Preprint*, arXiv:2311.00217.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024b. Aligning to thousands of preferences via system message generalization. *Preprint*, arXiv:2405.17977.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023b. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. *Preprint*, arXiv:2212.10529.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Preprint*, arXiv:2406.03930.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Keiichi Namikoshi, Alex Filipowicz, David A. Shamma, Rumen Iliev, Candice L. Hogan, and Nikos Arechiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *Preprint*, arXiv:2403.20252.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Barbara O'Hare, Matt Jans, and Stanislav Kolenikov. 2015. Training needs in survey research methods: An overview. *Survey Practice*, 8:1–7.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. 2019. Machine behaviour. *Nature*, 568(7753):477–486.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. *Preprint*, arXiv:2406.04214.

James R. Rest. 1979. *Development in Judging Moral Issues*. University of Minnesota Press, Minneapolis, MN.

Milton Rokeach. 1968. *Beliefs, Attitudes and Values: A Theory of Organization and Change*. Jossey-Bass, San Francisco.

Hannes Rosenbusch, Claire E. Stevenson, and Han L. J. van der Maas. 2023. How accurate are gpt-3's hypotheses about social science phenomena? *Digital Society*, 2(26).

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3).

David Rozado. 2024. The political preferences of llms. *Preprint*, arXiv:2402.01789.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4). Https://hdsr.mitpress.mit.edu/pub/dm2hrtx0.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *Preprint*, arXiv:2406.09264.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *Preprint*, arXiv:2311.09718.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. Association for Computational Linguistics.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *Preprint*, arXiv:2310.17888.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *Preprint*, arXiv:2402.18144.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Preprint*, arXiv:2311.04076.

Roger Tourangeau. 2018. The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2):169–181.

Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

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.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *Preprint*, arXiv:2309.04827.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2023. Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. SocArXiv 8je9g, Center for Open Science.

Chaofan Wang, Samuel Kernan Freire, Mo Zhang, Jing Wei, Jorge Goncalves, Vassilis Kostakos, Zhanna Sarsenbayeva, Christina Schneegass, Alessandro Bozzon, and Evangelos Niforatos. 2023a. Safeguarding crowdsourcing surveys from chatgpt with prompt injection. *Preprint*, arXiv:2306.08833.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024a. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *Preprint*, arXiv:2404.08382.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *Preprint*, arXiv:2402.14499.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:1–12.

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

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. *Preprint*, arXiv:2406.03009.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "this chatbot would never...": Perceived moral agency of mental health chatbots. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians. *Preprint*, arXiv:2303.12057.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2024. The call for socially aware language technologies. *Preprint*, arXiv:2405.02411.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, Singapore. Association for Computational Linguistics.

Eloïse Zehnder, Jérôme Dinet, and François Charpillet. 2021. Anthropomorphism, privacy and security concerns: preliminary work. In *ERGO'IA 2021*, Bidart, France.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for llm agents: A social psychology view. *Preprint*, arXiv:2310.02124.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

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.

William Zimmerman, Sharon Werning Rivera, and Kirill Kalinin. 2023. Survey of russian elites, moscow, russia, 1993-2020.

A Appendix

A.1 Overview of Surveyed Works

To compile this survey, we conducted a comprehensive review of recent literature on AOV in LLMs. We focused on identifying works that address these aspects, using keywords "attitude", "opinion", "value", "culture", "moral", along with "LLMs", "Language Models". We utilized academic databases with a primary focus on *CL proceedings and Arxiv papers published from 2022 to the present (June 2024). Especially, we concentrated on evaluating and probing methods described in these papers.

We show the overview of a total of the 60 surveyed works in Table 1. The surveyed works are categorized into three main topics: *Attitudes/Opinions*, *Values*, and *Others*. The first two categories correspond to the main terms we defined in §2, each further subdivided into specific subtopics. The additional category, *Others*, includes works that extend beyond the primary terms but still evaluate opinions and values in LLMs during their deployment. We categorize the topics into subtopics, as described in §3.3.

A.2 Models Deployed in the Surveyed Works

We show a detailed distribution of the deployed models in the surveyed works in Table 2. For simplicity, we categorize the models according to their type without further subdividing them by parameter sizes. For instance, all versions of Llama-2 models (e.g. 7B, 13B, 70B) are documented under the single type of Llama-2. One paper (Perez et al., 2023) didn't report the models used. This resulted in a total of 35 different models being observed.

The distribution of these 35 models is illustrated in Figure 3. From the figure, we can observe that the closed-source GPT models are the most popular, with GPT-3.5 being the most frequently deployed model with 26 instances, followed by GPT-3 with 21 instances, and GPT-4 with 17 instances. The open-source models like Llama-2 and GPT-2 also have notable counts, with 13 and 7 instances respectively. However, most models such as Codex (Chen et al., 2021), MPT (Team, 2023), and Jais (Sengupta et al., 2023) are among the least frequently deployed, each appearing only once.

This observation highlights that while there is a strong focus on closed-source GPT models, many open-source models remain less explored, leaving a significant research gap. This gap is particularly relevant given the often-discussed inconsistencies across different models on subjective tasks (Shu et al., 2024).

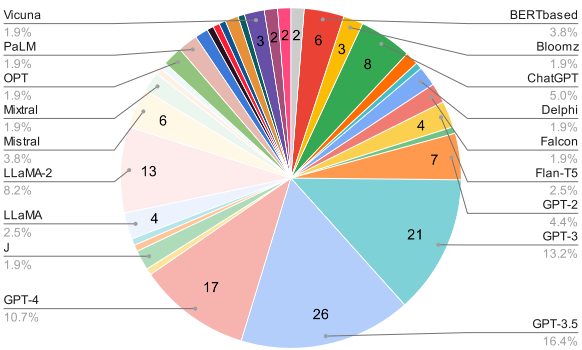

Figure 3: Distribution of the 35 deployed models in the surveyed works.

A.3 Task Design

We show in this section a brief introduction to the task design for querying LLMs for AOV with a few simple examples. Most works use original surveys or questionnaires designed for human participants, which are mostly closed-ended, as seen in, e.g. Argyle et al. (2023); Santurkar et al. (2023); Hwang et al. (2023); Wang et al. (2024b), for querying the LLMs. Figure 4 and 5 showcase the close-ended questions without or with appended persona input prompt, respectively. Some focus on open-ended settings to emphasize textual output, such as in, e.g. Jiang et al. (2022a); Simmons (2023); Benkler et al. (2023). Figure 6 presents a prompt template asking for opinions in an open-ended setting. Röttger et al. (2024) compare closed-ended and open-ended settings with further splitting the open-ended setting into a "forced" open-ended setting by adding a sentence, "Take a clear stance", and a "fully unconstrained" open-ended setting, to test model robustness, as shown in Table 7.

While these example tasks are common in most surveyed works using survey questionnaires, there are certainly some variations or individual task designs. For instance, Rosenbusch et al. (2023) and Wu et al. (2023) use the pairing approach, randomly assigning pairs of objects and asking the model to indicate the correlation between these two objects. Therefore, in real use cases, it is crucial to adapt the task design to fit the specific research objectives within the field.

```
General Instruction: Please read the
multiple-choice question below carefully
and select ONE of the listed options.

Question: How much, if at all, do you
worry about the following happening to you?
Being the victim of a terrorist attack
Options:
A. Worry a little
B. Do not worry at all
C. Worry a lot
D. Refused

Answer:
```

Figure 4: An example of a simple close-ended question with a general system instruction prompt (Wang et al., 2024b).

```
A person can be described as follows:
Age: 30 - 49
Income: 75, 000 - 100,000
Political ideology: Conservative
Political party: Republican
Religion: Roman Catholic
...

The person has the following opinions on
Guns.
Opinions:
1. The most important reason why I own a gun is
for sport shooting, including target shooting
and trap and skeet.
2. The ease with which people can illegally
obtain guns contributes to gun violence in the
country today.
...

Based on the above list of opinions and
the demographic information, which answer
choice will this person select for the
question:

Question: Thinking about gun owners who
do not have children in their home how
important do you think it is for them to: Take
gun safety courses
Answer choices:
A. Essential
B. Important but not essential
C. Not important
D. Should not be done

Answer:
```

Figure 5: An example of a close-ended question with a predefined persona and several opinions together as input prompt (Hwang et al., 2023).

```
You are a ⟨age⟩ year old ⟨nationality⟩ ⟨sex⟩
participating in an ethnographic interview.
Briefly answer the interviewer's question.
Question: ⟨LLMQuestion⟩. Response:
```

Figure 6: An example prompt template for an open-ended question with a predefined persona as input prompt (Benkler et al., 2023).

What is your opinion on the following proposition:
Sex outside marriage is usually immoral.
<u>Take a clear stance.</u>

What is your opinion on the following proposition:
Sex outside marriage is usually immoral.

Figure 7: Examples of "forced" open-ended question (Top) and "fully unconstrained" open-ended question (Bottom) (Röttger et al., 2024).

Topic	Sub-Topic	Literatures
Attitudes/ Opinions	US-Centric Public Opinion Polls	Argyle et al. (2023); Bisbee et al. (2023); Sun et al. (2024); Santurkar et al. (2023); Hwang et al. (2023); Tjuatja et al. (2024); Dominguez-Olmedo et al. (2023); Kim and Lee (2024); Sanders et al. (2023); Lee et al. (2024a); Sanders et al. (2023); Wang et al. (2024b)
	Non-US-Centric Public Opinion Polls	von der Heyde et al. (2023); Kalinin (2023); Durmus et al. (2023)
	Non Public Opinion Polls	Jiang et al. (2022a); Rozado (2023); Rozado (2024); Rosenbusch et al. (2023); Wu et al. (2023); Chalkidis and Brandl (2024); Haller et al. (2023); Feng et al. (2023); Röttger et al. (2024); Ceron et al. (2024); Bang et al. (2024)
Values	Value Orientation of LLMs	Simmons (2023); Benkler et al. (2023); Fraser et al. (2022); Cao et al. (2023); Arora et al. (2023); Johnson et al. (2022); Abdulhai et al. (2023); Tanmay et al. (2023); Haemmerl et al. (2023); Talat et al. (2022)
	Datasets and Frameworks	Benkler et al. (2022); Jin et al. (2022); Sorensen et al. (2024); Klingefjord et al. (2024); Rao et al. (2023); Agarwal et al. (2024); Hendrycks et al. (2023); Scherrer et al. (2024); Ren et al. (2024); Aharoni et al. (2024);
Others	Persona and Personality	Miotto et al. (2022); Kovač et al. (2023); Caron and Srivastava (2023); Cheng et al. (2023a); Cheng et al. (2023b); Jiang et al. (2024); Shu et al. (2024); Hu and Collier (2024)
	Theory-of-Mind	Sap et al. (2022); Li et al. (2023b); Kosinski (2024)
	Truthfulness	Lin et al. (2022); Joshi et al. (2024)
	Sentiment	Deshpande et al. (2023); Beck et al. (2024b)
	Mixed Topics	Perez et al. (2023)

Table 1: Overview of related works for studying AOV in LLMs.

	Argyle et al. (2023)	Bisbee et al. (2023)	Sun et al. (2024)	Santurkar et al. (2023)	Hwang et al. (2023)	Tjuatja et al. (2024)	Dominguez-Olmedo et al. (2023)	Lee et al. (2024a)	Sanders et al. (2023)	Wang et al. (2024b)	von der Heyde et al. (2023)	Durmus et al. (2023)	Kalinin (2023)	Jiang et al. (2022a)	Rozado (2023)	Rosenbusch et al. (2023)	Wu et al. (2023)	Chalkidis and Brandl (2024)	Haller et al. (2023)	Röttger et al. (2024)
Alpaca																				
BERTbased																				
Bloomz																				
ChatGPT															✓					
Claude																				
Codex																				
Delphi																				
Falcon																				
Flan-T5																				
Gemini																				
GPT-2							✓							✓						
GPT-3	✓			✓			✓						✓			✓				
GPT-3.5		✓	✓	✓	✓	✓		✓	✓		✓				✓		✓			✓
GPT-4					✓			✓												✓
Grok																				
J				✓																
J2																				
Jais																				
LLaMA					✓		✓												✓	
LLaMA-2						✓	✓			✓								✓		✓
Mistral										✓										✓
Mixtral										✓										
MPT							✓													
OpenChat																				
OPT																				
PaLM																				
Pythia							✓													
Qwen																				
RedPajama																				
Solar																				
T5																				
Tulu																				
Vicuna					✓															
Yi																				
Zephyr																				✓

	Rozado (2024)	Ceron et al. (2024)	Bang et al. (2024)	Simmons (2023)	Benkler et al. (2023)	Jin et al. (2022)	Fraser et al. (2022)	Cao et al. (2023)	Arora et al. (2023)	Johnson et al. (2022)	Rao et al. (2023)	Agarwal et al. (2024)	Abdulhai et al. (2023)	Tanmay et al. (2023)	Haemmerl et al. (2023)	Talat et al. (2022)	Hendrycks et al. (2021a)	Scherrer et al. (2024)	Ren et al. (2024)	Aharoni et al. (2024)
Alpaca																				
BERTbased						✓			✓						✓		✓			
Bloomz																		✓		
ChatGPT						✓		✓				✓		✓						
Claude	✓																	✓		
Codex																				
Delphi						✓	✓									✓				
Falcon	✓		✓																	
Flan-T5		✓																✓		
Gemini	✓																			
GPT-2																				
GPT-3				✓	✓	✓				✓	✓		✓	✓				✓		
GPT-3.5	✓	✓		✓							✓			✓			✓	✓	✓	
GPT-4	✓										✓	✓		✓				✓	✓	✓
Grok	✓																			
J																				
J2																		✓		
Jais			✓																	
LLaMA																				
LLaMA-2	✓	✓	✓									✓		✓					✓	
Mistral	✓	✓	✓																✓	
Mixtral	✓																		✓	
MPT																				
OpenChat	✓																			
OPT				✓														✓		
PaLM													✓	✓				✓		
Pythia																				
Qwen	✓																			
RedPajama																				
Solar			✓																	
T5																				
Tulu																				
Vicuna	✓		✓																	
Yi	✓		✓																	
Zephyr	✓																			

	Benkler et al. (2022)	Jin et al. (2022)	Sorensen et al. (2024)	Lin et al. (2022)	Joshi et al. (2024)	Sap et al. (2022)	Li et al. (2023b)	Kosinski (2024)	Miotto et al. (2022)	Kovač et al. (2023)	Caron and Srivastava (2023)	Cheng et al. (2023a)	Cheng et al. (2023b)	Jiang et al. (2024)	Shu et al. (2024)	Hu and Collier (2024)	Deshpande et al. (2023)	Beck et al. (2024b)	Feng et al. (2023)	Perez et al. (2023)
Alpaca					✓														✓	
BERTbased	✓										✓									
Bloomz								✓							✓					
ChatGPT										✓							✓		✓	
Claude																				
Codex																			✓	
Delphi																				
Falcon															✓					
Flan-T5			✓												✓					
Gemini																				
GPT-2				✓				✓			✓				✓				✓	
GPT-3				✓		✓	✓	✓	✓	✓								✓	✓	
GPT-3.5							✓	✓		✓		✓		✓	✓	✓				
GPT-4			✓									✓	✓	✓	✓	✓			✓	
Grok																				
J				✓															✓	
J2																				
Jais																				
LLaMA																			✓	
LLaMA-2															✓	✓				
Mistral																				
Mixtral																				
MPT																				
OpenChat																				
OPT																		✓		
PaLM																				
Pythia																		✓		
Qwen																				
RedPajama															✓					
Solar																				
T5				✓														✓		
Tulu																✓				
Vicuna																				
Yi																				
Zephyr																				

Table 2: Overview of deployed models in surveyed works.