Review

Not peer-reviewed version

Towards Pluralistic Alignment of LLMs: A Comprehensive Survey

Anudeex Shetty [†] , Usman Naseem [*,†] , Nikolaos Aletras , Mark Dras , Heng Ji , Preslav Nakov

Posted Date: 24 March 2026

doi: 10.20944/preprints202603.1876.v1

Keywords: large language models (LLMs); LLM alignment; pluralistic alignment

 doi:10.20944/preprints202603.1876.v1

Review

Towards Pluralistic Alignment of LLMs: A Comprehensive Survey

Anudeex Shetty [1,2,†], Usman Naseem [1,*,†], Nikolaos Aletras [3], Mark Dras [2], Heng Ji [4] and Preslav Nakov [5]

[1] Macquarie University
[2] University of Melbourne
[3] University of Sheffield
[4] University of Illinois Urbana-Champaign
[5] MBZUAI
* Correspondence: usman.naseem@mq.edu.au
† The first two authors are as listed, and the remaining authors are arranged alphabetically by surname.

Abstract

Pluralistic alignment instils large language models (LLMs) with the capacity to reflect diverse human values and preferences. It offers safe deployment that avoids LLMs collapsing into monolithic perspectives while ensuring that these systems operate in accordance with ethical standards and safety protocols. In this survey, we provide a comprehensive analysis of pluralistic alignment for LLMs based on the different capabilities that models support. We review existing literature covering different methods, datasets, and evaluation metrics, highlighting their strengths, limitations, and open challenges. Finally, we identify and encourage future areas for research to increase the scope and impact of pluralistic modelling in contemporary AI systems.

Keywords: large language models (LLMs); LLM alignment; pluralistic alignment

1. Introduction

As large language models (LLMs) exert growing influence, ensuring their beneficial and safe use has become increasingly important (Askell et al. 2021; Ji et al. 2023a; Minaee et al. 2024; Shen et al. 2023; Wang et al. 2024c). Accordingly, alignment has become central to responsible AI (Gabriel 2020; Ji et al. 2023a; Ouyang et al. 2022; Shen et al. 2023), which is the process of ensuring that AI model exhibits behaviour consistent with human values and intentions (Ji et al. 2023a; Shen et al. 2023). This has driven widespread adoption of aligned LLMs, mitigating issues such as privacy violations, toxicity, and misinformation observed in earlier systems (Bubeck et al. 2023; Carlini et al. 2023; Ji et al. 2023b; Weidinger et al. 2021). Despite these advances, alignment methods still face fundamental limitations. In particular, they tend to capture monolithic (or *average*) human values (Feng et al. 2024a; Shetty et al. 2025; Sorensen et al. 2024b), oversimplifying the alignment problem by failing to represent pluralistic human values (Conitzer et al. 2024; Sorensen et al. 2024b; Stanczak et al. 2025). Human values and requirements vary widely, not only across individuals, but also for the same individuals across different contexts (Bilsky et al. 2011).

To address this limitation, emerging research in pluralistic alignment (Conitzer et al. 2024; Sorensen et al. 2024b; Stanczak et al. 2025) examined how systems can be designed and deployed to reflect diverse human needs and values. Sorensen et al. (2024b) proposed a theoretical framework defining three modes how models could be pluralistic: *Overton*, *Steerable*, and *Distributional*. As illustrated in Figure 1, consider the query: "*Should the country drastically reduce immigration levels?*" Under the Overton mode, the model presents all reasonable answers, covering a range of perspectives based on political views, moral frameworks, *etc.* In Steerable mode, the model generates a response conditioned on a specified attribute, for instance, opposing the proposal from a far-left perspective or

 doi:10.20944/preprints202603.1876.v1

supporting restrictive immigration policies from a far-right perspective. Finally, in the Distributional mode, responses reflect population-level preferences. If 80% of the population opposes the proposal and 20% supports it, the model should respond "No" 80% of the time and "Yes" 20% of the time.

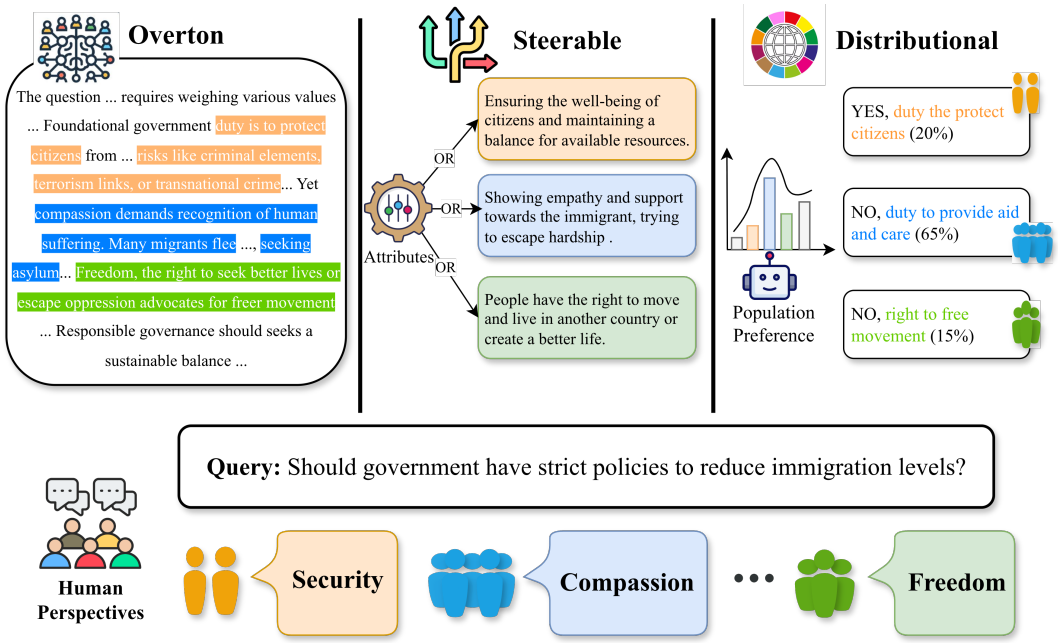

Figure 1. An overview of three pluralistic alignment modes (Overton, Steerable, and Distributional) for an example query: *"Should government have strict policies to reduce immigrant levels?"* Overton covers different perspectives, whereas Steerable responds as per attribute conditioned, and Distributional output match reflecting population preference distribution.

Over the past few years, pluralistic alignment has attracted significant attention, including numerous papers (Castricato et al. 2025; Chen et al. 2024a; Feng et al. 2024a; Shetty et al. 2025; Sorensen et al. 2024b, *inter alia)* and dedicated workshops at major AI and ML conferences (Shen et al. 2025; Terekhov et al. 2024). However, rapid progress has led to limited clarity about the field's current state, open gaps, and future directions. Existing work has demonstrated the potential for achieving pluralism in constrained settings and modes. Our goal is to provide a systematic and comprehensive guide that helps researchers identify appropriate tools and techniques and contribute to this rapidly evolving area. We are also the first to examine pluralism beyond LLMs. These definitions are generalisable to broader AI systems, and we include other modalities such as images, audio, and agents.

To this end, our survey presents an overview of the current landscape of pluralistic alignment. For each alignment mode (defined in §3), we provide a comprehensive overview of existing methods (§4) and evaluation approaches (§5), together with current shortcomings and directions for future research (§6). We aim to provide a clear entry point for researchers and facilitate further work on pluralistic LLM systems. We believe this survey is timely, offering a bird's-eye view of the literature while supporting continued progress in pluralistic alignment.[1]

The key contributions of this survey are as follows:

- To the best of our knowledge, this is the first survey dedicated exclusively to pluralistic alignment, providing a bird's-eye view of three modes—Overton, Steerable, and Distributional—which are often studied in isolation. We consolidate existing research and identify pluralistic methodologies, providing a practical roadmap for researchers.

[1] We additionally organise the papers covered by this survey at https://github.com/anudeex/Awesome-Pluralistic-Alignment .

 doi:10.20944/preprints202603.1876.v1

- We provide a detailed review of pluralistic alignment techniques, together with relevant datasets and evaluation metrics, helping researchers make informed methodological choices. For each category, we synthesise the literature and highlight open challenges and key takeaways.
- We examine understudied aspects of pluralistic alignment and its connections to adjacent fields, outlining promising directions for future work, including multi-turn interactions, agentic AI, temporal and geo-alignment, mechanistic interpretability, and safety risks. This survey aims to advance the development of generalist models that reflect pluralism and enable safe and responsible deployment.

2. Related Surveys and Key Differences

As AI systems are deployed to broader populations, prevailing preference-based alignment methods (Ouyang et al. 2022; Rafailov et al. 2023) that model average opinions risk treating value diversity as noise (Aroyo et al. 2023; Siththaranjan et al. 2024), contributing to bias (Benkler et al. 2023), fairness concerns (Gallegos et al. 2024), and under-representation (Ovalle et al. 2025). Equally concerning is that models may be aligned to reflect the biased preferences of their creators (Bender et al. 2021; Gabriel 2020; Weidinger et al. 2021, *inter alia)*, thereby supporting particular agendas rather than supporting the full spectrum of human values. Pluralistic alignment instead seeks to model this diversity directly. Despite its growing relevance, no prior survey provides a comprehensive overview of pluralistic alignment. Existing reviews focus on general alignment (Cao et al. 2024; Ji et al. 2023a; Shen et al. 2023), multimodal alignment (Shu et al. 2025; Yu et al. 2025), reward design (Ji et al. 2025), cultural alignment (Pawar et al. 2025a), or personalisation (Xie et al. 2025b; Zhang et al. 2024). We address this gap by presenting a systematic overview of alignment through the lens of supporting pluralism: consolidating current methods, clarifying distinctions, and outlining open challenges (shown in blue boxes at the end of each subsection), also connecting pluralistic alignment to broader LLM research.

3. Pluralistic Alignment

> *A modern democratic society is characterised not simply by a pluralism of comprehensive religious, philosophical, and moral doctrines but by a pluralism of incompatible yet reasonable comprehensive doctrines.*
>
> —Rawls (1971)

We define a pluralistic model as one that serves different users and adapts to their varied queries across various contexts. It is paramount to have a single foundation model that is *pluralistic* (Bai et al. 2022b; Gordon et al. 2022; Sorensen et al. 2024b). We study pluralistic behaviour in LLMs using a framework from Sorensen et al. (2024b), which we describe next, along with motivation. Then, we formally define each part of the framework in Section 3.2.

3.1. Sorensen's Framework

Figure 1 illustrates all three modes for an example query in a pluralistic AI system. The motivation behind these modes are as follows:

- **Overton** alignment expects the model's output to cover all feasible responses. Usually, there are many reasonable answers to a question (Min et al. 2020; Scherrer et al. 2023), and outputting a single reasonable answer promotes bias and limits representation. An Overton-pluralistic model instead covers all reasonable answers. The most relevant application of Overton pluralism is advice giving (Jakesch et al. 2023; Krügel et al. 2023), where responses should cover all perspectives and opinions.
- **Steerable** alignment allows the model to successfully orient itself to a particular attribute (*e.g.*, political view, religion, culture, among others). Customisation is key in steerable alignment where users often want to personalise models towards characteristic properties and perspectives (Chen et al. 2024c; Li et al. 2023; Ma et al. 2024).

 doi:10.20944/preprints202603.1876.v1

- **Distributional** alignment steers the model's response to match the underlying population distribution. Distributional alignment applies in scenarios where the model must reflect the views of a population. Example applications include social simulations (Park et al. 2022; Törnberg et al. 2023), autonomous vehicles (Cui et al. 2024), and related tasks.

3.2. Problem Formulation

To formalise the definition of pluralistic alignment, throughout we assume an AI system $\mathcal{M}$, that receives a query x and generates a response y.

Overton.

We define the set of reasonable answers $\mathcal{R}$ for a given query x as those responses that reflect an interpretation or solution that the majority of the population would agree upon. However, additional filters (such as safety, fairness, *etc.*) are applicable. We denote the set of all such reasonable answers by the Overton window $W(x)$ (OED 2026):

$$W(x) = \{y \in \mathcal{Y} \mid (x, y) \in \mathcal{R}\}.$$

A response set $\{y\}$ is Overton-pluralistic for a given query x if it consists of all reasonable answers (*i.e.*, $W(x)$). Similarly, a model $\mathcal{M}$ is Overton-pluralistic if for a given input x,

$$\mathcal{M}(x) = W(x).$$

Steerable.

This is the case where the model must faithfully align (or steer) a response y to a given query x towards a specified attribute or perspective. Steering attributes $\mathcal{A}$ denote the set of such attributes, including values, perspectives, populations, frameworks, and others.

We say that a response $y_{x,a}$ reflects an attribute $a \in \mathcal{A}$ if the response y to query x is consistent with a. Accordingly, a model $\mathcal{M}$ is steerable-pluralistic with respect to attributes $\mathcal{A}$ if, for every input x :

$$\forall a \in \mathcal{A}, \quad \mathcal{M}(x, a) = y_{x,a}.$$

Distributional.

Here, we define a reference population $\mathcal{G}$ as the set of people that a distributionally-pluralistic model aims to represent. A distributionally aligned model $\mathcal{M}$ then responds y to a given query x according to the reference population distribution. In other words, $\mathcal{M}$ is well calibrated over reasonable answers y with respect to the reference population.

$$P_{\mathcal{M}}(y \mid x) = P_{\mathcal{G}}(y \mid x)$$

4. Pluralistic Alignment Methodologies

Given Sorensen's framework, we categorise methodologies introduced in prior work and provide avenues for future research. Figure 2 organises existing literature as per pluralistic alignment modes supported.

 doi:10.20944/preprints202603.1876.v1

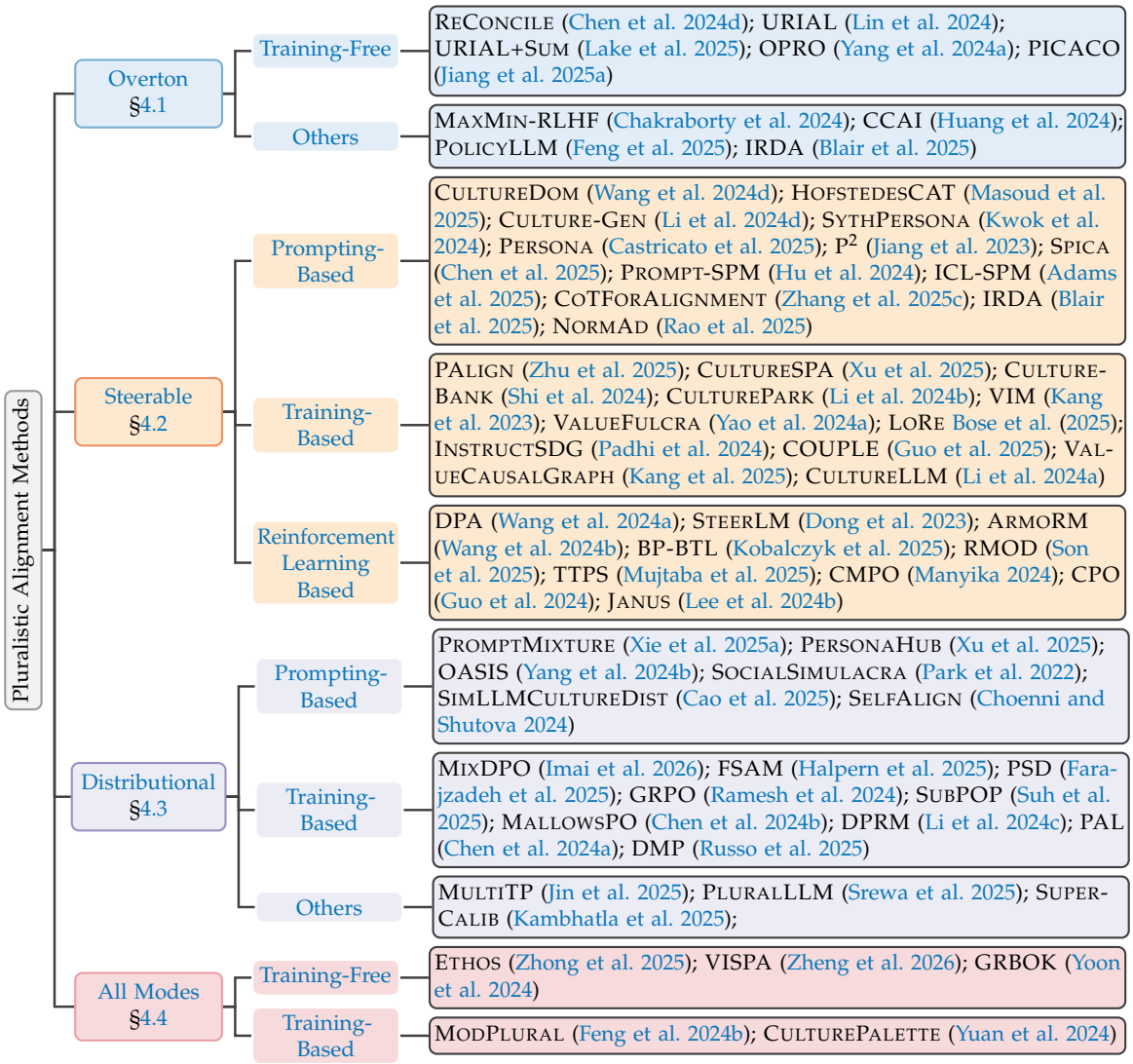

Figure 2. An overview of prior work in pluralistic alignment through the lens of the different modes supported. The leafs are hyperlinked to corresponding subsection.

4.1. Overton

Training-Free.

ReConcile (Chen et al. 2024d) is one of the initial works supporting Overton mode. It is an inference-time-based multi-agent method involving collaborative discussion (*i.e.,* multiple rounds) among LLMs, leveraging diversity across different models to synthesise a response with a higher Overton window coverage. PICACO (Jiang et al. 2025a) is the most recent in-context alignment (ICA) method for Overton mode (or value coverage) working solely on inference time without any training. At its core, it is a prompting-based technique trying to elicit diverse responses embedded in the LLM (Cheng et al. 2024; Ganguli et al. 2023). However, unlike prior ICA approaches, it does not need multiple LLMs as in OPRO (Yang et al. 2024a) for iterative prompt optimisation, sophisticated time-consuming human-crafted meta-instructions as in URIAL (Lin et al. 2024), or the strong LLM extension of URIAL in URIAL+Sum (Lake et al. 2025). PICACO maximises the total correlation between target values—comprising Helpful and Harmless (Bai et al. 2022a; Weidinger et al. 2021) and the Schwartz Value Theory (Schwartz 2012)—and generated responses using a feedback loop to construct effective prompts. This also combats the instruction bottleneck, *i.e.,* LLMs struggle to comprehend multiple values in a single prompt (Chen et al. 2024e; Jiang et al. 2024). Including the well-known prompting issues (Chakraborty et al. 2024; Sahoo et al. 2024; Zhao et al. 2024), this work only evaluates a limited set of values and uses a potentially biased LLM scoring for value coverage.

 doi:10.20944/preprints202603.1876.v1

Others.

Feng et al. (2025) propose a policy-based work primarily focused on Overton mode. It is grounded in a real-world 15-week observational study of LLM policymaking workshops in an industrial AI lab. They advocate that LLM policy prototyping could be an interesting avenue for pluralistic alignment because policy prototyping encourages discussion with contrasting preferences and opinions. Similarly, Collective Constitutional AI (CCAI) (Huang et al. 2024) proposes a multi-stage process that sources diverse rules from the public and incorporates them in LLM development. This builds on the foundations of Constitutional AI and can be seen as an extension of reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022), promoting that the responsibility of LLM should not rest only with developers.

Overton Takeaways

- Overton responses are inherently long-form and are therefore not applicable to one-word responses from LLMs.
- The formulation also does not account for uncertainty or intent, and is more suitable for advice-giving scenarios.
- Defining reasonable Overton window and reasonable answers remains challenging, as both depend on context and situation.
- An ill-defined Overton window can lead to false balance. We encourage the research community to work more on Overton mode, considering the lack of current research and importance of *generalist* of the frontier LLMs.

4.2. Steerable

Prompting-Based.

LLMs exhibit a degree of steerable pluralism via prompting (Sorensen et al. 2024b). Initial works achieved steerability by simply setting different contexts in the system roles of LLMs in a zero-shot manner. One simple approach is "role-playing", where LLMs simulate a person from a particular culture, i.e., CULTUREDOM (Wang et al. 2024d), CULTURE-GEN (Li et al. 2024d), and HOFSTEDESCAT (Masoud et al. 2025), or demographics, i.e., NORMAD (Rao et al. 2025) and SYTHPERSONA (Kwok et al. 2024). Similarly, methods that include user history in prompts lead to better steerability, e.g., PERSONA (Castricato et al. 2025).

The relevance of ICA methods for Overton (and in general) alignment (§4.1) has also been studied for steerability. One of the first works showed that a simple few-shot personality prompting (P^2) (Jiang et al. 2023) using Big Five traits was effective in inducing personality. Adams et al. (2025) improved ICA for the steerable mode. It proposed ICL-SPM, a few-shot comparative regression-based approach. It uses in-context prompting for improved steerable accuracy, with chain-of-thought reasoning for explainability, and LLM-as-Judge for reducing bias in candidate selection. Overall, as in ICA methods, no fine-tuning or model merging is needed. Expectedly, ICL-SPM is shown to improve on PROMPT-SPM (Hu et al. 2024) (a zero-shot prompt-based alignment method) for MIC (Ziems et al. 2022) and HELPSTEER2 (Wang et al. 2024e) benchmarks. SPICA (Chen et al. 2025) is another in-context alignment technique for the steerable mode. It accounts for group-level differences during example retrieval for in-context prompting, rather than only similarity. They show improved steerability on VALUEPRISM (Sorensen et al. 2024a) and PRISM (Kirk et al. 2024a) datasets. However, the study is limited to only four demographic groups and lacks strong steerable baselines.

Training-Based.

The culture and personality sub-dimension of pluralistic alignment (AlKhamissi et al. 2024) has been of interest in the community, especially their adaptability (or steerability). There have been several works (Bose et al. 2025; Li et al. 2024a,b; Shi et al. 2024; Xu et al. 2025; Zhu et al. 2025, *inter alia)* collecting various cultural or personality data (with the hope of more representation) and then fine-tuning LLMs to achieve better steerability. Taking a different route, Guo et al. (2025) proposes COUPLE, an inference-time causal-based method, improving on previous causal-based work VALUECAUSALGRAPH (Kang et al. 2025). The paper argues that existing solutions consider all values to be independent and/or equally weighted, failing to capture their causal interdependencies between

 doi:10.20944/preprints202603.1876.v1

them. COUPLE builds a causal graph between values and model response using an LLM, then applies counterfactual reasoning to generate outputs aligned with a specific given value. They demonstrate better performance compared to existing prompting and tuning methods, but it is limited to the culture sub-dimension of pluralism and need a strong reasoning model.

Reinforcement Learning Based.

In need of controllable LLMs, previous work has explored fine-tuning for multi-objective using scalar feedback for user-steerable LLMs as an alternative to RLHF or DPO (single-objective). STEERLM (Dong et al. 2023) was one such prominent supervised fine-tuning method. This method conditions its generation to conform to a set of attributes. The reward model in RLHF is replaced with an attribute prediction model, which predicts the human preference for different attributes. This work was further improved and extended using better multi-attribute datasets, HELPSTEER (Wang et al. 2024f) and HELPSTEER2 (Wang et al. 2024e). STEERLM has the same limitation of reward mispecification as in RLHF due to the use of reward models. This was rectified in several follow-up works (Guo et al. 2024; Manyika 2024; Wang et al. 2024a,b, *inter alia*). Similarly, there have been several position works (Harland et al. 2024; Vamplew et al. 2024; Xiong and Singh 2025) sketching out the connection between multi-objective reinforcement learning and steerable pluralistic alignment.

Steerable Takeaways

- Because steerable pluralism conditions responses on attributes, it reduces ambiguity relative to unconditioned settings.
- However, important questions remain, such as which attributes to steer towards. If an attribute is too broad (for example, a broad demographic category), substantial variation within the attribute can make steering internally conflicting.
- Steerable pluralism raises concerns about dual use when attributes are harmful (for example, self-hate), some degree of reasonable filtering may therefore be necessary.
- Counterfactual reasoning coupled with causal graph direction might be effective to capture interdependency among values and perspectives.
- Steerability in LLM is far from solved, although it has gotten the most attention. Further work is needed on the steerability of LLM agents and others.

4.3. Distributional

Santurkar et al. (2023) were among the first to study distributional alignment, contributing OPINIONQA, an instrumental survey-based benchmark for evaluating pluralistic alignment. Expectedly, they established notable misalignment between the views reflected in LLMs and those of US demographic groups. Meister et al. (2025) empirically study the distributional mode on GLOBALOPINIONQA, OPINIONQA, and their new dataset, NYTBOOKOPINIONS, that expands measurements beyond political and cultural values. Results reveal systematic weaknesses in LLMs as demographic simulacra, i.e, describing human distributions better than sampling them. This work also coined the term: "distributional alignment", laying out the evaluation protocols and datasets.

Prompting-Based.

Xie et al. (2025a) designs a prompt mixture modelling solution (PROMPTMIXTURE) for distributional alignment, focusing on social simulation (or distributional mode). It comprises a pool of prompts, each representing a subpopulation, and then learns an optimal mixture for these prompts (Park et al. 2022). Finally, prompts are sampled according to the weights to match the underlying human distribution, simulating the actual population. The method offers improvements over existing similar social simulation techniques PERSONAHUB (Xu et al. 2025) and OASIS (Yang et al. 2024b). PERSONAHUB sets a persona as a system prompt sampled from a pool of personas, while OASIS uses demographic-inspired personas from the Big Five dataset. Likewise, Kambhatla et al. (2025) studies distributional calibration of LLMs for different social science surveys. Simple supervised calibration of probabilities generated by LLMs improve distributional alignment. The LLMs tested for different variations of socio-demographic prompts (e.g., SIMLLMCULTUREDIST (Cao et al. 2025), SELFALIGN (Choenni and Shutova 2024)), showing effective calibration. A core issue is that heavy

 doi:10.20944/preprints202603.1876.v1

reliance on demographics and heuristics for persona sampling could lead to oversimplification and under-representation (Beck et al. 2024; Cheng et al. 2023a,b).

Training-Based.

PAL (Chen et al. 2024a) is a training-time framework for pluralistic alignment, capturing and preserving heterogeneity, addressing the core limitation of RLHF (single, universal preference). It introduces the lens of the ideal point model and metric learning, using a ground-up mixture modelling approach. It inherently supports distributional and steerable modes in this order. Likewise, MALLOWSPO (Chen et al. 2024b) leverages dispersion-based modelling, and DPRM (Li et al. 2024c) uses multi-attribute annotations along with optimal-transport-based loss to better align with diverse human preferences. MIXDPO (Imai et al. 2026) is a generalisation of DPO, a prominent RLHF technique that optimises models by directly comparing preferred and less-preferred outputs in preference data. In DPO and previous approaches, all human preferences have uniform strength, leading to average (or monolithic) human preference learning. In contrast, MIXDPO has variable preference strength (based on individual and context), which is more suitable for pluralistic alignment. They demonstrate that MIXDPO leads to improvement compared to DPO. Although it lacks evaluation of the modes of pluralistic alignment specifically, it is most applicable to the distributional mode because it captures heterogeneous preference strength. Halpern et al. (2025) proposes an ensemble of rewards heuristic, FSAM, instead of a single reward model to better represent pluralism. It treats annotator disagreement as soft labels instead of noise. They learn this diverse human preference distribution via pairwise calibration, where the proportion of rewards models preferring one preference captures the fraction of annotators (or population) with that preference. The work demonstrated its effectiveness both theoretically and empirically. However, all these works need further empirical investigation on standard distributional benchmarks.

Distributional Takeaways

- A central challenge in distributional alignment lies in defining the target population. For example, for systems such as ChatGPT, what is the target population?
- Similarly, for open-ended questions, the relevant population may be ill-defined. There is also a risk of generating harmful responses through population reflection.
- Human label variation field (Plank 2022) is relevant for considering variability in labels and some ideas could be borrowed.
- Considering the importance of language and representation for diverse human perspectives and values, we need future works studying pluralism in different languages.
- The PLURALLLM work (Srewa et al. 2025) presents the first study of pluralistic alignment in the federated learning space; more research is needed to explore other modes of pluralism.

4.4. All Modes

Training-Based.

MODPLURAL (Feng et al. 2024b) was one of the foundational works tackling all three modes of pluralistic alignment. It is a model-collaboration technique where the main LLM works in collaboration with a set of specialised smaller LLMs to achieve pluralism. They show for different types, sizes, and architectures of main LLMs empirically better compared to standard baselines: off-the-shelf model, pluralistic prompting, and mixture of experts (pool of specialised LLMs). As per the mode of alignment, collaboration is conducted. For Overton, the main LLM acts as a summariser for specialised LLM responses and compiles a final representative (higher Overton window coverage) answer. In Steerable, one relevant response from the pool is chosen, and the main LLM is steered to output conditioned on this chosen response. Finally, for distributional, the process is similar, where the main LLM generates conditioned on each pool of responses. The point to note is that instead of focusing solely on text, it aggregates multiple answer probability distributions resembling the underlying population preference distribution. In the paper, they evaluate two sets of specialised LLMs, perspectives-informed across left/centre/right for social media texts and community-based covering different continents. However, the specialised LLMs are basically fine-tuned 7B LLMs on custom datasets. The main selling point of MODPLURAL was due to the decoding-time collaboration; a new specialised LLM could be consulted

 doi:10.20944/preprints202603.1876.v1

or patched as required. Still, the requirement of special datasets and computationally heavy fine-tuning bottlenecks remains a challenge. There have been subsequent works trying to replace the pool of specialised LLMs with training-free alternatives, we discuss cover next.

Training-Free.

Follow-up work, ETHOS (Zhong et al. 2025), replaced this fine-tuned pool of LLMs with a pool of personas appropriate for the query. The core tenet is to elicit different answers from the same secondary LLM. Though scalable and training-free, the known issues of prompting (Chakraborty et al. 2024; Sahoo et al. 2024; Zhao et al. 2024) and its brittleness still persist. A recent work, VISPA (Zheng et al. 2026), improves on ETHOS and MODPLURAL with state-of-the-art performance on VITAL (Shetty et al. 2025) and general pluralistic alignment benchmarks (from Feng et al. (2024a)). VISPA is inspired by the success of activation steering for capturing different concepts and values, and better-controlled generation compared to prompting. The method is not heavily impacted by the choice of different activation steering instantiations. Therefore, with further improved methods in activation steering, it is a potential scalable and interpretable pathway to pluralistic alignment. This work also takes the first step for combating representation and bias issues by incorporating a comprehensive set of values considered for internal steering, including non-WEIRD (Western, Educated, Industrialised, Rich, and Democratic).

Overall Methodology Takeaways

- The majority of work has focused on Steerable and Distributional modes in contrast to the Overton mode.
- Social values can vary significantly at sub-country levels; most studies concentrate solely on country-level analyses (Pawar et al. 2025b; Qiu et al. 2025). We need more granular research capturing local diversity in social values beyond national averages.
- Many proposals remain theoretical or are merely thought experiments at this point.
- We encourage the community to contribute on all fronts of data, evaluation, and methods for Overton pluralism and realise true generalist models.
- Further evaluations for multi-lingual and multi-turn settings are needed to extend this line of work.
- To address the negative effect of instruction-tuning on pluralistic alignment, first steps were taken in Sorensen et al. (2025). However, further studies on its effectiveness on large-scale models, different architectures, hyperparameter tuning, and full-scale data need to be conducted.
- Additional other post-training and pre-training methods remains an open intriguing research question.
- Ideas from parallel fields of continual learning (Shi et al. 2025) and machine unlearning (Nguyen et al. 2025) might be applicable to support evolving dynamic pluralistic alignment.

5. Evaluation of Pluralistic Alignment

5.1. Datasets

In this section, we cover datasets in the pluralistic alignment literature. We also go beyond and cover relevant potential datasets from related literature. We predominantly cover benchmark datasets across different dimensions, which might be of interest to readers, like modes, languages supported, modalities, turn settings covered, and others. A detailed table Appendix Table A1 covers all datasets and metadata, providing a bird's-eye view. To be consistent with previous discussion on methodologies, we cover benchmarks as per alignment modes.

Overton.

An Overton benchmark should comprise a scenario/situation and corresponding applicable and reasonable answers as defined in Section 3.2. VALUEPRISM (Sorensen et al. 2024a) was the first Overton-specific dataset that consists of a scenario and different values applicable to it. These different values applicable to a scenario represent the Overton window. It is a large-scale dataset formed from a human online platform, DELPHI (Jiang et al. 2025b), with diligent dataset preparation. However, it only encompasses value pluralism, and this is not conclusive of all aspects of pluralism.

Steerable.

A steerable benchmark consists of situations that contain a question and a list of responses, where each response is labelled with a set of attributes (this can be values, perspectives, or properties of

 doi:10.20944/preprints202603.1876.v1

interest). And, we expect the model to steer towards a specific response for a given attribute and question. Therefore, surveys (Durmus et al. 2024; Inglehart et al. 2000; Santurkar et al. 2023) are relevant because they document different preferred answers (as per demographics, country, culture, and other attributes) for the same question. Similarly, the literature has numerous datasets on moral dilemmas (Chiu et al. 2025; Ren et al. 2024; Scherrer et al. 2023; Sorensen et al. 2024a), conditioning different and often conflicting values, making them suitable as steerable datasets. Recently, chatbot preferred response datasets (Kirk et al. 2024a; Zhang et al. 2025a) with rich user information have been contributed. As in other fields of current research, we have several synthetic datasets contributed. Castricato et al. (2025) introduced a large-scale synthetic dataset PERSONA to evaluate steerability across diverse user preferences using LLM-as-Judge. Recently, in Jiang et al. (2025c), they propose the INDIEVALUECATALOG dataset, a transformed version of World Values Survey (WVS) to study individual steerability.

Distributional.

For distributional-pluralistic system evaluation, we need a dataset with a query and a suite of answers along with a target distribution over these answers (usually collected by many annotators). Again, existing survey data is ideal for such an evaluation (Durmus et al. 2024; Jiang et al. 2023; Santurkar et al. 2023). The works here leverage a multiple-choice dataset and try to capture whether the model distribution matches the underlying golden distributions on limited buckets of multiple-choices. GLOBALOPINIONQA is a prominent cross-national survey capturing opinions on global matters. Whereas the Machine Personality Inventory (MPI) is a collection of 120 questions measuring human personality traits, these are some prominent datasets used for distributional alignment evaluation. Similarly, there are also several training datasets comprising answer preference pairs from traditional alignment literature applicable in distributional alignment (see Appendix Table A1), both for evaluation and training.

All Modes.

Feng et al. (2024b) proposed first benchmark covering all thee modes pluralistic alignment. They also re-purposed existing datasets, contributing a benchmark using VALUEKALEIDO for Overton and Steerable; OPINIONQA for Steerable; MORALCHOICE and GLOBALOPINIONQA for Distributional mode evaluation. VITAL (Shetty et al. 2025) has taken the first steps, being the first comprehensive pluralistic alignment benchmark. And, subsequent works (Zheng et al. 2026; Zhong et al. 2025) have built on this benchmark and contributed improved pluralistic alignment methods. Similarly, a few culture-focused datasets: CULTURALPALETTE (Yuan et al. 2024) and CULTURALKALEIDO (Banerjee et al. 2025) have been contributed. Although these works do not evaluate all three modes, given the nature of the datasets, there is scope to support all three modes. However, they are limited to cultural aspects. Similarly, VITAL is healthcare-focused.

Dataset Takeaways

- There is a lack of benchmarks covering all three modes of pluralistic alignment.
- Most of the datasets used in this field are survey-based and are re-purposed from existing datasets, comprising multiple-choice question types which do not fully encapsulate real-world perspectives and values.
- One more issue is that most of the evaluation datasets are relatively old compared with high possibility of data contamination in LLMs. Therefore, we need to have evaluation benchmarks that are evolving with time and not fixed.
- Some of these preference-based datasets could be used to impart pluralism during the pre-training stage.
- We need more benchmarks representative of other domains (such as VITAL for healthcare) for a holistic evaluation of the pluralistic alignment.
- Most of these datasets are English-based and in static single-turn setting. To be truly pluralistic, it should cover all the languages and multi-turn conversations. We advocate for datasets in such settings as they will bring additional complexity and challenges as values and objectives evolve.
- We need more datasets and additional studies on methodologies for other modalities. LLMs are no longer limited to text, they are applied to other modalities such as images, videos, audio, and others.

 doi:10.20944/preprints202603.1876.v1

5.2. Evaluation Metrics

Evaluation is important for alignment research, especially for the development of empirical alignment methods. We survey metrics that have been developed and used to evaluate pluralistic alignment. Fundamentally, these metrics are specific to alignment mode being evaluated. Other than standard text generation metrics (such as BERTSCORE) or LLM-as-Judge evaluation, the most prominent and best Overton window coverage evaluation was defined in Feng et al. (2024b). It proposes a metric using natural language inference (NLI) model entailing whether a set of golden values is covered in the response. Auxiliary, they recommend pairwise win-rate evaluation among two responses (*i.e.*, which response seems more pluralistic) using human and LLM annotators, though not scalable. The same was adopted in several subsequent works (Zheng et al. 2026; Zhong et al. 2025). Recently, Poole-Dayan et al. (2025) proposed the first framework to measure Overton pluralism, OVERTONSCORE. They formalise it as a set-coverage metric with each value weighted as per the fraction of survey participants in support of it (instead of equal weights). Additionally, they propose a standard benchmark and demonstrate scope for further improvement for Overton coverage in current LLMs (achieving ≈0.4 of the maximum 1.0). More importantly, there is a strong correlation between OVERTONSCORE and human judgement, providing a scalable and reliable metric for Overton evaluation and future system progress. The steerability of models is usually measured by the accuracy of such correct steering in the LLMs. Similarly, for Distributional alignment mode the model distribution with the target human distribution using the Jensen-Shannon distance, where lower values mean higher distribution similarity. This distance is averaged for each sample in the benchmark. Similarly, some works have used Earth-mover distance (or Wasserstein distance) for measuring similarity between distributions.

Metrics Takeaways

- Almost all evaluations strongly assumes each value has the same weightage. This is not true in the real world, where values might hold varying importance.
- Similarly, most of the works using LLM-as-Judge evaluation, which could be biased and not representative enough for alignment evaluation.
- We advocate for human-centred evaluations whenever possible such as OVERTONSCORE.

6. Discussion and Future Directions

While the research community has made substantial strides, pluralistic alignment is far from a solved problem. There are many under-explored and promising research directions in pluralistic alignment. In this section, we outline directions for future research in addition to highlighted points above (denoted by pointers).

Dataset Challenges.

Most existing work assumes a single-turn setting with static user representations. However, in a practical setup, user representations should evolve across multiple interactions (*i.e.*, multi-turn) (Chakraborty et al. 2024; Durmus et al. 2024; Klassen et al. 2024a). For example, a user's stance on a topic may shift after exposure to new information, or certain value preferences may change depending on prior interactions (Berinsky 2017; Grimm 2010). As a first step, future work should prepare a dataset for evaluating pluralism in multi-turn settings, followed by methods to handle it. A comprehensive multi-modal benchmark remains challenging, similar to traditional alignment (Wu et al. 2023; Yin et al. 2024b). Moreover, pluralistic alignment must account for diverse perspectives across cultures and languages. However, as shown in Table Appendix Table A1, current datasets are predominantly English and Western-centric. Advancing this field requires focused efforts on better data representation.

Better Pluralistic Methods.

In addition to the takeaways from Section 4. Current works treat alignment as a static solution; however, human preferences and values evolve over time. We suggest recent approaches evaluating fairness over time could be applied to temporal pluralism (Klassen et al. 2024b). There is also a lack

 doi:10.20944/preprints202603.1876.v1

of work ensuring pluralistic alignment in the context of agentic AI. Alamdari et al. (2024) sketches a connection in agents learning policies while being considerate of others in the environment and the future well-being. However, the paper lacks empirical demonstrations, and further work is needed to study the effectiveness of such alignment policies, ensuring collective social welfare while achieving their goals. Considering that AI agents are deployed in the wild, a pluralistic geo-alignment is also important where agents' actions vary depending on geography, culture, legality, systems, and evolve over time. Such spatio-temporal alignment is advocated in Janowicz et al. (2025), and there is a need for future work on all fronts.

Advancing Pluralistic Evaluation.

We highlight the need for multi-dimensional benchmarks (*e.g.*, culture, norms, and demographic groups) for evaluating performance across varying contexts. For instance, medical advice might change depending on the culture (AlKhamissi et al. 2024; Shetty et al. 2025). Future work should develop improved pluralistic sensitivity metrics. Similarly, the reasoning integrity of pluralistic models remains under-explored. Evaluations should consider logical consistency, epistemic calibration, and causal understanding (Lanham et al. 2023; Wei et al. 2022). Finally, evaluation should go beyond utility and assess potential alignment taxes (Ouyang et al. 2022; Yuan et al. 2023) introduced by pluralistic alignment, including over-refusal (Cui et al. 2025), increased jailbreak susceptibility (Greenblatt et al. 2024), reduced expressivity (Kirk et al. 2024c; Meister et al. 2025), and related trade-offs (Perez et al. 2023; Wolf et al. 2024). We encourage the use of automated red-teaming (Wang et al. 2023), as well as jury (Bai et al. 2022b; Sorensen et al. 2024b) and debate-based (Michael et al. 2023; Xu et al. 2024) evaluation frameworks for studying value conflicts and their resolution (Kirk et al. 2024b). Improved evaluation also requires reliable auditing; interpretability might help, which we discuss next.

Mechanistic Interpretability.

As these "black-box" LLMs are used for high-stakes decisions, it is crucial to understand their internal workings (or *interpretability*) (Lipton 2018; Shen et al. 2023; Vilone and Longo 2020). Although interpretability might not seem directly applicable to alignment, techniques in this area could still be beneficial. One particular direction, *mechanistic interpretability* (Bereska and Gavves 2024; Nanda et al. 2023), is an interesting avenue whose objective is to reverse engineer the reasoning processes of LLMs. It elucidates the internal mechanisms by which LLMs transform inputs into outputs, providing deeper insight into their internal workings. VISPA (Zheng et al. 2026) showed promising results by applying ideas from mechanistic interpretability at inference time to achieve pluralism. They steer models' internal activations at run time to satisfy different pluralistic objectives depending on the context. We advocate for further research on the interplay between mechanistic interpretability and pluralistic alignment. This work could extend beyond post-hoc applications and also be used to monitor and guide the pre-training process, ensuring that desired objectives (*i.e.*, interpretability evaluation metrics) are met (Naseem 2026; Sharkey et al. 2025).

Other Framework Extensions.

Although the jury-pluralistic benchmark was briefly touched on in Sorensen et al. (2024b), it is another way of aggregating diverse preferences. In jury learning, the model aims to resolve these disagreements explicitly via a jury, *i.e.*, groups in some proportions defined as per task determine the final prediction label. Other extensions include POPE (Huang et al. 2025), the first framework for offline pluralistic policy alignment in LLMs, which targets diverse rather than averaged preferences; deliberative approaches inspired by social science and law, sketching consensus-seeking in LLMs (Blair et al. 2025); and temporal pluralism to handle evolving values over time, adapting temporal fairness evaluation methods so AI reflects different stakeholders' values at different periods (Klassen et al. 2024a,b). In agentic AI, work remains limited despite agents' growing deployment: Alamdari et al. (2024) proposes that agents learning policies considerate of others' future well-being can foster pluralistic alignment, though empirical evidence is lacking and further study is needed for collective

 doi:10.20944/preprints202603.1876.v1

welfare; PLURALS (Ashkinaze et al. 2025) provides a multi-agent deliberation framework with persona-based social ensemble simulations, showing promise in case studies; and geo-alignment (Janowicz et al. 2025) advocates spatio-temporal adaptation to geographic, cultural, and legal contexts for agents in real-world spaces. Related ideas draw from legal theories on democratic legitimacy and deliberation (Caputo 2024), and political pluralism seeking (approximations of) neutrality amid subjectivity (Fisher et al. 2025). These conceptual extensions merit more empirical validation and integration into pluralistic alignment research.

Misuse of Pluralistically Aligned AI.

There is a risk that AI systems could be used for manipulative purposes, such as influencing people's political opinions and social behaviours, if they are customised to closely match personal values. Additionally, users' autonomy may be diminished by hyper-individualised human-AI interaction, endangering independent thought. Safeguards should be in place to guarantee that AI systems empower users rather than manipulate them based on their personal values, preserving diversity and fairness in the process. The current research in this space have only evaluated pure performance, attacks are not evaluated (Ji et al. 2023a). Although there is a need for more datasets for pluralistic alignment, one must be aware of potential privacy infringement. These alignment techniques use data that contains private details, such as individuals' values and preferences (Jiang et al. 2025d). This is further exacerbated by the anthropomorphisation of LLMs through user tailoring to their values, which increases the risk of further private information leakage. Furthermore, when using real-world data, researchers must obtain consent and be transparent about the data usage to the users.

7. Conclusion

We highlighted the critical role of pluralistically aligned LLMs for improving their overall safety. We analysed pluralistic alignment through a framework in three modes: Overton, Steerable, and Distributional, providing a comprehensive overview of current strategies and their applications. By reviewing the relevant literature, datasets, and evaluation measures, we identified important research gaps and outlined the core limitations of current methodologies. Future research should aim to broaden pluralism, encompassing a wider range of applications and dynamic contexts. By further advancing pluralistic alignment approaches to be more adaptive and context-aware, we can support the development of AI systems that better reflect diverse human perspectives while remaining robust, reliable, and appropriately balanced in terms of user helpfulness.

 doi:10.20944/preprints202603.1876.v1

Appendix A

Table A1. Overview of datasets for pluralistic alignment covering different dimensions. 'Modes' represents different pluralistic alignment modes supported: O - Overton, S - Steerable, D - Distributional. 'Annotator' captures how this dataset was annotated, 🤖: LLM, 👤: human, 🤖👤: mostly LLM-based, some human-based. 'Question Type' covers ✎ - Free Text, ⚖ - Preference Pairs, ☰ - QnA type. 'Modality' of samples are denoted as: 🗎 = Text, 🖻 = Image. 'Lang.' denotes languages in dataset: ABC - English, 🌐 - Multilingual, and Kr - Korean. 'Multi-turn' represents if samples in the dataset are for multi-turn setting.

Dataset	Modes	Size	Annotator	Question Type	Modality	Lang.	Multi-Turn?	Source
OvertonScore(2025)	● Ⓢ Ⓓ	60	👤	✎	🗎	ABC	×	US Human Study
IndieValueCatalog(2025d)	Ⓞ ● Ⓓ	93k	👤	✎	🗎	ABC	×	WVS
DailyDilemmas(2025)	Ⓞ ● Ⓓ	1.3k	👤	☰	🗎	ABC	×	Handcrafted
ValueBench(2024)	Ⓞ ● Ⓓ	453	👤	✎	🗎	ABC	×	Psychometric inventories
ComPRed(2025)	Ⓞ ● Ⓓ	15M	👤	✎ ⚖	🗎	ABC	×	Reddit
EValueSteer(2025)	Ⓞ ● Ⓓ	165k	🤖	⚖	🗎	ABC	×	WVS
MoralChoice(2023)	Ⓞ Ⓢ ●	1.8K	👤	☰	🗎	ABC	×	Surveys
MPI(2023)	Ⓞ Ⓢ ●	120	👤	☰	🗎	ABC	×	Human psychometric tests
Prism(2024b)	Ⓞ Ⓢ ●	68k	👤	✎	🗎	ABC	✓	Multi-turn dialogue (from surveys) with LLMs
AlignCure(2025)	Ⓞ Ⓢ ●	27k	👤	⚖	🗎	ABC	×	Human Study (US and Gemany)
MDD(2025)	Ⓞ Ⓢ ●	1.6K	👤	☰ ✎	🗎	ABC	×	Reddit
MIC(2022)	Ⓞ Ⓢ ●	114k	👤	✎	🗎	ABC	×	Reddit
MultiTP(2025)	Ⓞ Ⓢ ●	98k	👤	☰	🗎	🌐	×	Moral Machine Experiment
MID-Space(2024)	Ⓞ Ⓢ ●	42k	👤	⚖	🗎 🖻	ABC	×	Study in Montreal CA
NoveltyBench(2025b)	Ⓞ Ⓢ ●	1.1k	👤	✎	🗎	ABC	×	NB-CURATED+NB-WILDCHAT
NYTBookOpinions(2025)	Ⓞ Ⓢ ●	9k	👤	☰	🗎	ABC	×	New York Times
ConflictQA(2023)	Ⓞ Ⓢ ●	1.2k	🤖	☰	🗎	ABC	×	Existing QA datasets
DebateQA(2024)	Ⓞ Ⓢ ●	2.9K	👤	✎	🗎	ABC	×	Debate forums
CIVICS(2024)	Ⓞ Ⓢ ●	700	👤	✎	🗎	🌐	×	Hand-crafted
NormBank(2023)	Ⓞ Ⓢ ●	155k	🤖👤	✎	🗎	ABC	×	Existing social norms
KorNAT(2024a)	Ⓞ Ⓢ ●	10k	👤	☰	🗎	Kr	×	Korean surveys
ModelSlant(2025)	Ⓞ Ⓢ ●	180k	👤	⚖	🗎	ABC	×	Handcrafted political queries
GlobalOpinionQA(2024)	Ⓞ Ⓢ ●	2.5k	👤	☰	🗎	ABC	×	Surveys (GAS, WVS)
ValuePrism(2024a)	● Ⓢ Ⓓ	218k	🤖👤	✎	🗎	ABC	×	Delphi Platform
CulturePark(2024b)	Ⓞ ● ●	41k	🤖	✎	🗎	🌐	✓	Existing cultural sources
PAPI(2025)	Ⓞ ● ●	320k	👤	✎ ⚖	🗎	ABC	×	Psychometric inventories
CultureBank(2024)	Ⓞ ● ●	23k	🤖	✎	🗎	ABC	×	Tiktok and Reddit
Multifaceted(2024b)	Ⓞ ● ●	197k	🤖	✎ ⚖	🗎	ABC	×	Existing instruction-following datasets
CulturePark(2024b)	Ⓞ ● ●	41k	🤖	✎	🗎	🌐	✓	LLM multi-agent simulations
Fulcra(2024b)	Ⓞ ● ●	20k	🤖👤	✎	🗎	🌐	×	Seeded from Schwartz's Theory
PluriHarms(2026)	Ⓞ ● ●	15k	👤	☰	🗎	ABC	×	ValueKaleido, SafetyAnalyst
OpinionQA(2023)	Ⓞ ● ●	1.5k	👤	☰	🗎	ABC	×	Opinion polls
SafeWorld(2024a)	Ⓞ ● ●	2.5K	👤	✎ ☰	🗎	ABC	×	Human Study
Persona(2025)	Ⓞ ● ●	317k	🤖	⚖	🗎	ABC	×	Synthetic
HelpSteer2(2024e)	Ⓞ ● ●	21k	👤	✎ ⚖	🗎	ABC	×	ShareGPT
DIVE(2025)	Ⓞ ● ●	38k	👤	✎ ⚖	🗎 🖻	ABC	×	Human Study (on T2I images)
LIVS(2025)	Ⓞ ● ●	37k	👤	⚖	🗎 🖻	ABC	×	2 year human study
CulturalPalette(2024)	● ● ●	40k	🤖	✎	🗎	ABC	×	PRISM
CulturalKaleido(2025)	● ● ●	30k	👤	☰ ⚖	🗎	ABC	✓	Existing datasets and hand-crafted
Vital(2025)	● ● ●	18.8K	🤖👤	☰ ✎	🗎	ABC	×	Surveys, Polls

References

2026. Oxford English Dictionary, s.v. "Overton window (n.)".

Jadie Adams, Brian Hu, Emily Veenhuis, David Joy, Bharadwaj Ravichandran, Aaron Bray, Anthony Hoogs, and Arslan Basharat. 2025. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 15–25.

Parand A. Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2024. Being considerate as a pathway towards pluralistic alignment for agentic AI. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Dalia Ali, Aysenur Kocak, Dora Zhao, Allison Koenecke, and Orestis Papakyriakopoulos. 2025. A sociotechnical perspective on aligning ai with pluralistic human values. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.

Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2023. The reasonable effectiveness of diverse evaluation data. *arXiv preprint arXiv:2301.09406*.

 doi:10.20944/preprints202603.1876.v1

Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A system for guiding llms via simulated social ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7580–7617.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. 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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

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

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

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

Wolfgang Bilsky, Michael Janik, and Shalom H Schwartz. 2011. The structural organization of human values-evidence from three rounds of the european social survey (ess). *Journal of cross-cultural psychology*, 42(5):759–776.

Carter Blair, Kate Larson, and Edith Law. 2025. Reflective verbal reward design for pluralistic alignment. In *IJCAI '25*.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. 2025. Lore: Personalizing LLMs via low-rank reward modeling. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, and 1 others. 2024. Towards scalable automated alignment of llms: A survey. *arXiv preprint arXiv:2406.01252*.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154.

Nicholas A Caputo. 2024. Rules, cases, and reasoning: Positivist legal theory as a framework for pluralistic ai alignment. *arXiv preprint arXiv:2410.17271*.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

 doi:10.20944/preprints202603.1876.v1

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6116–6135.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024a. PAL: Pluralistic alignment framework for learning from heterogeneous preferences. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. 2024b. Mallows-DPO: Fine-tune your LLM with preference dispersions. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, and 1 others. 2024c. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024d. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.

Quan Ze Chen, Kevin Feng, Chan Young Park, and Amy X Zhang. 2025. Spica: Retrieving scenarios for pluralistic in-context alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 748–765.

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten Rijke. 2024e. The sifo benchmark: Investigating the sequential instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1691–1706.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219.

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.

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.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2025. Dailydilemmas: Revealing value preferences of LLMs with quandaries of daily life. In *The Thirteenth International Conference on Learning Representations*.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, and 1 others. 2024. Position: social choice should guide ai alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9346–9360.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, and 1 others. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 958–979.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-bench: An over-refusal benchmark for large language models. In *Forty-second International Conference on Machine Learning*.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288.

Esin Durmus, Karina Nguyen, Thomas 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. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit N. Shah, Aadirupa Saha, and Brian D Ziebart. 2025. Imitation beyond expectation using pluralistic stochastic dominance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kevin Feng, Inyoung Cheong, Quan Ze Chen, and Amy X Zhang. 2025. Policy prototyping for LLMs: Pluralistic alignment via interactive and collaborative policymaking. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

 doi:10.20944/preprints202603.1876.v1

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024a. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Jillian Fisher, Ruth E Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret E Roberts, Jennifer Pan, and 1 others. 2025. Political neutrality in ai is impossible-but here is how to approximate it. *arXiv preprint arXiv:2503.05728*.

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

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, and 1 others. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Kshitish Ghate, Andy Liu, Devansh Jain, Taylor Sorensen, Atoosa Kasirzadeh, Aylin Caliskan, Mona T Diab, and Maarten Sap. 2025. Evaluesteer: Measuring reward model steerability towards values and preferences. *arXiv preprint arXiv:2510.06370*.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, and 1 others. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing*.

Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2025. Counterfactual reasoning for steerable pluralistic value alignment of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, and 1 others. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1454.

Daniel Halpern, Evi Micha, Ariel D Procaccia, and Itai Shapira. 2025. Pairwise calibrated rewards for pluralistic alignment. *arXiv preprint arXiv:2506.06298*.

Hadassah Harland, Richard Dazeley, Peter Vamplew, Hashini Senaratne, Bahareh Nakisa, and Francisco Cruz. 2024. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic AI. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Brian Hu, Bill Ray, Alice Leung, Amy Summerville, David Joy, Christopher Funk, and Arslan Basharat. 2024. Language models are alignable decision-makers: Dataset and application to the medical triage domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 213–227.

Chengkai Huang, Junda Wu, Zhouhang Xie, Yu Xia, Rui Wang, Tong Yu, Subrata Mitra, Julian McAuley, and Lina Yao. 2025. Pluralistic off-policy evaluation and alignment. *arXiv preprint arXiv:2509.19333*.

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 *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417.

Saki Imai, Pedram Heydari, Anthony Sicilia, Asteria Kaeberlein, Katherine Atwell, and Malihe Alikhani. 2026. Mixdpo: Modeling preference strength for pluralistic alignment. *arXiv preprint arXiv:2601.06180*.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

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*, pages 1–15.

 doi:10.20944/preprints202603.1876.v1

Krzysztof Janowicz, Zilong Liu, Gengchen Mai, Zhangyu Wang, Ivan Majic, Alexandra Fortacz, Grant McKenzie, and Song Gao. 2025. Whose truth? pluralistic geo-alignment for (agentic) ai. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pages 799–803.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023a. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. 2025. A survey on progress in llm alignment from the perspective of reward design. *arXiv preprint arXiv:2505.02666*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Han Jiang, Dongyao Zhu, Zhihua Wei, Xiaoyuan Yi, Ziang Xiao, and Xing Xie. 2025a. Picaco: Pluralistic in-context value alignment of llms via total correlation optimization. *arXiv preprint arXiv:2507.16679*.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, and 1 others. 2025b. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, 7(1):145–160.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025c. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025d. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. In *The Thirteenth International Conference on Learning Representations*.

Gauri Kambhatla, Sanjana Gautam, Angela Zhang, Alex Liu, Ravi Srinivasan, Junyi Jessy Li, and Matthew Lease. 2025. Improving the distributional alignment of llms using supervision. *Preprint*, arXiv:2507.00439.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. 2025. Are the values of LLMs structurally aligned with humans? a causal perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23147–23161, Vienna, Austria. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024a. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024b. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

 doi:10.20944/preprints202603.1876.v1

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024c. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Toryn Q. Klassen, Parand A. Alamdari, and Sheila A. McIlraith. 2024a. Pluralistic alignment over time. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Toryn Q. Klassen, Parand A. Alamdari, and Sheila A. McIlraith. 2024b. Pluralistic alignment over time. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Kasia Kobalczyk, Claudio Fanconi, Hao Sun, and Mihaela van der Schaar. 2025. Few-shot steerable alignment: Adapting rewards and LLM policies with neural processes. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A Smith, and Hannaneh Hajishirzi. 2025. Compo: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279.

Louis Kwok, Michal Bravansky, and Lewis Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

Thom Lake, Eunsol Choi, and Greg Durrett. 2025. From distributional to overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6794–6814.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024a. Kornat: Llm alignment benchmark for korean social values and common knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11177–11213.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024b. Aligning to thousands of preferences via system message generalization. *Advances in Neural Information Processing Systems*, 37:73783–73829.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach llms to personalize–an approach inspired by writing education. *arXiv preprint arXiv:2308.07968*.

Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. 2024c. Aligning crowd feedback via distributional preference reward modeling. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024d. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Jing-Jing Li, Joel Mire, Eve Fleisig, Valentina Pyatkin, Anne Collins, Maarten Sap, and Sydney Levine. 2026. Pluri-harms: Benchmarking the full spectrum of human judgments on ai harm. *arXiv preprint arXiv:2601.08951*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

 doi:10.20944/preprints202603.1876.v1

Julian Manyika. 2024. *Steerable Alignment with Conditional Multiobjective Preference Optimization*. Ph.D. thesis, Massachusetts Institute of Technology.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Dena Mujtaba, Brian Hu, Anthony Hoogs, and Arslan Basharat. 2025. Aligning machiavellian agents: Behavior steering via test-time policy shaping. *arXiv preprint arXiv:2511.11551*.

Rashid Mushkani, Shravan Nayak, Hugo Berard, Allison Cohen, Shin Koseki, and Hadrien Bertrand. 2025. LIVS: A pluralistic alignment dataset for inclusive public spaces. In *Forty-second International Conference on Machine Learning*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Usman Naseem. 2026. Mechanistic interpretability for large language model alignment: Progress, challenges, and future directions. *Preprints*.

Shravan Nayak, Rashid Mushkani, Hugo Berard, Allison Cohen, Shin Koseki, and Hadrien Bertrand. 2024. MID-space: Aligning diverse communities' needs to inclusive public spaces. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2025. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 16(5).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Kai-Wei Chang, Adina Williams, and Levent Sagun. 2025. The root shapes the fruit: on the persistence of gender-exclusive harms in aligned language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 3094–3105.

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Manish Nagireddy, Pierre Dognin, and Kush R. Varshney. 2024. Value alignment from unstructured text. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025a. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025b. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

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, and 44 others. 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.

 doi:10.20944/preprints202603.1876.v1

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1132–1144.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elinor Poole-Dayan, Jiayi Wu, Jiaxin Pei, and Michiel A. Bakker. 2025. Benchmarking overton pluralism in LLMs. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Haoyi Qiu, Alexander Richard Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of llm web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Diaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, Vinodkumar Prabhakaran, Verena Rieser, and Lora Aroyo. 2025. Whose view of safety? a deep DIVE dataset for pluralistic alignment of text-to-image models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

John Rawls. 1971. A theory of justice.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040.

Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. 2025. The pluralistic moral gap: Understanding judgment and value differences between humans and large language models. *arXiv preprint arXiv:2507.17216*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

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

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*. Survey Certification.

Hua Shen, Ziqiao Ma, Reshmi Ghosh, Tiffany Knearem, Michael Xieyang Liu, Sherry Wu, Andrés Monroy-Hernández, Diyi Yang, Antoine Bosselut, Furong Huang, Tanu Mitra, Joyce Chai, Marti Hearst, Dawn Song, and Yang Li. 2025. Iclr 2025 workshop on bidirectional human-ai alignment. Workshop at the Thirteenth International Conference on Learning Representations (ICLR 2025). Singapore.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. VITAL: A new dataset for benchmarking pluralistic alignment in healthcare. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

 doi:10.20944/preprints202603.1876.v1

Linguistics (Volume 1: Long Papers), pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5).

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*.

Seongho Son, William Bankes, Sangwoong Yoon, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. 2025. Robust multi-objective controlled decoding of large language models. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, and 1 others. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38-18, pages 19937–19947.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024b. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghallah, Liwei Jiang, and Yejin Choi. 2025. Spectrum tuning: Post-training for distributional coverage and in-context steerability. *arXiv preprint arXiv:2510.06084*.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 2025. Pluralllm: pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pages 64–69.

Karolina Stanczak, Nicholas Meade, Mehar Bhatia, Hattie Zhou, Konstantin Böttinger, Jeremy Barnes, Jason Stanley, Jessica Montgomery, Richard Zemel, Nicolas Papernot, Nicolas Chapados, Denis Therien, Timothy P Lillicrap, Ana Marasovic, Sylvie Delacroix, Gillian K Hadfield, and Siva Reddy. 2025. Societal alignment frameworks can improve LLM alignment. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *80th Annual AAPOR Conference*. AAPOR.

Mikhail Terekhov, Moksh Jain, Ruyuan Wan, Maarten Sap, Mitchell Gordon, Dongyeop Kang, Caglar Gulcehre, Amy Zhang, and He He. 2024. Pluralistic alignment workshop at neurips 2024. Workshop at the Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS 2024). Vancouver, Canada.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Peter Vamplew, Conor F. Hayes, Cameron Foale, Richard Dazeley, and Hadassah Harland. 2024. Multi-objective reinforcement learning: A tool for pluralistic alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592.

 doi:10.20944/preprints202603.1876.v1

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024c. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024d. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024e. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and 1 others. 2024f. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Sean J Westwood, Justin Grimmer, and Andrew B Hall. 2025. Measuring perceived slant in large language models through user evaluations. *modelslant.com*.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental limitations of alignment in large language models. In *Forty-first International Conference on Machine Learning*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. 2025a. Distributional alignment for social simulation with LLMs: A prompt mixture modeling approach. In *First Workshop on Social Simulation with LLMs*.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, and 1 others. 2025b. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*.

Nuoya Xiong and Aarti Singh. 2025. Projection optimization: A general framework for multi-objective and multi-group RLHF. In *Forty-second International Conference on Machine Learning*.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. 2024. Debateqa: Evaluating question answering on debatable knowledge. *arXiv preprint arXiv:2408.01419*.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024a. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, and 1 others. 2024b. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024a. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024b. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the*

 doi:10.20944/preprints202603.1876.v1

North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8762–8785.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024a. Safeworld: Geo-diverse safety alignment. *Advances in Neural Information Processing Systems*, 37:128734–128768.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024b. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Sangwoong Yoon, William Bankes, Seongho Son, Anja Petrovic, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. 2024. Group robust best-of-k decoding of language models for pluralistic alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, and 1 others. 2025. Aligning multimodal llm with human preference: A survey. *arXiv preprint arXiv:2503.14504*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. 2024. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, and 1 others. 2025a. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025b. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025c. Exploring chain-of-thought reasoning for steerable pluralistic alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25647–25660.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*.

Shenyan Zheng, Jiayou Zhong, Anudeex Shetty, Heng Ji, Preslav Nakov, and Usman Naseem. 2026. Vispa: Pluralistic alignment via automatic value selection and activation. *arXiv preprint arXiv:2601.12758*.

Jiayou Zhong, Anudeex Shetty, Chao Jia, Xuanrui Lin, and Usman Naseem. 2025. Pluralistic alignment for healthcare: A role-driven framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31308–31331.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773.