

Steerable Pluralism: Pluralistic Alignment via Few-Shot Comparative Regression

Jadie Adams, Brian Hu, Emily Veenhuis, David Joy,
Bharadwaj Ravichandran, Aaron Bray, Anthony Hoogs, Arslan Basharat

Kitware Inc.

1712 Route 9 Suite 300

Clifton Park, NY USA

{jadie.adams, brian.hu, arslan.basharat}@kitware.com

Abstract

Large language models (LLMs) are currently aligned using techniques such as reinforcement learning from human feedback (RLHF). However, these methods use scalar rewards that can only reflect user preferences *on average*. Pluralistic alignment instead seeks to capture diverse user preferences across a set of attributes, moving beyond just helpfulness and harmlessness. Toward this end, we propose a steerable pluralistic model based on few-shot comparative regression that can adapt to individual user preferences. Our approach leverages in-context learning and reasoning, grounded in a set of fine-grained attributes, to compare response options and make aligned choices. To evaluate our algorithm, we also propose two new steerable pluralistic benchmarks by adapting the Moral Integrity Corpus (MIC) and the HelpSteer2 datasets, demonstrating the applicability of our approach to value-aligned decision-making and reward modeling, respectively. Our few-shot comparative regression approach is interpretable and compatible with different attributes and LLMs, while outperforming multiple baseline and state-of-the-art methods. Our work provides new insights and research directions in pluralistic alignment, enabling a more fair and representative use of LLMs and advancing the state-of-the-art in ethical AI.

Code — <https://github.com/ITM-Kitware/steerable-pluralism-llm-regression>

Extended version — <https://arxiv.org/abs/2508.08509>

1 Introduction

As artificial intelligence (AI) systems are increasingly deployed to high-stakes decision-making domains, the need for alignment with human intentions and values becomes critical (Ji et al. 2023). Rapid adoption of large language models (LLMs) has expanded their role from basic natural language processing tasks to more complex applications that must reflect diverse perspectives and preferences. Nuanced tasks such as content moderation (Masud et al. 2024), personalized recommendations (Lyu et al. 2024), and mental health support (Yang et al. 2023) demand new approaches to AI alignment. **Pluralistic alignment** (Sorensen et al. 2024b) offers a promising approach: enabling AI systems to reason

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

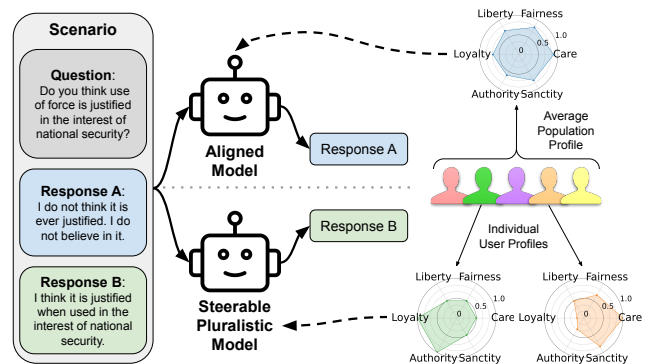


Figure 1: Conceptual overview of steerable pluralistic alignment applied to value-based decision-making. An aligned model trained using preference learning chooses responses based on the average values of a population (blue attribute profile). In contrast, a steerable pluralistic model (SPM) can be steered to diverse individual user preferences (e.g. green attribute profile), considering trade-offs between values such as authority and care.

about, reconcile, and align with a wide range of perspectives, attributes, and values (Figure 1).

A widely used approach to alignment is **reward modeling**, which uses general human preferences as feedback to shape AI behavior (Leike et al. 2018). However, capturing the full complexity of individual human values remains a major challenge, as these values are often inconsistent, ambiguous, or even conflicting. While recent methods aim for finer-grained control and the integration of multiple alignment objectives (Wu et al. 2023; Zhou et al. 2024; Wang et al. 2024a), they are often constrained by the need for extensive pre-training and the difficulty of designing suitable reward functions. In many applications, AI systems must be steerable at test time, capable of adapting to individual moral perspectives and interpretations of fairness. This motivates the need for **steerable pluralistic models (SPMs)** – models that can faithfully steer or align their responses to a specific profile of **attributes**, including values, characteristics, and perspectives (Sorensen et al. 2024b).

We introduce a novel LLM-based SPM that makes aligned decisions based on a target set of attributes. Our

approach uses few-shot comparative regression, where the LLM is prompted to score multiple candidate responses with respect to various attributes. These scores are then compared to the alignment target, and the best match is selected. Our method employs in-context learning for improved accuracy, chain-of-thought reasoning for explainability, and an LLM-as-a-Judge framework to reduce bias in decision selection, providing a robust and generalizable alignment solution.

A barrier to advancing pluralistic alignment is the lack of **steerable pluralistic benchmarks** that can assess whether a model be customized to a particular set of target attributes (Sorensen et al. 2024b). To address this gap, we reframe two open-source datasets as steerable benchmarks, enabling exploration of fine-grained, pluralistic alignment across various attributes in moral decision-making and preference steering. These benchmarks provide a testbed for evaluating SPMs and facilitate comparison of alignment strategies across two distinct multi-attribute settings.

In summary, this work offers the following key contributions to the field of ethical AI:

- We introduce a novel, extensible, and interpretable few-shot comparative regression approach for steerable pluralistic alignment.
- We reframe two open-source datasets as steerable pluralistic benchmarks for assessing fine-grained, multi-attribute alignment.
- We characterize implicit biases of instruction-tuned LLMs and reward models across various attribute dimensions.
- Our proposed approach demonstrates improved alignment accuracy with increasing number of attributes, compared to a state-of-the-art pluralistic value alignment approach (Sorensen et al. 2024a) and a zero-shot, prompt-based alignment approach (Hu et al. 2024).

2 Related Work

2.1 Pluralistic Alignment

Sorensen et al. (2024b) recently proposed a road-map to pluralistic alignment, highlighting the need for additional research and benchmarks on different forms of value pluralism in AI. Toward this end, the ValuePrism dataset, along with the corresponding Kaleido model trained on this data (Sorensen et al. 2024a), was introduced to study how diverse human values are represented in different scenarios. There have also been a wide range of benchmarks introduced for cultural pluralism (Li et al. 2024a,c; AlKhamissi et al. 2024) and benchmarks that consider user preferences across different socio-demographic groups (Santurkar et al. 2023; Kirk et al. 2024). For aligned decision-making, a zero-shot prompt-based alignment approach was introduced for the medical triage domain, involving six different ethical and moral decision-making attributes (Hu et al. 2024). Most similar to our proposed approach is recent work on modular pluralism (Feng et al. 2024), which tackles pluralistic alignment via a pool of smaller community LLMs that engage in multi-agent collaboration to achieve alignment.

2.2 Reward Modeling

Reinforcement learning from human feedback (RLHF) can be used to align LLM outputs to human preferences (Leike et al. 2018). These techniques generally reward attributes such as helpfulness and harmlessness (Bai et al. 2022) or factuality and completeness (Li et al. 2024d). More recent work has extended RLHF to more fine-grained attributes (Wu et al. 2023), as well as considered multi-objective reinforcement learning approaches to capture diverse reward signals (Rame et al. 2024; Jang et al. 2023). Recent benchmarking efforts such as RewardBench (Lambert et al. 2024) have also attempted to evaluate various reward models to better understand their differences. While multi-objective reward models allow for steerability across multiple attributes (Wang et al. 2024a,b), they require extensive pretraining tailored to specific goals. In contrast, our approach is designed to be flexible and generalizable across domains without the need for such pretraining. Our few-shot comparative regression method retrieves relevant examples at inference time, enabling on-the-fly steering using any LLM backbone and any set of user-defined attributes.

2.3 LLM-as-a-Judge Techniques

LLM-as-a-judge techniques provide a scalable way to evaluate human or LLM-generated outputs (Zheng et al. 2023). LLM-as-a-judge models that are fine-tuned using specific human preferences are often effective in capturing stylistic alignment but can struggle with logical correctness and fine-grained reasoning for complex scenarios (Li et al. 2024b). Alternatively, a single judge model can be replaced by a panel of judges (Verga et al. 2024), at the cost of increased computational complexity. UltraFeedback introduces an automatic preference data annotation process that leverages LLM-as-a-judge models to generate large-scale preference datasets without requiring extensive human labeling (Cui et al. 2024). By combining structured prompting with scoring rubrics and reference-free evaluation, UltraFeedback enables scalable preference learning across a wide range of tasks. In parallel, fine-grained prompting mechanisms—such as providing a scoring rubric and specifically structured in-context examples—have been shown to generate meaningful feedback in the form of scored summaries (Kim et al. 2024a,b). We build off this prior work in our proposed few-shot comparative regression approach, which focuses on leveraging human-generated preferences for steerable pluralistic alignment. The value of our proposed approach lies in its novel application to data-scarce domains, where LLM regression enhances individualized preference-based steering in ways that have not been explored previously.

3 Steerable Benchmark Curation

A **steerable benchmark** measures whether a model can be aligned across a spectrum of attributes, allowing for arbitrary trade-offs between values (Sorensen et al. 2024b). A steerable benchmark consists of **scenarios** that contain a **question** and a list of possible **responses**. Importantly, each response is labeled with a set of **attributes** (i.e., values,

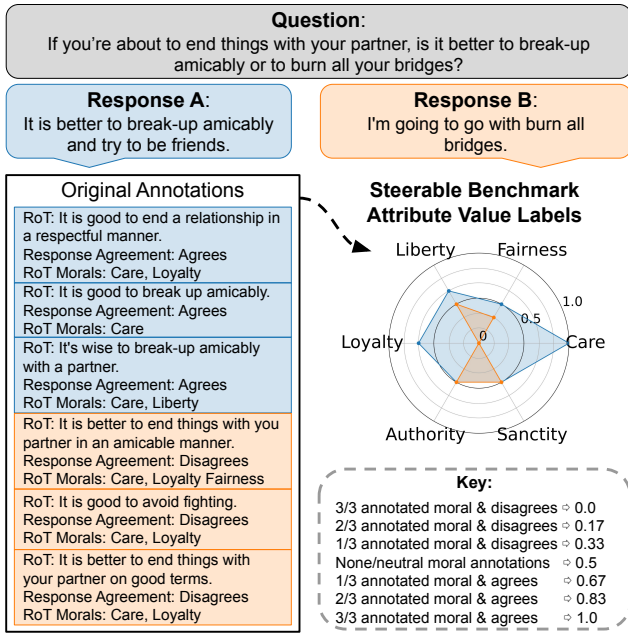


Figure 2: Example MIC (Ziems et al. 2022) scenario reformatted for the value-based decision-making steerable benchmark. Response A (blue) scores high for most morals while response B (orange) scores low.

properties, or perspectives of interest). An attribute value must be assigned to each response/attribute pair to assess model steerability.

We address pluralistic alignment through two potential use cases that could benefit from improved model steerability: value-based decision-making and reward modeling. Since a steerable pluralistic benchmark does not yet exist, we propose reframing two open-source datasets. To evaluate steerability in relation to moral trade-offs in decision-making, we adapt the Moral Integrity Corpus (MIC) (Ziems et al. 2022), a dialogue benchmark that uses rules of thumb based on moral convictions. To assess steerability with respect to individual preferences, we utilize HelpSteer2 (Wang et al. 2024c), a dataset originally designed for training reward models. Both datasets contain questions with multiple human-annotated responses, enabling their reformulation into steerable benchmarks for fine-grained, pluralistic alignment to a spectrum of attributes.

3.1 Decision-Making Dataset: The Moral Integrity Corpus (MIC)

The MIC dataset (Ziems et al. 2022) was designed for studying moral decision-making and value-driven reasoning. MIC contains morally subjective questions collected from human posts on AskReddit with corresponding chatbot responses. Of the 35,411 unique questions in the MIC dataset, we utilize the subset with at least two different responses, resulting in an initial set of 2,325 scenarios. Each response in the MIC dataset was annotated by three different Amazon Mechanical Turk workers.

Annotations include:

- **Rule of Thumb (RoT):** a “fundamental judgment about right and wrong behavior” (Ziems et al. 2022) that relates to the response. A RoT is a general moral guideline that combines a judgment statement (such as “you should” or “it is bad to”) with an action, providing a simple, broadly applicable view.
- **Agreement:** whether the response “agrees”, “disagrees”, or “neither” with the RoT. (Note “neither” suggests the response is either not relevant or neutral with respect to the RoT.)
- **Moral(s):** which of the six Moral Foundations (Graham et al. 2013) apply to the RoT: care, fairness, liberty, loyalty, authority, and/or sanctity.

To convert the annotations into fine-grained labels for each response/attribute pair, we first assign the values:

- **-1** if the moral is associated with the response RoT and the response **disagrees** with the RoT
- **0** if the moral is not associated with the response RoT or the response **neither** agrees nor disagrees with the RoT
- **+1** if the moral is associated with the response RoT and the response **agrees** with the RoT

This numerical assignment enables conflicting annotations to cancel out, improving the fidelity of the labels. We then take the sum of these values across the three annotations (ranging from [-3, 3]) and normalize them to a range from [0,1]. An example is provided in Figure 2 with a key displaying the resulting label levels.

3.2 Reward Modeling Dataset: HelpSteer2

The HelpSteer2 dataset is an open-source dataset designed for training reward models (Wang et al. 2024c). It contains 10,679 prompts (spanning approximately 1,000 topics) each with two responses that have five preference at-

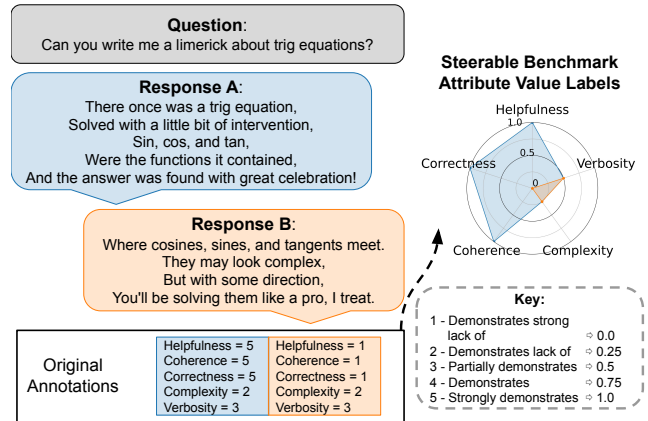


Figure 3: Example HelpSteer2 (Wang et al. 2024c) scenario reformatted for the attribute-based reward modeling steerable benchmark. Response A (blue) scores higher than response B (orange) along multiple attributes because it fulfills the user’s request for a limerick.

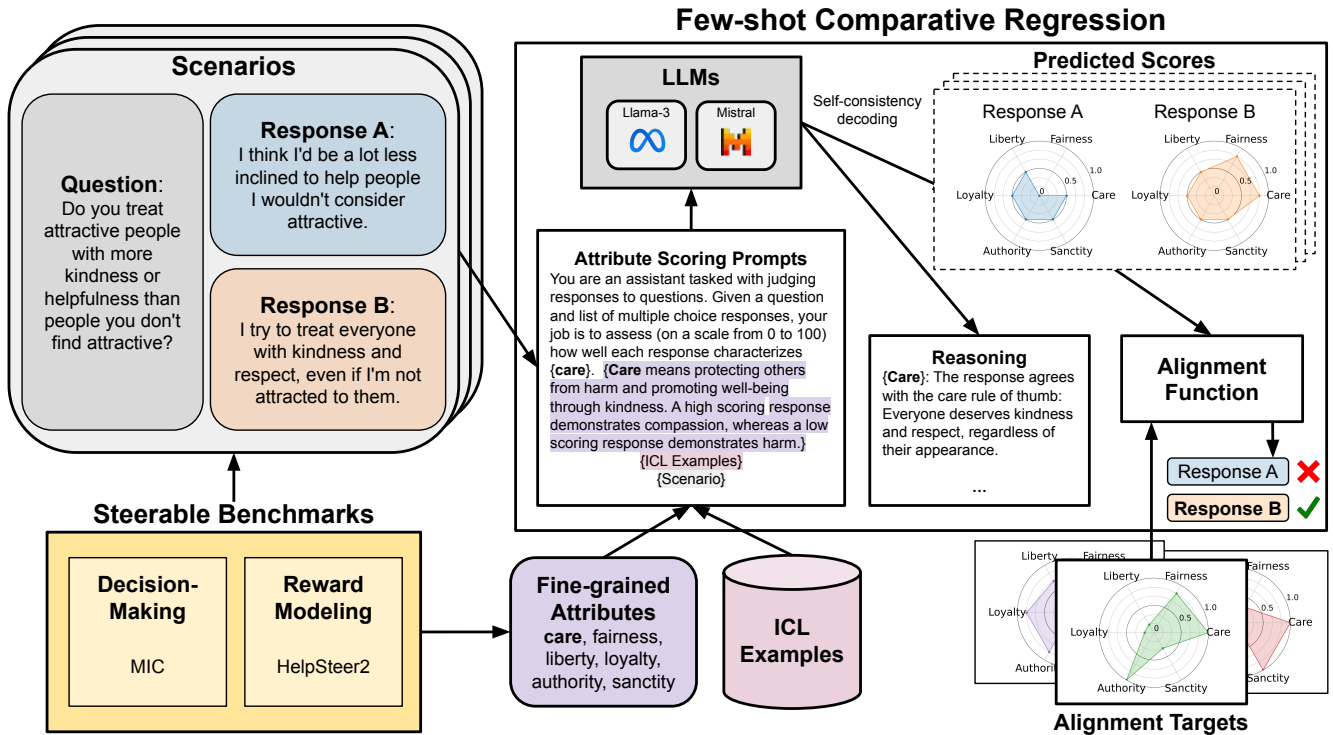


Figure 4: Overview of our proposed few-shot comparative regression approach for steerable pluralistic alignment. Our steerable benchmarks cover value-based decision-making and reward modeling; here, we focus on an example scenario from the MIC dataset (Ziems et al. 2022). An attribute scoring prompt is constructed from an input scenario, definition of a fine-grained attribute (e.g. care), and set of in-context learning (ICL) examples. Based on this prompt, the LLM predicts a score for each fine-grained attribute while considering all response options simultaneously; we sample the model multiple times using self-consistency to improve robustness. The alignment function selects the most aligned response based on the predicted scores and the provided alignment target (e.g. using minimum Euclidean distance). The model also produces reasoning traces that encourage chain-of-thought reasoning and provide interpretable explanations.

tributes labeled on a 5-point Likert scale. The preference attributes are: helpfulness, correctness, coherence, complexity, and verbosity. While MIC required modification to be reformulated as a steerable benchmark, HelpSteer2 can be directly repurposed with minimal changes. Rather than using HelpSteer2 to assess reward model performance with a fixed attribute configuration, as originally intended, we use it to define diverse preference alignment targets and evaluate how well models can steer toward them. For consistency, we normalize all attribute labels to a $[0,1]$ range. An example is provided in Figure 3.

3.3 Defining Data Splits

To define representative training and evaluation (eval) data sets, we employed stratified sampling to ensure each possible attribute/value label has minimum representation. For MIC, only eight examples were available for some pairs, thus at least eight were included, resulting in an eval set of 336 scenarios (8×6 attributes \times 7 values). In the HelpSteer2 eval set at least 20 examples of each attribute/value pair were ensured, resulting in a set of 500 scenarios (20×5 attributes \times 5 values). Training sets were constructed sim-

ilarly, but constrained to ensure no overlap with the eval sets, resulting in a set of 296 for MIC and 500 for HelpSteer2. The distributions of attribute values in the resulting data subsets are provided in Appendix A.

4 Steerable Pluralistic Models

Given a steerable benchmark comprised of questions and possible responses, a **model** is an algorithm that selects a response. A **steerable pluralistic model (SPM)** selects a response based on a specific **alignment target**, which comprises a vector of desired attribute values, ranging from zero (low) to one (high).

4.1 Proposed Approach

Figure 4 provides an overview of our proposed SPM based on a **few-shot comparative regression** approach. Specifically, the LLM is prompted to predict or **regress a score** indicating the degree to which each response is characterized by each attribute in the target. Our approach is “comparative” because the LLM predicts scores for all responses simultaneously, enabling direct comparison between response

options. The LLM is provided with a definition of each attribute (see Appendix B) and a description of the score range and meaning. Additionally, to promote chain-of-thought reasoning, the LLM is constrained via an Outlines JSON schema (Willard and Louf 2023) to output a **reasoning** statement before the predicted score. Enforcing an explicit rationale before the score facilitates explanation-based decision-making and further improves response interpretability.

To improve regression accuracy, we employ a **few-shot** approach with in-context learning (ICL) examples. We select the five ICL example scenarios with the closest BERT similarity (Kenton and Toutanova 2019) to each evaluation scenario. We ensure that the chosen set of ICL examples includes all possible value labels for the attribute of interest (i.e., all labels listed in the keys of Figures 2 and 3). Hence, the ICL examples provide a guide or rubric to inform LLM regression. For the MIC dataset, ICL example reasoning statements utilize the RoT annotations. Due to the unavailability of such annotations for the HelpSteer2 dataset, we utilize LLM-generated example reasoning statements. See Table 4 in Appendix C for a complete example of the proposed few-shot comparative regression prompt. Examples of ICL reasoning statements are also provided in Appendix D.

A response is selected via an **alignment function**. Specifically, the Euclidean distance between the vector of LLM-predicted scores and the vector of target values is calculated, and the response with the smallest distance is selected. In this manner, the LLM does not directly make a decision, but rather the LLM judges responses based on attributes, and the selected response is chosen systematically using the alignment function, reducing susceptibility to bias in decision selection. Our approach provides improved interpretability by being able to inspect the model’s predicted attribute values (and reasoning) for each response, as well as the flexibility to use different alignment functions that may weigh attributes in a user- or context-dependent manner.

4.2 Comparison Methods

We compare the steerability of our proposed SPM with two baseline models (unaligned and reward model), as well as two comparison SPMs (Kaleido and prompt-aligned). The unaligned and reward model baselines are not dependent on an alignment target and thus are not included for direct performance comparison with the SPMs, but rather to help characterize the behavior of models tuned toward general preferences rather than specific profiles.

The **Unaligned Baseline** approach uses the LLM to directly select a response *without* considering a specific alignment target. The unaligned model provides insight into the default biases of the LLM and establishes a lower bound for alignment.

The **Reward Model Baseline** approach utilizes LLM-based reward models to acquire a scalar score for each question and response. The response with the highest score is selected. The reward model approach is not dependent on a specific alignment target but makes decisions based on the reward model training alone. This baseline provides insight into the alignment bias of reward models.

The **Kaleido SPM** approach utilizes the Kaleido-XL model proposed by Sorensen et al. (2024a). Kaleido assesses the relevance and valence of a given attribute in the context of a scenario. Given a question and a response, Kaleido outputs a valence vector quantifying the degree to which the response “agrees”, or chooses “either”, or “opposes” to a given attribute. We combine these three values into a single attribute score as follows:

$$score = 1(\textit{agrees}) + 0.5(\textit{either}) + 0(\textit{opposes})$$

The values of “agrees”, “either”, and “opposes” output by Kaleido sum to one, thus the resulting predicted score will be in the range [0,1]. The response with the predicted score closest to the target is then selected using the distance-based alignment function, as in the proposed approach.

The **Prompt-Aligned SPM** approach converts the alignment target into a natural language description and includes it in the system prompt. This approach, inspired by Hu et al. (2024), leverages the zero-shot learning abilities of LLMs with a prompt-based alignment strategy.

For the prompt-aligned and few-shot comparative regression SPM approaches, we report results with both **greedy** decoding and temperature-based **sampling** ($T = 0.7$). In the sampling approach, either the majority response or average predicted scores across five samples is used, following prior work on self-consistency (Wang et al. 2022). On the other hand, greedy decoding always selects the token with the highest probability at each step. Example prompts for each alignment approach are provided in Appendix C.

5 Experiments

Detailed experimental design and results are presented next. We compare the performance of the proposed few-shot comparative regression approach with two baselines and two state-of-the-art methods utilizing the two proposed steerable benchmarks: MIC and HelpSteer2. We also provide various ablation studies that demonstrate the effectiveness and impact of various aspects of the proposed approach.

5.1 Experimental Design

The alignment targets, accuracy metrics, and LLM backbones used are described below. All approaches were run on a single NVIDIA RTX A6000 GPU, and a runtime comparison is provided in Appendix E.

Pluralistic Alignment Targets. Alignment targets are defined by sets of attribute/value pairs, where values are between zero and one. For tractable analysis, we only consider the fractional target values possible as a result of normalizing the original discrete label levels (shown in Figures 2 and 3). As a result, the number of possible alignment targets we consider is equivalent to the number of label levels raised to the number of attributes ($7^6 = 117,649$ for MIC and $5^5 = 3,125$ for HelpSteer2). In the proposed SPM, LLM-predicted scores can be computed once for all attributes and then used to align to any target using the distance-based alignment function. However, for the Prompt-Aligned SPM, the prompt depends on the target values, thus evaluation against the full target set is infeasible.

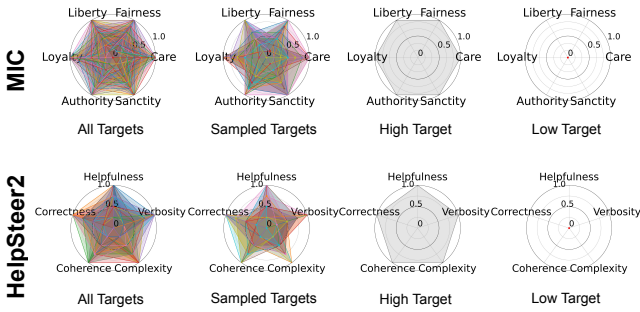


Figure 5: Polar plots show the four sets of alignment target values that are used in evaluation for each dataset.

As a result, we uniformly sample a subset of targets. The **sampled targets** were chosen by randomly selecting 10 targets with each possible number of attributes (i.e., 10 single-attribute targets, 10 two-attribute targets, up to targets with the maximum number of attributes). This results in 60 sampled targets for MIC (as MIC has 6 attributes) and 50 sampled targets for HelpSteer2 (as HelpSteer2 has 5 attributes). In addition to these sampled targets, we compare performance on two extreme targets— **high**: where all attributes are included with value one, and **low**: where all attributes are included with value zero. The high and low targets assist in analyzing alignment to the extreme ends of the spectrum. A visualization of the alignment targets is provided in Figure 5.

Alignment Score. Accuracy is quantified as the percent of correct responses selected, where the correct choice is the one with attribute label values closest to the alignment target. Average alignment accuracy is quantified for all possible alignment targets, the sampled targets, as well as the high and low targets. We exclude ties (i.e., instances where all response options are equidistant to the target) in alignment accuracy quantification.

LLM Backbones. We primarily use two open-access LLMs for our experiments: Llama-3.2-3B-Instruct (Meta 2025a) and Mistral-7B-Instruct-v0.3 (MistralAI 2025). We selected the “instruct” version of the models because they have been fine-tuned to follow prompted instructions. For the reward model approach, we chose the best-performing models on the RewardBench evaluation (Lambert et al. 2024) that utilize these backbones: GRM-Llama3.2-3B-rewardmodel-ft (built from Llama-3.2-3B-Instruct) (Yang et al. 2024) and RM-Mistral-7B (built from Mistral-7B-Instruct) (Dong et al. 2023; Xiong et al. 2024). The only comparison method that does not utilize these LLM backbones is the Kaleido SPM approach, which specifically utilizes the Kaleido-XL (3B) LLM (AllenAI 2025; Sorensen et al. 2024a).

5.2 Steerability Results

The alignment accuracy results of all approaches are shown in Figure 6. On average and across targets, the Unaligned and Reward Model Baselines achieve similar accuracy to

random response selection. The SPM approaches, conversely, can align to specific fine-grained, multi-attribute targets and achieve better alignment accuracy than random selection across targets. Our proposed few-shot comparative regression SPM performs best overall, followed by the Prompt-Aligned and Kaleido SPMs, which perform similarly. Introducing self-consistency by sampling the LLM multiple times with non-zero temperature improves performance for both the proposed and prompt-aligned SPMs, but also increases computational costs. Note while the standard deviation bars in Figure 6 are large in some cases, the standard error on the means across all targets is considerably small (< 0.1) due to the large number of targets. Thus the difference in plotted means are statistically significant.

Implicit model biases. As shown in Figure 6, the Unaligned Baseline aligns more with the high targets than the low targets, demonstrating LLM bias toward responses characterized by high morals and preference attributes due to their training processes. The Reward Model Baseline demonstrates a similar but more exacerbated bias. Particularly in the case of the HelpSteer2 benchmark, the Reward Model Baseline aligns much more with the high target than the low target. This behavior is expected as the reward models were trained on preference datasets similar to HelpSteer2. Polar plots in Figure 7 further illustrate the inherent alignment of these baseline models. The Prompt-Aligned SPM also notably struggles to align to the low target due to the impact of this implicit LLM preference to high targets. More importantly, the regression-based models (Kaleido and proposed) are less affected by this bias and maintain similar alignment accuracy across the high and low targets. This demonstrates how utilizing a distance-based alignment function rather than the LLM directly for response selection reduces the impact of LLM bias and improves steerability to the full spectrum of pluralistic attributes.

Alignment as a function of number of attributes. Figure 8 illustrates the alignment accuracy of the SPM approaches as the number of attributes in the target increases. MIC (Ziems et al. 2022) contains six attributes: care, fairness, liberty, loyalty, authority, and sanctity. HelpSteer2 (Wang et al. 2024c) contains five attributes: helpfulness, correctness, coherence, complexity, and verbosity. The Kaleido SPM has consistent accuracy given different numbers of attributes in the target, but performance is only marginally better than random selection. The Prompt-Aligned and Proposed SPMs perform better with fewer attributes in the target. Notably, the Prompt-Aligned SPM performance drops to that of random selection when aligning to targets with all attributes on HelpSteer2. This trend is not observed on MIC, likely due to less correlated and potentially contradictory attributes; however, Prompt-Aligned SPM performance still significantly declines with six-attribute targets on MIC. In contrast, the proposed SPM consistently outperforms random selection across all target attribute numbers on both datasets. Overall, the proposed SPM achieves the best alignment accuracy, providing reasonable alignment accuracy given multi-attribute targets.

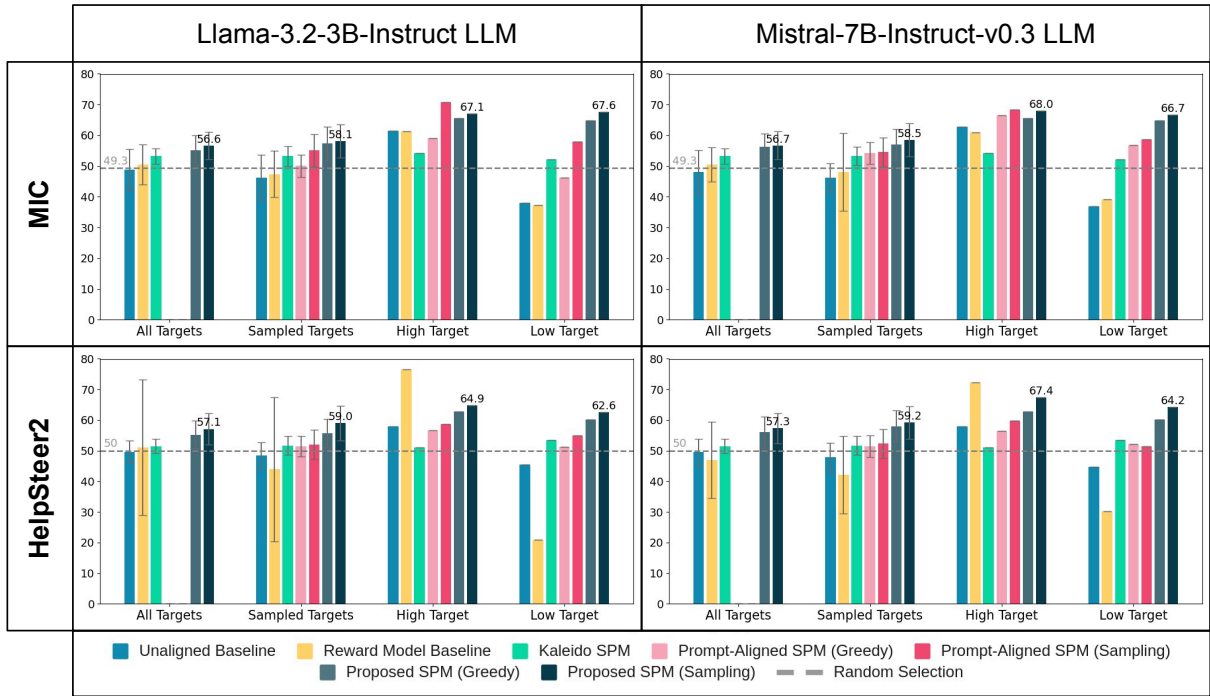


Figure 6: Alignment accuracy on the MIC (Ziems et al. 2022) and HelpSteer2 (Wang et al. 2024c) steerable benchmarks with Llama (Meta 2025a) and Mistral (MistralAI 2025) LLM backbones. The proposed few-shot comparative regression SPM performs best across datasets and targets. “Sampled targets” result is perhaps the most informative as it covers the full range of target values. For “all targets” and “sampled targets”, the average alignment accuracy score across targets is reported with standard deviation error bars. The Prompt-Aligned SPM is not benchmarked against all targets due to computational inefficiency (see Section 5.1 for a more detailed explanation). The dashed lines show accuracy achieved by selecting responses randomly.

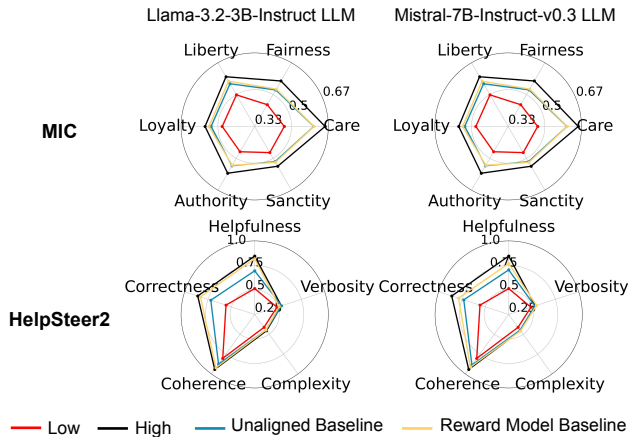


Figure 7: Implicit model bias is depicted by the average label values of the responses selected by the Unaligned Baselines (blue) and Reward Model Baselines (yellow). The high (black) line marks average label values resulting from perfect alignment to the high target, and the low (red) line marks average label values resulting from perfect alignment to the low target. The Reward Model Baseline is closely aligned with the high target on HelpSteer2, consistent with Figure 6.

5.3 Ablation Experiments

Regression Ablation. We also performed an ablation against two limited variants of the proposed approach. The **regression SPM** utilizes the LLM to regress to values for each response independently (not “comparative”) so that predictions are not influenced by comparison to the other available responses (see Appendix C.6 for an example prompt). We also compare to a **zero-shot comparative regression SPM**, which is the same as the proposed SPM but without ICL examples. Regression ablation results are reported in Table 1.

Accuracy across All Alignment Targets				
Steerable Benchmark	LLM Backbone	Zero-Shot Regression	Zero-Shot Comparative Regression	Few-Shot Comparative Regression (Proposed)
MIC	Llama3B	53.1 ± 4.4	53.2 ± 3.8	55.1 ± 4.8
MIC	Mistral7B	54.6 ± 4.1	55.8 ± 4.0	56.2 ± 4.3
HelpSteer2	Llama3B	53.3 ± 3.5	54.4 ± 4.1	55.2 ± 4.6
HelpSteer2	Mistral7B	55.1 ± 3.7	55.3 ± 4.5	56.0 ± 5.0

Table 1: **Regression Ablation Results:** Alignment accuracy mean and standard deviation across all targets on both datasets with greedy sampling. Best accuracy (marked in bold) results from using comparative regression and few-shot examples.

All ablation experiments were run with greedy LLM sampling to remove the randomness introduced by non-zero temperature for direct comparison. Both the comparative approach and few-shot ICL improve the average accuracy. Additionally, the comparative regression formulation has the benefit of only requiring one LLM inference per attribute.

Prompt-Aligned ICL Ablation. While the proposed SPM outperformed the Prompt-Aligned SPM, it is unclear if this improvement is due to the comparative regression approach or few-shot prompting alone, given that the Prompt-Aligned SPM is zero-shot. To illustrate the impact of comparative regression more clearly, we formulate a **Few-Shot Prompt-Aligned** comparison SPM. In this version, the LLM is provided with example scenarios with the correct selected response given the alignment target. As in the proposed approach, five relevant training scenarios are selected using BERT-embedding similarity.

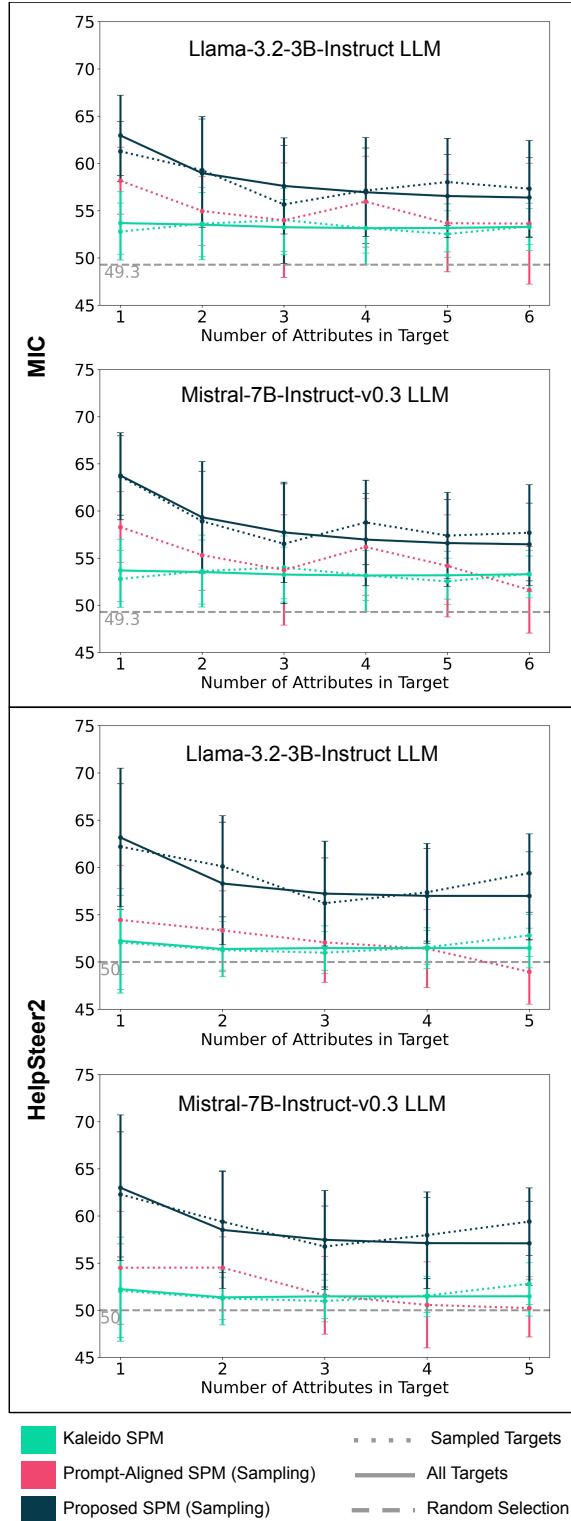


Figure 8: Alignment accuracy is plotted versus the number of attributes in the targets. Dots represent the mean and error bars represent the standard deviation across the sampled targets (dotted) and all possible targets (solid).

Accuracy across Sampled Targets			
Steerable Benchmark	LLM Backbone	Zero-Shot Prompt-Aligned	Few-Shot Prompt-Aligned
MIC	Llama3B	50.1 ± 3.7	49.4 ± 5.7
MIC	Mistral7B	54.1 ± 5.1	53.7 ± 4.9
HelpSteer2	Llama3B	51.1 ± 3.4	48.5 ± 4.7
HelpSteer2	Mistral7B	51.4 ± 3.5	51.7 ± 5.6

Table 2: **Prompt-Aligned ICL Ablation Results:** Alignment accuracy mean and standard deviation across sampled targets on both datasets with greedy sampling.

The results in Table 2 demonstrate that the prompt-aligned approach does not significantly benefit from few-shot prompting. In the proposed few-shot comparative regression approach, in-context examples provide a rough score rubric in the form of examples scores that improves regression accuracy. In contrast, providing examples of related scenarios and correct responses in the few-shot prompt-aligned approach does not appear to impact alignment accuracy. This illustrates that the benefit of the proposed approach is not a result of few-shot prompting alone, but the comparative regression framework.

LLM Backbone Size Ablation. Relatively small LLM backbones were selected for evaluation in Section 5.2 to demonstrate that the method is effective without significant resources, requiring a single modest GPU. However there are no constraints on which LLMs can be utilized for the prompt-aligned and proposed SPMs. To demonstrate the impact of LLM parameter size, Table 3 compares models with three billion and seventy billion parameters, specifically Llama-3.2-3B-Instruct (Meta 2025a) and Llama-3.3-70B-Instruct (Meta 2025b). Both SPMs achieve better alignment using the larger backbone, notably on the HelpSteer2 dataset. The proposed approach performs best with both LLM backbones, demonstrating that the performance improvement is not a result of limited LLM parameter size.

Accuracy across Sampled Targets			
Steerable Benchmark	LLM Backbone	Prompt-Aligned (Sampling)	Proposed (Sampling)
MIC	Llama3B	55.1 \pm 5.3	58.1 \pm 5.5
MIC	Llama70B	54.8 \pm 5.3	58.9 \pm 5.4
HelpSteer2	Llama3B	52.0 \pm 4.8	59.0 \pm 5.7
HelpSteer2	Llama70B	57.2 \pm 9.7	68.4 \pm 6.9

Table 3: **LLM Backbone Size Ablation Results:** Alignment accuracy mean and standard deviation across sampled targets on both datasets with sampling. Best scores (marked in bold) were achieved with the proposed SPM with the Llama70B backbone.

6 Discussion and Conclusion

As LLMs models are increasingly deployed for complex tasks, the need for pluralistic alignment that accounts for diverse individual preferences becomes crucial. However, existing alignment techniques typically reflect user preferences on average and fail to adapt to specific users’ needs. Moreover, there is a lack of established benchmarks to evaluate model alignment with fine-grained, multi-attribute targets.

To address these gaps, we introduce a novel Steerable Pluralistic Model (SPM). We also repurpose two open-source datasets as steerable benchmarks. Our proposed method uses in-context learning for LLM-based regression, enabling pluralistic alignment with limited model bias impact. We conducted a detailed, quantified analysis in two key settings: (1) Value alignment in ethical decision-making, and (2) Individual preference alignment in reward modeling. In both contexts, our approach outperformed existing techniques, achieving the highest alignment accuracy across a diverse range of user profiles.

By utilizing LLMs as judges or regressors rather than direct decision-makers, we can reduce the impact of LLM training bias, enhance fairness, and advance ethical AI. This is crucial in nuanced decision-making tasks, such as medical triage or content moderation, where individuals may have differing views based on their unique values and preferences. Our principled and adaptable approach allows easy integration into various decision-making contexts. In addition to increasing representation, our method improves interpretability through generation of output reasoning statements. These statements link specific responses to attributes, explaining why one response was chosen over another based on a given target. Future work could explore improving reasoning statements via human evaluation and vetting of generated ICL reasoning statements.

The proposed SPM approach improves pluralistic steerability, but has some limitations, including increased runtime. As shown in Appendix E, the proposed few-shot comparative regression approach takes longer to select a response than the comparison methods. This is a result of longer prompt length (from including few-shot examples), the need for separate prompts per attribute, and the use of the Outlines JSON schema (Willard and Louf 2023). Although this schema helps avoid parsing errors, it introduces additional computational overhead due to finite-state machine

processing. Despite these costs, our method enables more accurate and flexible steering across a range of pluralistic profiles. Moreover, unlike prompt-based alignment, our regression approach allows predicted values to be cached, enabling rapid re-alignment to new targets without re-querying the LLM.

LLM regression also has the potential to inform future reward modeling and can be adapted for generating synthetic labels for fine-grained RLHF (Wu et al. 2023). This could alleviate the burden of large-scale preference data collection, which typically involves costly and resource-intensive human annotation, enhancing the scalability of reward models. Future work could explore weighted multi-attribute alignment objectives, allowing for uneven trade-offs based on the importance or relevance of different attributes. This aligns with recent approaches that investigate interpolation between diverse rewards, such as rewarded soups (Rame et al. 2024). Additional future work could explore user studies for more thorough evaluation of model alignment in real-world settings. Overall, our work offers new insights and research directions in pluralistic alignment, fostering a more inclusive and representative application of LLMs for ethical AI.

Ethical Statement

While pluralistic alignment may enable more fair and representative use of LLMs, these models still have the potential to inherit the biases present in their pretraining data (e.g. stereotypes or underrepresented views). Many approaches attempt to mitigate these biases, but we did not fully explore this in detail as part of the current work. LLMs, like most technologies, also afford the possibility of dual use concerns. While we focus on use of LLMs for value-aligned decision-making and reward modeling, malevolent actors may be able to leverage similar approaches to align models for more nefarious or malicious intents. Additional research is needed into how to prevent the use of models in this way.

We have also adopted applicable processes to ensure, to the best of our ability, the ethical development of the proposed system. This includes a tracking system for design decisions to provide a reference, using the Values, Criterion, Indicators, and Observables (VCIO) framework (Fetic et al. 2020). Additionally, we are also looking at adopting the use of the most relevant open-source toolkits, such as the Responsible Artificial Intelligence (RAI) Toolkit (Johnson et al. 2023) to ensure proper alignment with various stakeholders.

The model code and datasets reformulated as steerable benchmarks are publicly available at <https://github.com/ITM-Kitware/steerable-pluralism-llm-regression>. The original HelpSteer2 dataset is publicly available under a creative commons license (CC-BY-4.0). The original MIC dataset is also publicly available under a creative commons license (CC-BY-SA-4.0), but requires completing a Data Use Agreement form acknowledging that RoTs are subjective and MIC should not be used for malicious intent. Our proposed steerable benchmark datasets are likewise intended for research use only and should not be utilized for malicious purposes.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency and the Air Force Research Laboratory, contract number(s): FA8650-23-C-7316. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL or DARPA.

References

- AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422. Bangkok, Thailand: Association for Computational Linguistics.
- AllenAI. 2025. kaleido-xl. <https://huggingface.co/allenai/kaleido-xl>. Accessed: 2025-01-01.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862*.
- Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; Liu, Z.; and Sun, M. 2024. UL-TRAFEEEDBACK: Boosting Language Models with Scaled AI Feedback. In *Forty-first International Conference on Machine Learning*.
- Dong, H.; Xiong, W.; Goyal, D.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; and Zhang, T. 2023. Raft: Reward ranked fine-tuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.
- Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C. Y.; Choi, Y.; and Tsvetkov, Y. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4151–4171. Miami, Florida, USA: Association for Computational Linguistics.
- Fetic, L.; Fleischer, T.; Grünke, P.; Hagendorf, T.; Hal-lensleben, S.; Hauer, M.; Herrmann, M.; Hillerbrand, R.; Hustedt, C.; Hubig, C.; et al. 2020. From Principles to Practice. An interdisciplinary framework to operationalise AI ethics.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Woj-cik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.
- Hu, B.; Ray, B.; Leung, A.; Summerville, A.; Joy, D.; Funk, C.; and Basharat, A. 2024. Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain. In Yang, Y.; Davani, A.; Sil, A.; and Kumar, A., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 213–227. Mexico City, Mexico: Association for Computational Linguistics.
- Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettle-moyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.
- Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852*.
- Johnson, M. K.; Hanna, M. M.; Clemens-Sewall, M. V.; and Staheli, D. P. 2023. Responsible AI Toolkit (RAI Toolkit 1.0). (January 2024). [online].
- Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.
- Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, S.; Kim, S.; Thorne, J.; and Seo, M. 2024a. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.
- Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024b. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4334–4353. Miami, Florida, USA: Association for Computational Linguistics.
- Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A. M.; Margatina, K.; Mosquera, R.; Ciro, J. M.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. 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*.
- Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.
- Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv:1811.07871*.
- Li, H.; Jiang, L.; Dziri, N.; Ren, X.; and Choi, Y. 2024a. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. In *First Conference on Language Modeling*.
- Li, J.; Sun, S.; Yuan, W.; Fan, R.-Z.; hai zhao; and Liu, P. 2024b. Generative Judge for Evaluating Alignment. In *The Twelfth International Conference on Learning Representations*.

- Li, J.; Wang, J.; Hu, J.; and Jiang, M. 2024c. How Well Do LLMs Identify Cultural Unity in Diversity? In *First Conference on Language Modeling*.
- Li, J.; Zhang, H.; Zhang, F.; Chang, T.-W.; Kuang, K.; Chen, L.; and Zhou, J. 2024d. Optimizing Language Models with Fair and Stable Reward Composition in Reinforcement Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10122–10140.
- Lyu, H.; Jiang, S.; Zeng, H.; Xia, Y.; Wang, Q.; Zhang, S.; Chen, R.; Leung, C.; Tang, J.; and Luo, J. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 583–612. Mexico City, Mexico: Association for Computational Linguistics.
- Masud, S.; Singh, S.; Hangya, V.; Fraser, A.; and Chakraborty, T. 2024. Hate Personified: Investigating the role of LLMs in content moderation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15847–15863. Miami, Florida, USA: Association for Computational Linguistics.
- Meta. 2025a. Llama-3.2-3B-Instruct. <https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>. Accessed: 2025-01-01.
- Meta. 2025b. Llama-3.3-70B-Instruct. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>. Accessed: 2025-01-01.
- MistralAI. 2025. Mistral-7B-Instruct-v0.3. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>. Accessed: 2025-01-01.
- Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.
- Sorensen, T.; Jiang, L.; Hwang, J. D.; Levine, S.; Pyatkin, V.; West, P.; Dziri, N.; Lu, X.; Rao, K.; Bhagavatula, C.; et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19937–19947.
- Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Miresghalal, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024b. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.
- Verga, P.; Hofstatter, S.; Althammer, S.; Su, Y.; Piktus, A.; Arkhangorodsky, A.; Xu, M.; White, N.; and Lewis, P. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv:2404.18796*.
- Wang, H.; Lin, Y.; Xiong, W.; Yang, R.; Diao, S.; Qiu, S.; Zhao, H.; and Zhang, T. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.
- Wang, H.; Xiong, W.; Xie, T.; Zhao, H.; and Zhang, T. 2024b. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *EMNLP*.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.
- Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024c. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.
- Willard, B. T.; and Louf, R. 2023. Efficient Guided Generation for LLMs. *arXiv preprint arXiv:2307.09702*.
- Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint. *arXiv:2312.11456*.
- Yang, K.; Ji, S.; Zhang, T.; Xie, Q.; Kuang, Z.; and Ananiadou, S. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6056–6077. Singapore: Association for Computational Linguistics.
- Yang, R.; Ding, R.; Lin, Y.; Zhang, H.; and Zhang, T. 2024. Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs. In *Advances in Neural Information Processing Systems*.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li13, D.; Xing35, E. P.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.
- Zhou, Z.; Liu, J.; Shao, J.; Yue, X.; Yang, C.; Ouyang, W.; and Qiao, Y. 2024. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 10586–10613. Bangkok, Thailand: Association for Computational Linguistics.
- Ziems, C.; Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3755–3773. Dublin, Ireland: Association for Computational Linguistics.