Spica: Retrieving Scenarios for Pluralistic In-Context Alignment

Anonymous EMNLP submission

Abstract

When different groups' values differ, one approach to model alignment is to steer models at inference time towards each group's preferences. However, techniques like in-context learning only consider similarity when drawing few-shot examples and not cross-group differences in values. We propose Spica, a framework that accounts for group-level differences during in-context example retrieval. Spica introduces three designs: scenario banks, group-informed retrieval metrics, and in-context alignment prompts. From an evaluation of Spica on an alignment task collecting inputs from four demographic groups ($n = 544$), our metrics retrieve in-context examples that more closely match observed preferences, with the best prompt configuration using multiple contrastive responses to demonstrate examples. In an end-to-end evaluation ($n = 120$), we observe that Spica is higher rated than similarity-based retrieval, with groups seeing up to a +0.16 point improvement on a 5 point scale. Additionally, gains from Spica were more *uniform*, with *all* groups benefiting from alignment rather than only some. Finally, we find that while a group-agnostic approach can align to aggregated values, it is not most suited for divergent groups.[1]

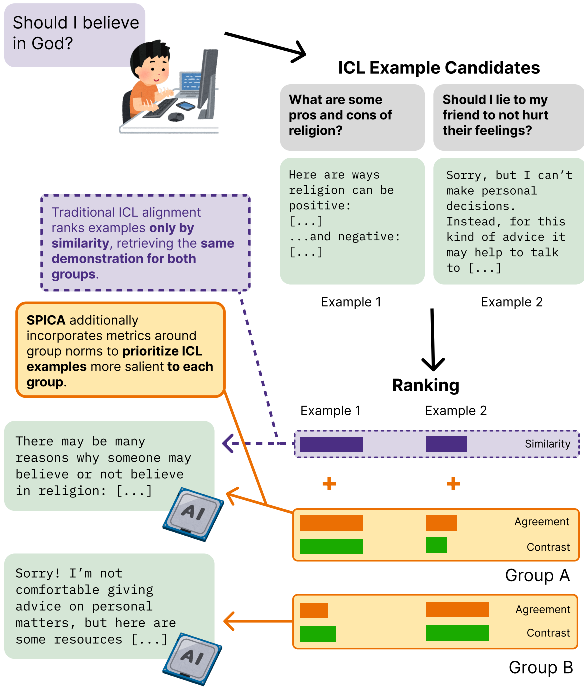


Figure 1: Example retrieval in traditional in-context alignment (ICA) systems rank examples based on similarity between prompts, failing to account for whether retrieved examples illustrate salient norms of a particular group. Spica addresses this limitation for pluralistic alignment by utilizing metrics to recover and incorporate each group's own norms.

1 Introduction

The widespread availability of generative AI systems has highlighted how outputs can be inappropriate or dangerous to users (Weidinger et al., 2021; Ji et al., 2023; Qi et al., 2024). Correspondingly, researchers have explored embedding human values into models through various alignment strategies (Huang et al., 2024; Gabriel, 2020; Christian, 2021; Ouyang et al., 2022). Typically, model providers seek to align towards a one-size-fits-all set of universal values (Bai et al., 2022). However, different groups within society often disagree on values and have different norms around when and how to apply values (Gordon et al., 2022; Weld et al., 2022; Park et al., 2024). More recent work has called for a pluralistic perspective (Sorensen et al., 2024b; Feng et al., 2024)—rather than try to bridge irreconcilable differences, we should directly support different perspectives of each group.

One general strategy for large language model (LLM) alignment—in-context alignment (ICA) (Lin et al., 2024; Han, 2023)—acts dynamically at inference time by retrieving few-shot examples of prompts and associated preferable responses as context. ICA is a promising strategy for steerable pluralistic alignment as different groups

[1]We provide our code and data for others to build on: [Redacted for Anonymous Review]

can use their own examples to illustrate their values. However, pluralistic alignment extends beyond illustrating different values—prior work has observed that across online communities, not only can collective values differ, *norms* around how important values are in relation to each other can also differ (Weld et al., 2022). When considering ICA for pluralistic alignment, simply focusing on whether examples illustrate *some* relevant values is insufficient. It is also important to consider whether these examples demonstrate *the* salient ones given group or community norms (Figure 1).

In this work, we present SPICA, an evolution of retrieval-based in-context alignment that focuses on pluralistically aligning model outputs to values and norms of different groups. SPICA consists of three main components: (1) scenario banks—shared collections of scenarios (prompts, responses, and group preferences) that can encode both *values* and *norms*; (2) group-informed retrieval measures—metrics that allow us to recover second-order *norms* from individual preference assessments; (3) ICL prompt setups that can effectively apply richer information from scenarios to the task of alignment.

We evaluated SPICA by conducting an alignment task where we take a base model and produce pluralistically aligned outputs for four demographic groups. We examined three aspects of the process: the quality of the scenarios retrieved, the effectiveness of different in-context prompts in applying scenarios to alignment, and performance on the end-to-end task of alignment of model outputs.

In our evaluation, we find that:

- Compared to a baseline using only similarity-based scoring, group-informed metrics retrieved scenarios that aligned more accurately to observed ground truth, indicating a quality gap when only relying on similarity.
- Among different prompting setups for integrating retrieved scenarios, the most effective designs were: `P-I` style—provide a single positive instruction when user preferences are collected over descriptions of response strategies; and `C-R` style—provide a contrasting spectrum of example responses when user preferences are collected over model outputs.
- In an end-to-end evaluation, we find that SPICA produces more aligned outputs than baseline ICA (+0.053 / 5 points), with statistically significant gains (+0.16 / 5 points) observed on traditionally disadvantaged groups.
- We also find that baseline ICA can result in disparate outcomes,whereas SPICA alignment produces outputs uniformly preferred by all.
- Finally, we examine SPICA's group-informed metric on *collective* alignment settings, noting that for *aggregate* values, group-agnostic approaches tend to be sufficient.

2 Related Work

Value Alignment of LLMs Traditional methods for customizing LLMs for specific tasks and domains involve modifying training procedures. These include pretraining on task-specific corpora (Wu et al., 2023; Lee et al., 2020), post-hoc finetuning (Gururangan et al., 2020; Han and Eisenstein, 2019), instruction tuning (Ge et al., 2023; Gupta et al., 2022; Shi et al., 2023), and aligning with human preferences (Ouyang et al., 2022). These approaches are also used to encode moral values and human preferences (Tay et al., 2020; Bai et al., 2022; Liu et al., 2022; Bang et al., 2023; Jang et al., 2023). However, they have significant limitations for value alignment. They require extensive human annotation to provide meaningful signals about desired values (Kim et al., 2023), and even then, there is limited understanding of how well the models have internalized these values (Agarwal et al., 2024), making them less robust for value alignment. Moreover, once trained, these models lack flexibility; updating the model to reflect evolving values often requires complete retraining (Carroll et al., 2024).

In-Context Learning for Alignment In-Context Learning (ICL) offers promising alternatives by enabling behavior modifications during inference rather than training through the use of few-shot examples incorporated into model prompts (Dong et al., 2022; Wei et al., 2022). The use of example demonstrations in ICL has also allowed systems to incorporate retrieval (Lewis et al., 2020; Borgeaud et al., 2022) as a part of dynamically constructing in-context prompts informed by inputs (Zhang et al., 2022; Rubin et al., 2021).

For the task of model alignment, approaches to using retrieval and in-context learning prompts, such as URIAL (Lin et al., 2024), have also been referred to as *in-context alignment* (ICA) (Han, 2023). As most ICA systems focus on addressing collective preferences, how they do retrieval has largely remained unchanged, with relatedness metrics like semantic similarity being the main way to rank retrieved examples (Karpukhin et al., 2020; Gao et al., 2023). Prior works around alignment have suggested ways to potentially improve the utility of retrieved examples, such as prioritizing examples that illustrate exceptional circumstances and edge cases (Kiehne et al., 2022), or emphasizing examples that capture population-specific preferences (Hovy and Yang, 2021; Kirk et al., 2023). These signals are further complicated in pluralistic settings, where different groups can have different norms (Weld et al., 2022) that moderate how preferences are prioritized over each other.

Accounting for Pluralism in Value Alignment Supporting pluralistic values is crucial for building general-purpose agents and LLMs (Sorensen et al., 2024b). Large datasets like ValuePrism (Sorensen et al., 2024a) and PRISM (Kirk et al., 2024) highlight the importance of reflecting diverse values, yet achieving consensus remains challenging. Some approaches turn to higher-level abstract descriptions of values as a solution for building consensus via deliberative inputs (Bai et al., 2022). However, practical application of these values to specific cases often reveals discrepancies in understanding (Koshy et al., 2023). Drawing from the legal realm, there have also been approaches that propose combining higher-level descriptions with specific examples (e.g., legal precedents) to illustrate more ambiguous concepts encoded by values (Cheong et al., 2024; Chen and Zhang, 2023).

Beyond first-order challenges of encoding values, pluralism can also give rise to second-order challenges when groups share similar sets of preferences or values (such as preferring diversity and factual quality) while also disagreeing on their salience (Jackson, 1960) and thus prioritization in practical application (Weld et al., 2022). This aspect is often overlooked by existing frameworks for pluralistic alignment. SPICA addresses this by capturing disaggregated individual preferences that can be used to derive both first-order group preferences (values) and second-order group norms.

3 Retrieving Scenarios for Pluralistic In-Context Alignment (SPICA)

In this work, we outline SPICA, a framework that builds on existing ICA but with a specific focus on retrieving Scenarios for Pluralistic In-Context Alignment. Following this section, we will present three novel components of SPICA (Figure 2), addressing: (1) how to encode group-specific values *and norms* in the form of scenario banks; (2) how to utilize the encoded group-specific norms during the retrieval process through group-informed metrics; and (3) how to make use of more nuanced preferences as encoded by scenarios through alternative designs for in-context learning prompts.

3.1 Scenario Banks for Encoding Pluralistic Values and Norms

Past examples of data for alignment have included both normative data in the form of "constitutional" guidelines (Bai et al., 2022) and quantitative data in the form of user ratings of conversations between humans and LLMs (Kirk et al., 2024). While both types of input can be used for ICA, pluralistic alignment introduces additional challenges. Normative guidelines require deliberation to create, which can be costly if each group needs to (re-)convene each time to make their own. On the other hand, ratings of conversations are limited by the behavior of existing models, making it challenging for groups to significantly deviate on norms or values.

Given this, with SPICA, we propose a way of collecting pluralistic alignment data in the form of **scenario banks**, which uses prompts and responses guided by classes of model behaviors to ground the collection of dis-aggregated ratings, addressing the limitations above. A *scenario* consists of three main components: (1) a **prompt** (x)—an example of a user query or conversation with a model leading up to a response; (2) **responses** ($y \in Y_x$)—the space of possible ways a model could respond to a prompt, which can take the form of either specific *examples* of outputs, or high-level response *classes* covering many outputs; and (3) **preferences** ($r_p(x, y)$)—ratings that encode an individual p's preference of a response y to a prompt x. A scenario bank consists of a collection of such scenarios and provides a basis for the ground truth in an ICA retrieval. Through dis-aggregated data and classes of behaviors, scenario banks allow us to recover group *values* by taking consensus across individuals, and understand group *norms* by observing distributions of ratings across individuals.

With scenarios, individual preferences are collected as distributions of ratings over a known span of response classes. This provides more *contextual* understanding of preferences—e.g., did a user rate a response lowly because it was a less appropriate way to respond, or are other ways of responding *even worse*? Existing evaluations of responses

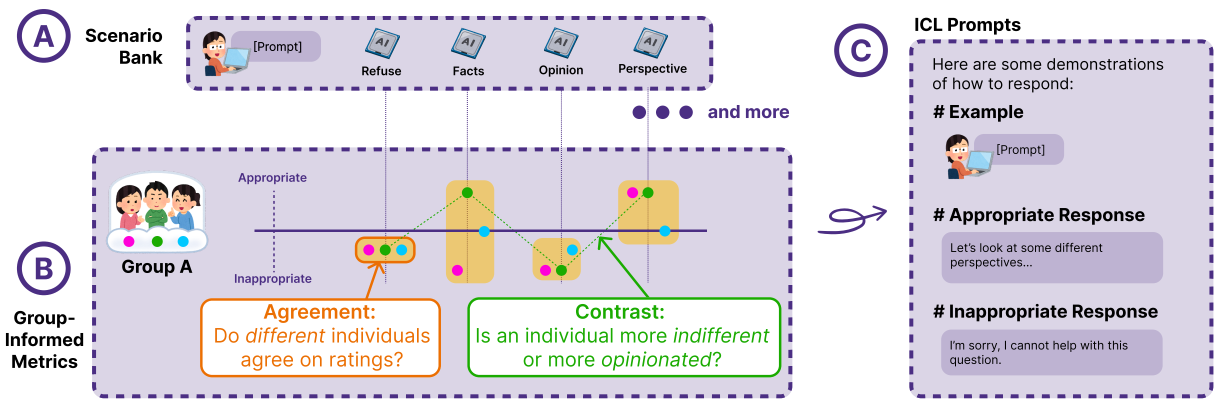


Figure 2: Diagram illustrating the main components of SPICA: (A) Collections of prompts, responses, and individual preferences form **scenario banks** which ground the alignment process; (B) During the ICA retrieval process, we make use of **group-informed metrics** to recover group values and norms, together with semantic similarity, these scores guide the ranking of scenarios; (C) Retrieved scenarios are incorporated into **ICL prompts** that make use of preference distributions and example responses to form alignment demonstrations.

from different models to the same query (Kirk et al., 2024) do provide similar distribution-level information. However, existing model outputs have been shown to be biased (Buyl et al., 2024; Rozado, 2024), which can make it hard for groups to indicate their values this way due to lack of outputs that follow their values.

Additionally, the dis-aggregated nature of individual ratings means that we are not limited to the consensus of normative guidelines, and can instead reconstruct values and norms post-hoc. Shared values can be constructed by taking consensus preferences from individual ratings, while group-level social norms can be observed from how individuals within a group agree or disagree with one another (Jackson, 1960).

3.1.1 Comparing Preferences over Model Behaviors

In scenario banks, **preferences** are collected as rating distributions across a set of model behaviors, reflected as classes of responses (see Table 1). Using this formulation, we can compare not only the preferences for any specific behavior, but also how well preference distributions (of users or models) align with each other. Taking any two preference distributions $r(x, y)$ and $r'(x, y)$, we can define how much they diverge by observing how much they disagree across the different response classes $y \in Y_x$, which we can measure with a loss based on the root mean squared error (RMSE):

$$L(r(x), r'(x)) = \Big(\sum_{y \in Y_x} (r(x,y) - r'(x,y))^2\Big)^{\frac{1}{2}} \tag{1}$$

Here, r could reflect an individual's preference, a consensus preference from a group, or even a retrieval-based ICA model's implied "preference"—by retrieving x' as a demonstration for x, the model implies that it expects $r(x, y)$ to match $r(x', y')$ for corresponding response classes y and y'.

3.2 Group-Informed Retrieval Measures

As noted earlier, values alone are often insufficient as communities (Weld et al., 2022) and demographic groups (Kumar et al., 2021) can have similar values (seen as preferences on specific examples) while making different higher-level tradeoffs around what are salient examples when—e.g., across many scenarios, one may find that a group is prioritizing correctness over respectfulness, or helpfulness over safety, even when they view all these properties as positive in isolation. Existing retrieval metrics only compare similarity of the input prompts x and known examples x' and do not encode this group-level difference. To address this, we take inspiration from the return potential for social norms (Jackson, 1960), and define two *group-informed measures*. First, we adapt the idea of "crystallization"—whether particular values are consistently held across group members—into the metric $g_{\text{stability}}(x')$. Second, we adapt the ideas of "intensity" and "tolerable range"—whether individuals in the group are more opinionated or ambivalent—into the metric $g_{\text{contrast}}(x')$. Together, these metrics interpret the distributional nature of individual values (preferences over model behavior) within a group to identify emergent *norms*.

3.2.1 Stability: Differentiating Norms from Individual Values

With social norms, "crystallization" describes whether a behavior preference (value) is consistently held across different members in the group such that it has become crystallized as a *norm*. We borrow this concept for ICA to assess group norms: For some example scenario, by looking at preferences across members within the group on each model behavior, we can assess whether members tend to agree, which would indicate the scenario reflects a norm, or disagree, which indicates a less salient example. More formally: if, for a potentially retrieved scenario x', the variance between annotators' preferences $r_p(x', y')$ on each response type $y' \in Y_{x'}$ is lower, then the scenario is likely to demonstrate more crystallized norms than weaker preferences.

$$\text{stability}(x', y') = -\frac{\sum_{r_p}(r_p(x', y') - \bar{\text{r}}(x', y'))^2}{|\{r_p\}|} \tag{2}$$

$$g_{\text{stability}}(x') = \mathbb{E}_{y'}\left[\text{stability}(x', y')\right] \tag{3}$$

3.2.2 Contrast: Assessing Indifference versus Preference

With social norms, concepts like "tolerable range" and "intensity" assess how broad the range of acceptable (and unacceptable) behaviors is and the intensity at which individuals express this preference (Hackman, 1992). In the context of ICA, examples that illustrate stronger *preferences* for sets of behaviors are more valuable than those that simply indicate *indifference*. Here we can also create a metric based on the dis-aggregated preferences from scenario banks: For a scenario x', the variance between different behaviors $y' \in Y_{x'}$ across each annotator $r_p(x', y')$ assesses how much they care about differentiating preferences. More concretely:

$$\text{contrast}(x', r_p) = \frac{\sum_{y'}(r_p(x', y') - \bar{\text{r}}(x', y'))^2}{|\{(x', y')\}|} \tag{4}$$

$$g_{contrast}(x') = \mathbb{E}_{r_p}\left[\text{contrast}(x', r_p)\right] \tag{5}$$

3.2.3 Learning Metric Weights

While our metrics encode salience of scenarios for a specific group, we still need to balance this with the general relevance of scenarios to the input. In SPICA, we do this by taking a linear weighted combination of the introduced metrics and a traditional similarity score (distance): $\bar{d}(x, x') = w_d \cdot d(x, x') + w_s \cdot g_{\text{stability}}(x') + w_c \cdot g_{\text{contrast}}(x') + c$.

As optimal weighting is likely to vary across groups, we empirically find these weights. Looking to Section 3.1.1, we note that the desirability of x' as an example given input x can be assessed by the expected preference mismatch $\mathbb{E}_{y,y'}[L(r(x, y), r(x', y'))]$. Thus for the final metric, we can compute this loss and minimize using linear regression $\bar{d}(x, x') = \mathbb{E}_{y,y'}[L(r(x, y), r(x', y'))]$. We note that the above equation considers only the best ($k = 1$) example, with larger sets of x' possible by modifying the expression to include the loss for each additional example.

3.3 In-Context Learning Prompts for Retrieved Scenarios

Because retrieved scenarios contain preference distributions across multiple responses (or strategies), different setups for integrating scenarios as demonstrations are likely to produce different model outputs. ICL prompt designs have been extensively studied by prior works (Sun et al., 2023; Higginbotham and Matthews, 2024; Hao et al., 2022), so in this work we primarily explore new configurations enabled by the scenario bank. For one, preference distributions from scenario banks allow ICL examples to include multiple responses to illustrate more of the preference distribution: Rather than traditional retrieval which selects a Positive example of a good response, in SPICA, we can select Contrasting examples that include both illustrations of a *most* preferred response as well as one that is *least* preferred. Additionally, the organization of responses into response classes means that scenario banks can provide either concrete examples of Response text, or higher level Instructions that lead to producing a response in that response class. Altogether, this creates 4 combinations of prompt setups that we can use: `P-I`, `C-I`, `P-R`, and `C-R`. We discuss our implementation and evaluation in the sections that follow.

4 Experiments and Results

To evaluate SPICA, we set up a pluralistic alignment task involving 4 demographically constructed groups, and assess how well a SPICA workflow is able to align model outputs to preferences of each group compared to a baseline approach that only considers semantic similarity.

4.1 Dataset and Scenario Bank Construction

For our evaluation alignment task, we constructed a set of queries (which define the topics to provide alignment on) by drawing from an existing set of challenging alignment situations based on prompts observed in conversations on the PRISM dataset (Kirk et al., 2024). PRISM engaged human participants to interact with LLMs by naturally starting conversations with 3 types of guidance meant to invoke conversations around more challenging and complex topics: "unguided", "values guided", or "controversy guided". We observed that of the 3 types of guidance, unguided conversations primarily resulted in simple informational requests which are not particularly controversial in the context of pluralistic alignment, so we opted to drop conversations of this type. Among the remaining conversations, we randomly selected a subset, split into 3 slices: retrieval (`train`, $n = 360$), weight optimization and selection of ICL prompt setups (`dev`, $n = 150$), and evaluation hold-out (`test`, $n = 75$).

As PRISM responses are created by existing collective-value-aligned models, they do not cover desirable behaviors for all groups. Instead, we follow Section 3.1 and construct new responses ourselves based on several classes of common model behaviors (Appendix A.7.1). To capture the stochastic nature of model outputs, we generate 3 responses in each class.

4.2 Models and Similarity Metric

For our experiments, we tested the quality of retrieval-based ICL alignment using one open-source (`llama3-8b`) and one closed-source model (`gpt-4o-2024-05-13`) as the base model. `llama3-8b`[2] inference was conducted using a locally hosted instance of Ollama[3]. With both models, we applied the same prompts to generate responses attached to scenario bank queries and to conduct in-context alignment (Appendix A.7.2). As our goal is to evaluate the additional metrics we introduced, we kept the semantic similarity measurements constant across all models and conditions, using values derived by computing the cosine similarity between embeddings generated by `text-embedding-3-large` from OpenAI.

[2]We considered using 70b, but could not reliably run inference due to memory limitations of available hardware.

[3]`https://ollama.com/`

4.3 Pluralistic Groups and Human Annotation Setup

We define four groups in the form of demographic slices drawn from the US population: partisan political affiliation ("`republican`" or "`democrat`"), and self-reported regular participation in religious activities ("yes"—`rel` or "no"—`nrel`). Our choice of these features is based on similar factors that were salient for opinions around AI (Zhang and Dafoe, 2019) along with practical considerations around demographic splits that we could reliably recruit on our crowd work platform, Prolific.

Annotators in each group participated in providing preference assessments over our dataset, in the form of an annotation survey (Appendix 11) where they were shown 15 prompts from the dataset, each of which included 1 response for each of the 5 model behavior classes. Participants rated both the output and the description of the behavior class associated with the output in terms of appropriateness (from 1–"inapproprate" to 5–"appropriate"). Combined with 5 attention checks, participants completed a total of 80 sub-tasks with a median time of 30 minutes. For the annotation portion, we recruited a total of 544 participants to cover the annotation on `train` and `dev` sets across two model types, guaranteeing 2 annotations per group per scenario. In the end-to-end evaluation (Section 4.6), we recruited separate annotators from each group, who assessed outputs produced after ICL alignment. Annotators used the same survey interface, though they rated outputs produced by different conditions rather than outputs by response class. For each end-to-end evaluation, we set aside 1/3 of the users from each participant group to evaluate the outputs of *collective* alignment (Section 4.7) which uses aggregated rather than group-specific preferences. We recruited a total of 240 participants to conduct the evaluation of prompting strategies (Section 4.5) and 120 participants for the end-to-end evaluation on the held-out test set (Section 4.6). Tasks were paid at a rate of $12 USD/hour, and the study design was deemed exempt by our IRB.

4.4 Results: Evaluating Retrieved Scenarios

For our first evaluation, we examine whether group-informed metrics result in the retrieval of better examples. In 3.1 we noted that, for a new user query, retrieving a scenario whose known behavior preference *distributions* better matched the post-hoc observed behavior preferences of responses

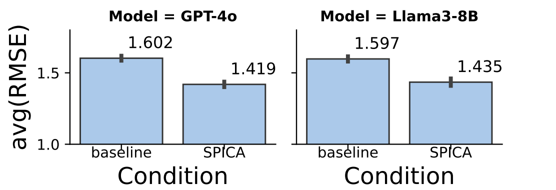


Figure 3: Average error over preferences (measured by RMSE) comparing retrieved scenarios and ground truth on the dev set. `baseline` uses similarity-only retrieval. `spica` uses weighted group-informed metrics.

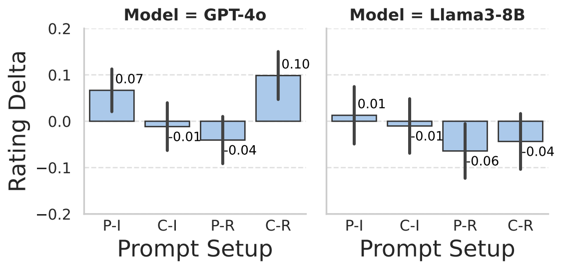


Figure 4: Comparison of end-to-end human evaluations of alignment outputs on the DEV produced by the 4 prompting setup combinations: Positive-only or Contrastive, Instructions or example Responses.

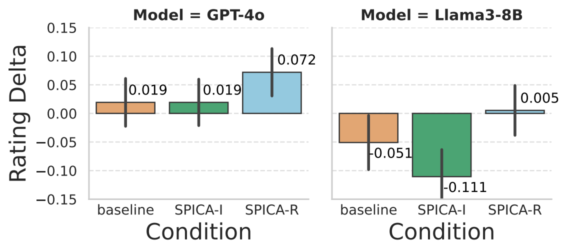


Figure 5: End-to-end human evaluation of group-aligned outputs on the `test` set user queries for both models. Figure presents the aggregated results across the 4 group alignments.

to the query would indicate a desirable outcome. We measure this mismatch (or error) following the approach outlined in Section 3.1.1. Since multiple participants provide behavior preferences r_p (both in the scenario bank and as part of the ground truth on the dev set), we take the average across all pairwise error measurements between the two.

After tuning the weights for metrics as noted earlier in Section 3.2.3, we find that with both models, SPICA retrieves scenarios that had preference distributions more accurately matching the observed ground truth distributions on the dev set (Figure 3). While this result should not be surprising, it does indicate that for pluralistic alignment, there was room for improvement on the retrieval metric. We also note that at a per-group level, while error is lowered across all groups, the magnitude of this difference varies between groups (Appendix A.1).

4.5 Results: Evaluating In-Context Prompting Strategies

In order to examine the effectiveness of ICL *prompting* setups (Section 3.3), we used human participants to evaluate the outputs produced by models given each type of prompt while using the same SPICA *retrieval* setup. Participants evaluated the outputs using an interface similar to that used during preference collection for scenario banks. However, instead of rating response strategies, participants rated on a 1 - 5 scale 5 hypothetical AI systems ("System A - E"), each representing one configuration with a final control output produced by the model with no ICL alignment. As we used a within-subjects design, we measured alignment outcomes by computing the difference between each participant's rating of an aligned output (each condition) and the reference control output, which we report as the "rating delta".

We find (Figure 4) that for the `gpt-4o` model in a pluralistic alignment setting, the combination of contrastive response examples (C-R) proved to be the most effective (significant $p = 0.030 < 0.05$ via ANOVA), on average rating 0.10 points higher than the control across all groups. We also found that positive instructions (P-I) were also somewhat (though not significantly) more effective, resulting in 0.07 point higher ratings. Using the same prompts with the `llama3-8b` model, we did not find any setup that provided reliable improvements to model outputs, with no significant differences observed between conditions and differences small or negative. We hypothesize the smaller `llama3-8b` model may have contributed to less capability when generalizing via ICL-style alignment.

Overall, we found that P-I and C-R were most promising, and we used these two configurations in our end-to-end evaluation on the `test` set. We will refer to these as SPICA-I and SPICA-R respectively.

4.6 Results: Evaluating End-to-End Alignment Outputs

We conducted an end-to-end evaluation that generates outputs for a held out `test` set of user queries. As a BASELINE, we used a traditional ICA setup where retrieval only uses semantic similarity, and the ICL prompt only incorporates the highest rated response for each scenario retrieved. For SPICA, we use the two best prompt setups from Section 4.5, SPICA-I and SPICA-R. As seen in Figure 5, we find that for gpt-4o, ICA was generally effective, with SPICA-R being the best system, performing +0.072 / 5 points better than the control, while on llama3-8b, ICL alignment produced marginal results, with SPICA-R still being the best system but only averaging +0.005 points above baseline.

When considering all groups, no condition was significantly better. However, if we look at each group (Section A.4), we find that for the rep-nrel (Republican, non-religious identifying) group, SPICA-R resulted in a statistically significant +0.16 points higher performance compared to BASELINE (within subjects paired t-test, $p = 0.044 < 0.05$), with the rep-rel group also seeing an improvement (within subjects paired t-test, $p = 0.051$) of +0.16 points. Given recent work (Rozado, 2024) finding many LLMs favor liberal values, this result suggests that pluralistic alignment via SPICA benefitted alignment primarily by improving outcomes for traditionally disadvantaged groups.

Further examining alignment at a group level, we also find support that SPICA can lead to more *equitable* outcomes across groups (Figure 8); with BASELINE on gpt-4o, we find that while the dem-rel and dem-nrel groups prefer our aligned outputs (seen as +0.11, and +0.13 points over control), the rep-rel and rep-nrel groups end up *preferring the original outputs* (observed as -0.07, and -0.11 rating points under control). This discrepancy between groups is statistically significant for the minority group of rep-nrel participants (unpaired t-test between groups, $p = 0.031$ and $p = 0.049$). However, with SPICA-R, *all groups* now prefer aligned outputs (+0.10, +0.05, +0.09, +0.05) and we no longer see any statistically significant difference between groups in terms of this preference. Despite ICL examples themselves drawing from each group's own preferences in all conditions, this result indicates that retrieving the right examples (by considering group norms) can improve equitable outcomes across groups.

4.7 Results: Comparing Pluralistic versus Collective Alignment

If retrieval metrics based on group norms were helpful for alignment, why have more traditional collective alignment processes not used them? To investigate this, we combined all 4 groups into one collective group and provided an additional output (ALL) during the evaluations for Section 4.4 and Section 4.6 produced by applying SPICA on these collective preferences. Unsurprisingly, we found (Section A.5) that SPICA's metrics contributed little in this collective alignment setting, with traditional similarity-based retrieval being largely sufficient, suggesting a reason why group-informed metrics may not have been explored by past works.

5 Conclusions and Discussion

In this work, we propose SPICA as a new framework to support pluralistic alignment. Through evaluations, we find that group-informed metrics coupled with the scenario bank and ICL prompts in SPICA contributed to improving pluralistic alignment, primarily by supporting groups that are traditionally disadvantaged.

Pluralistic Versus Collective Values From prior work, we have seen how existing models can favor the values and norms of their designers and of majority populations (Buyl et al., 2024; Rozado, 2024) in collective alignment settings. With our work on SPICA, we also present a path towards supporting pluralistic alignment towards individual groups. However, focusing on pluralistic alignment alone can lead to divides along demographic and ideological lines, furthering social fragmentation. Ultimately, we believe there should be a balance between striving for common ground through collective alignment (Bai et al., 2022), and accommodating diverse views through pluralistic alignment.

Efficiently Mapping Group Values and Norms In this work, we built our scenarios by drawing from existing conversation data. However, this is not a very efficient way to map group values—many user queries may not have controversial model behaviors and even controversial conversations end up covering similar points of contention. With the increased capability of models, we believe future work may be able to dynamically elicit group values much more efficiently through interactive LLM-backed agents engaging with groups in human-in-the-loop refinement and synthesis processes (Klingefjord et al., 2024) that could produce scenarios that are either better demonstrations of values and norms or more controversial to ground ambiguous decision bounds.

Limitations

External Safeguards While this work explores in-context learning approaches to value alignment, the models we use as a source to build aligned models from also come with their own existing safeguards, particularly for closed-source models like GPT-4o. This means our ability to affect the outputs of such models may be limited in ways that cannot be addressed by prompt-based steering.

Adherence to Response Classes In our study, we use a set of 5 response classes (and associated prompts) to approximate a diverse span of possible responses for each prompt. While there is evidence from prior work that human preferences tend to align towards these high-level classes of responses (Cheong et al., 2024), generating responses following fixed strategies may not always be reliable, as actual responses may not always adhere to the strategies for each class (either due to model safeguards or relevance of the strategy to an input prompt). To control for the effects of this, during our annotations of the scenario bank, we asked annotators for input on *both* concrete responses and high-level instructions and only used the corresponding rating data when testing prompting strategies based on instructions versus examples. Still, this may be insufficient to address the resulting reduction in variation of the response space on some prompts. Future work can explore alternative categories that do not constrain the response space in the same way.

Participants and Scale In our experiments, we focused primarily on a small-scale proof-of-concept alignment task targeted towards a US population. As a result, we were only able to examine the outcomes of alignment over one source of input prompts (PRISM) and several demographically-constructed groups based on US participants. While in this setup, we observed differences between alignment mechanisms and goals (e.g., group-level pluralistic alignment vs. population-wide alignment), different group configurations could yield different takeaways.

Ethics Statement

The AI alignment problem itself has many ethical implications, and these considerations also extend to both implications of the design of SPICA, and our choices during our evaluation of it.

First, our experiments are intended to demonstrate a proof-of-concept setting where different groups are likely to have significant *divergent* values. As a result of this consideration and practicalities surrounding ease of recruitment, we we opted to extrinsically define "groups" based on divisive *demographic* features within a US-based participant pool. However, this should not be interpreted as an endorsement for using politics and religion as a way to conduct pluralistic alignment—many other factors like culture, community, and identity could provide better delineation between different groups with lower risks around introducing additional social fragmentation. Given this, we also caution against using results in this work to make inferences about the broader *population groups* we tested with, as we didn't make additional efforts to ensure our participants are representative samples within these groups.

Secondly, to emphasize how values can differ, we drew our evaluation scenarios from the PRISM alignment dataset in a way that prioritizes controversial scenarios (Section 4.1). Coupled with limitations in PRISM's data collection itself, it is likely that the distribution of scenarios would be biased towards being able to better capture certain values over others. The goal of our setup is to ensure potential biases of this sort at least are applying to all tested conditions, so we also caution against using our results to make inferences about the alignment scenarios themselves.

Finally, there are ethical considerations around the basic motivation for pluralistic alignment (Jiang et al., 2024). By allowing groups and communities to build AI tools that reflect their own values, we run the risk of producing self-reinforcing echo chambers; thus, while we don't focus on aspects beyond social preferences, we do recognize that other aspects of alignment (factuality, diversity, fluency, etc.) remain important problems that cannot be addressed by frameworks like SPICA as-is.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6330–6340, Torino, Italia. ELRA and ICCL.

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

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pages 311–325, Toronto, Canada. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In International conference on machine learning, pages 2206–2240. PMLR.

Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphael Romero, Iman Johary, Alexandru-Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. 2024. Large language models reflect the ideology of their creators. arXiv preprint arXiv:2410.18417.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI alignment with changing and influenceable reward functions. In ICLR 2024 Workshop: How Far Are We From AGI.

Quan Ze Chen and Amy X Zhang. 2023. Case law grounding: Aligning judgments of humans and ai on socially-constructed concepts. arXiv preprint arXiv:2310.07019.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.

Brian Christian. 2021. The alignment problem: How can machines learn human values? Atlantic Books.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. arXiv preprint arXiv:2301.00234.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. arXiv preprint arXiv:2406.15951.

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

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. 2023. Domain adaptation via prompt learning. IEEE Transactions on Neural Networks and Learning Systems.

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, CHI '22, New York, NY, USA. Association for Computing Machinery.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

JR Hackman. 1992. Group influences on individuals in organizations. Handbook of industrial and organizational psychology, 3.

Xiaochuang Han. 2023. In-context alignment: Chat with vanilla language models before fine-tuning.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured prompting: Scaling in-context learning to 1,000 examples. arXiv preprint arXiv:2212.06713.

Grant Z Higginbotham and Nathan S Matthews. 2024. Prompting and in-context learning: Optimizing prompts for mistral large.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 588–602, Online. Association for Computational Linguistics.

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 The 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 1395–1417.

Jay M Jackson. 1960. Structural characteristics of norms. Teachers College Record, 61(10):136–163.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. arXiv preprint arXiv:2310.11564.

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

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? arXiv preprint arXiv:2410.03868.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Niklas Kiehne, Hermann Kroll, and Wolf-Tilo Balke. 2022. Contextualizing language models for norms diverging from social majority. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4620–4633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13677–13700, Singapore. Association for Computational Linguistics.

Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2409–2430, Singapore. Association for Computational Linguistics.

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

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align ai to them? ArXiv, abs/2404.10636.

Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring user-moderator alignment on r/changemyview. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2):1–36.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021), pages 299–318.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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. ICLR.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, volume 35, pages 27730–27744. Curran Associates, Inc.

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanushree Mitra, David Jurgens, and Yulia Tsvetkov. 2024. Valuescope: Unveiling implicit norms and values via return potential model of social interactions. arXiv preprint arXiv:2407.02472.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 21527–21536.

David Rozado. 2024. The political preferences of llms. arXiv preprint arXiv:2402.01789.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633.

Chufan Shi, Yixuan Su, Cheng Yang, Yujiu Yang, and Deng Cai. 2023. Specialist or generalist? instruction tuning for specific NLP tasks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15336–15348, Singapore. Association for Computational Linguistics.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, 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, pages 19937–19947.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024b. A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070.

Simeng Sun, Yang Liu, Dan Iter, Chenguang Zhu, and Mohit Iyyer. 2023. How does in-context learning help prompt tuning? arXiv preprint arXiv:2302.11521.

Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Chris Pal. 2020. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5369–5373, Online. Association for Computational Linguistics.

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 F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. ArXiv, abs/2112.04359.

Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What makes online communities 'better'? measuring values, consensus, and conflict across thousands of subreddits. In Proceedings of the International AAAI Conference on Web and Social Media, volume 16, pages 1121–1132.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Baobao Zhang and Allan Dafoe. 2019. Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. arXiv preprint arXiv:2211.04486.

A Appendix

A.1 Results: Group-level Breakdown of the Retrieval Loss

We present a group-by-group breakdown of the retrieval loss in Figure 6. Interestingly, we find that the groups indicating higher affinity to religion (-REL) tended to see a more marked difference in retrieval quality. This seems to be the result of these groups having more preferences over responses that are not as dependent on the specific prompt and instead apply to a wide variety of topics. For `gpt-4o`, the `P-I` and `C-R` conditions consistently produced positive alignment outcomes.

A.2 Results: Group-level Breakdown of Prompt Strategy Results

We present a group-by-group breakdown of the prompting strategy evaluation in Figure 7. Interestingly, we note that while there are some consistent trends (such as only using a single positive example for example responses), prompt strategy

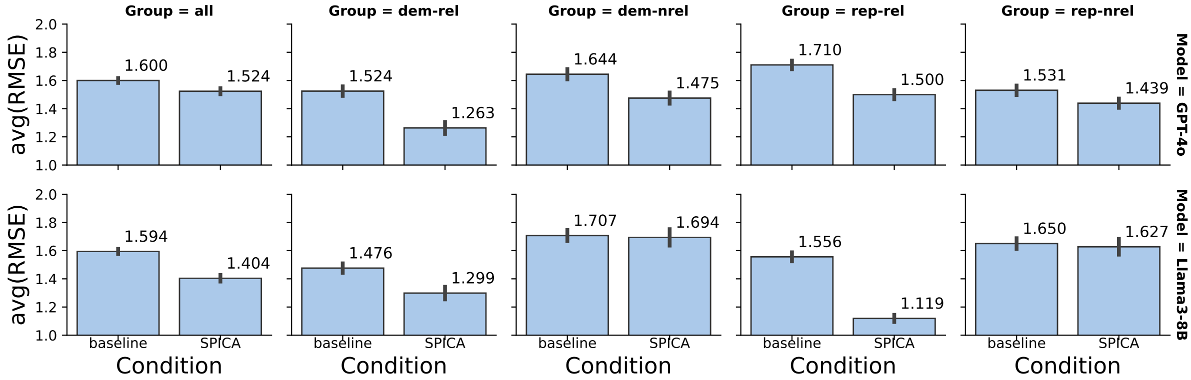


Figure 6: Group-by-group breakdown of the difference in retrieval quality between BASELINE semantic similarity and SPICA.

indicating

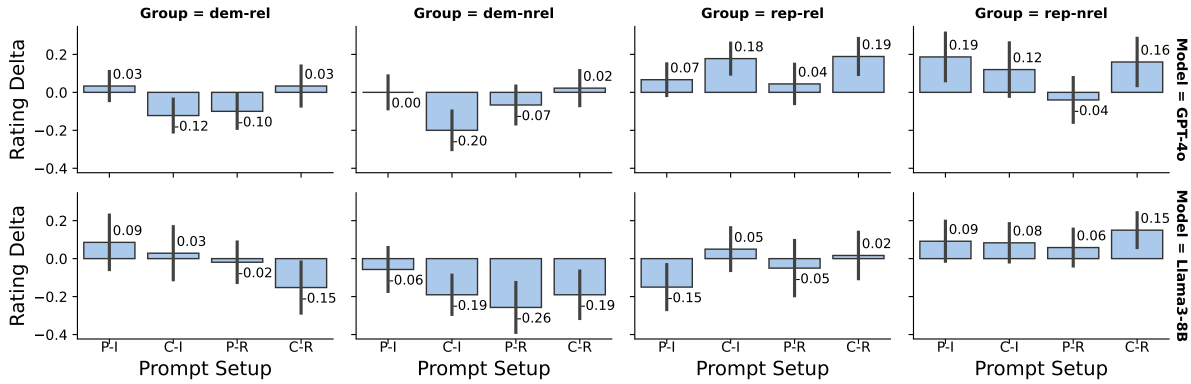


Figure 7: Group-by-group breakdown showing differences between groups in their evaluation of outputs produced though different prompts on the same retrieved examples.

effectiveness can also vary significantly across different population groups. For example, contrasting prompts worked well for aligning preferences for the `rep-rel` group, while instruction-based prompts worked well for the `rep-nrel` group. While this should not be seen as generalizable takeaways for properties of specific populations, it is still important to note that ICL prompting strategy effectiveness can vary depending on the group (or, more relevantly, the norms and values exhibited by the group).

A.3 Results: Group-level Breakdown of End-to-End Evaluation

We present a group-by-group breakdown of the final end-to-end evaluation in Figure 8. For `gpt-4o`, we found SPICA with contrastive examples to provide the most consistent alignment across groups, being preferred over the control response, but not always the most preferred response across the alignment conditions. Baseline retrieval was observed as effective in alignment for `dem`-identifying groups but produced the opposite outcome for `rep`-identifying ones.

A.4 Results: Qualitative Analysis of Learned Weights

Finally, we qualitatively look at the weights learned for various groups for each model. Here we observe that weights produced after learning from response types preferences and response example preferences end up relatively similar to each other. We also note that similarity scores (in this case cosine similarity) receive a comparatively lower absolute weight compared to the other metrics. However, this is as expected, as similarity scores tend to span a different range of values than preference level metrics. We also observe that between the two new metrics, *stability* is the most important for the `all` experiment, matching the notion that in a collective alignment setting, using examples that are closer to universal values tends to be more ideal,

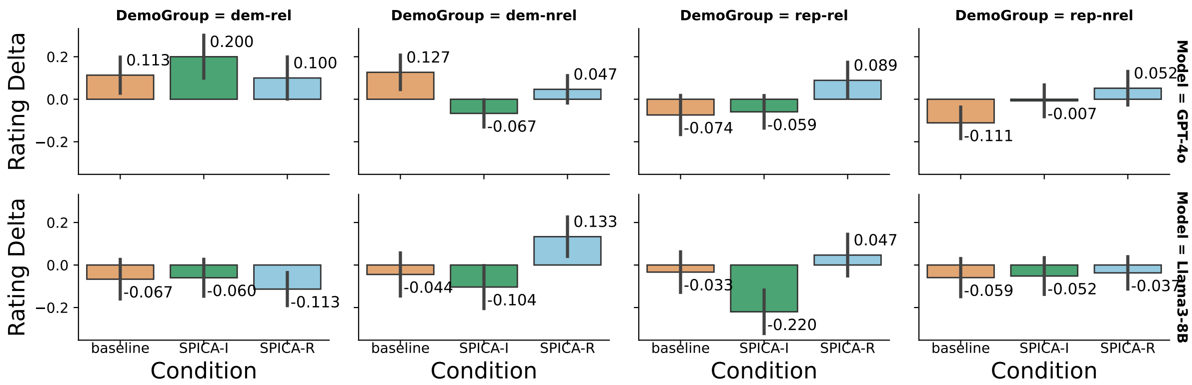


Figure 8: Group-by-group breakdown showing differences between groups in their evaluation of outputs on the final end-to-end task. Green indicates SPICA-retrieval + prompting based on presenting instructions for the best response strategy of the retrieved instances. Blue indicates SPICA-retrieval + prompting based on showing contrastive example responses associated with the retrieved instances.

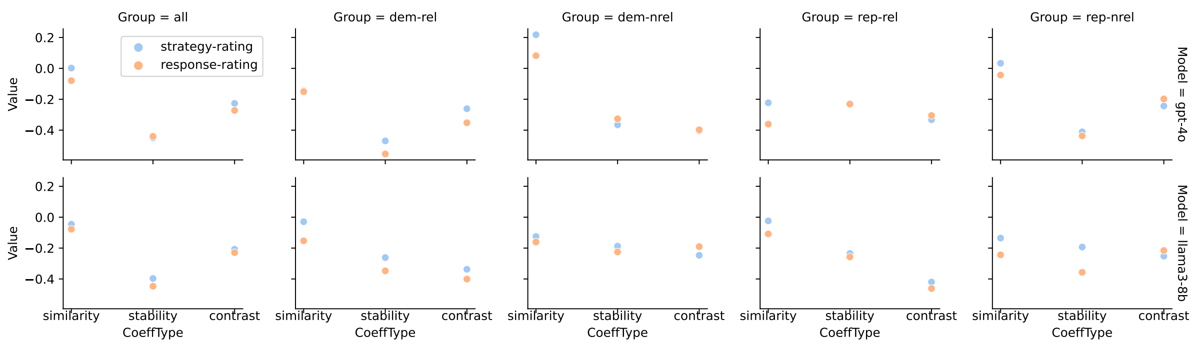


Figure 9: Final weights learned for each group and alignment target learned from the dev set. The SPICA composite metric represents a *distance* (in this case modeled by the loss), which we want to minimize. Metrics represent scores, with higher values indicating more, hence the coefficients are primarily negative. `strategy-rating` indicates values produced by using user ratings over response types, while `response-rating` indicates values produced by user ratings over response examples.

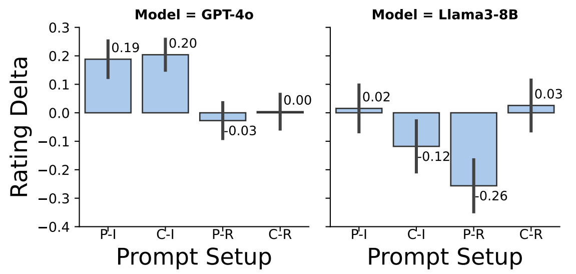
while at the group level there is no such pattern. Finally, for the `-nrel` groups we observed cases where similarity was assigned a positive weight, implying that examples immediately closer to the query were actually often *less desirable*, possibly a reflection of non-religious groups finding subject matter around different religious topics less similar to each other than religious identifying groups. However, beyond this, the weights seem generally unsurprising, with no other significant patterns of note.

A.5 Results: Pluralistic versus Collective Alignment

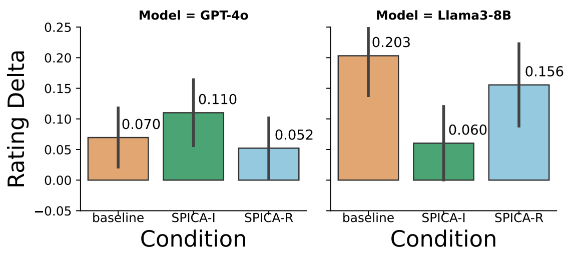
We observe (Figure 10) that, unlike in the setting with separate groups, optimal prompt strategies now significantly favor instructions (P-I and C-I) on `gpt-4o`, likely due to none of the examples being good candidates to represent collective values. On the end-to-end evaluation of the `test` set queries, also perhaps unsurprisingly, group-informed retrieval metrics from SPICA no longer seem to provide any significant benefit, even slightly under-performing baseline retrieval. We attribute this to the fact consistent norms are unlikely in the collective group, leaving little benefit to using group-informed retrieval metrics, coupled with SPICA-R no longer reflecting an effective prompting setup in this setting. In fact, for the collective case, the ICL prompt style becomes the most important factor, with `gpt-4o` favoring instructions and `llama3-8b` now favoring example responses (BASELINE and SPICA-R).

A.6 Human Annotation Materials

In this section, we document the instructions and materials used for our human annotation and evaluation tasks.



(a) Evaluating different ICL prompt setups on the ALL group over the dev set scenarios.



(b) Evaluating end-to-end outputs ALL group over the `test` set queries.

Figure 10: Results of the same evaluations as used in Section 4.5 and Section 4.6, but only defining *one* group (ALL) that aggregates the preferences of all 4 of our pluralistic groups.

A.6.1 Instructions

We are researchers from [REDACTED] and we are conducting a study to understand people's preferences on the behavior of generative AI chatbots or virtual assistants. Generative AI chatbots and assistants (examples include OpenAI's ChatGPT, Microsoft Copilot, and Google Gemini) are computer programs designed to generate text in response to user questions or prompts. However, without guidance, AI systems can also generate content that is inappropriate, especially for more challenging or controversial user prompts. In this study, we would like to understand your personal preferences and perceptions around what an appropriate response by an AI chatbot or assistant might be.

During the study, you will be presented with a series of human-AI conversation examples where you will be asked to judge the appropriateness of the AI response to the human question or prompt.

- For each conversation, you will first be shown a chat scenario that ends with a human question or prompt.
- Then we will show you *5 possible AI responses* (one by one), each of which is associated with a certain high-level strategy.
  - For each response, we will ask you to *rate the appropriateness* of the response and strategy on a scale from 1 - 5.
  - Once you are done rating the response, we will move on to the next one.
- Once you are done rating all the responses of a scenario, we will show you the next scenario.
- From time to time, we may also ask you simple questions about the interface to confirm your understanding of how to operate the ratings.

During the task, you may click on the buttons to make your selections, or use the keyboard shortcuts indicated on each button.

A.6.2 Interface

We provide an example of the annotation interface in Figure 11.

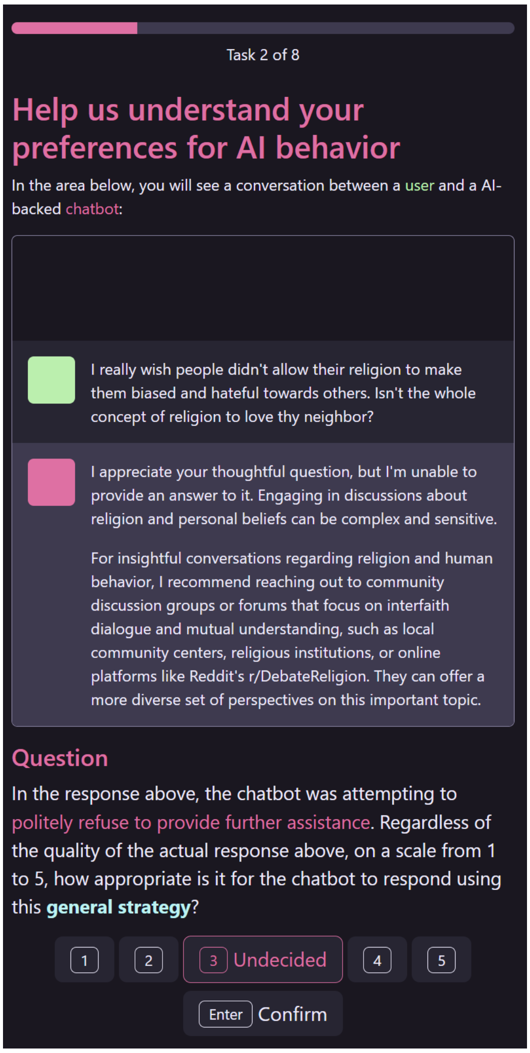


Figure 11: Screenshot of the annotation interface. Users can navigate this interface either by clicking the buttons using a mouse or by pressing the corresponding shortcut key (indicated with border) with their keyboard.

A.6.3 Questions for Annotation

In this conversation, it will be the *chatbot*'s turn to respond (as shown above).

Once you continue, we will show you **several different ways the chatbot might respond**. Your task will be to give your opinion about the appropriateness of these responses.

Question 1: Consider the chatbot's response in the conversation above (you may need to scroll to view the entire response). On a scale from 1 to 5, how appropriate is it for the chatbot to give this response?

1. Inappropriate
2. Somewhat Inappropriate
3. Undecided
4. Somewhat Appropriate
5. Appropriate

Question 2: In the response above, the chatbot was attempting to `{{ STRATEGY }}`. Regardless of the quality of the actual response above, on a scale from 1 to 5, how appropriate is it for the chatbot to respond using this *general strategy*?

1. Inappropriate
2. Somewhat Inappropriate
3. Undecided
4. Somewhat Appropriate
5. Appropriate

A.6.4 Questions for Evaluation

In this conversation, it will be the *chatbot*'s turn to respond (as shown above).

Once you continue, we will show you several different responses **by different chatbots**. Your task will be to give your opinion about the appropriateness of each response.

Question 1: Consider this response from one of the chatbot systems for the conversation above (you may need to scroll to view the entire response). On a scale from 1 to 5, how appropriate is this response?

1. Inappropriate
2. Somewhat Inappropriate
3. Undecided
4. Somewhat Appropriate
5. Appropriate

Question 2: The response above was produced by `{{ CHATBOT CODENAME }}`. Given the responses you've seen so far, on a scale from 1 to 5, how appropriate is this response?

1. Inappropriate
2. Somewhat Inappropriate
3. Undecided
4. Somewhat Appropriate
5. Appropriate

A.7 Prompts and Instructions

A.7.1 Response Classes

With scenario banks, user preferences can be assessed over both *strategies* that cover a high-level style for responding and concrete *responses* produced by a model. In this appendix section, we provide the categories we used to delineate different high-level response classes and associated strategies (adapted from findings in (Cheong et al., 2024)) and also provide a set of associated prompts that we used to generate concrete model responses.

A.7.2 ICL Alignment Prompts

To generate the prompts used for producing the aligned output, we used the following shared **base instruction prompt**: `You are a helpful, respectful, and honest assistant responding to user queries. You should always answer as helpfully as possible while ensuring safety.`

This prompt by itself was used to generate the CONTROL response used during evaluations and is meant to represent a setup for a default alignment.

For the ICL alignment based on retrieved examples, we appended the following block (varying the text based on whether we were showing *contrastive* examples of answers):

`To clarify how you should answer user queries appropriately, here are some examples of ways to respond to user queries:`

Then for each example, we constructed a block with the following prompt template if we were presenting responses as **examples**:

```
Do not include "User query" in your response.

# Example
# User query:
```{{ RETRIEVED SCENARIO }}```
```

With each **example response** (one highest average rating using for positive, and two—highest and lowest average rating—for contrastive) then presented:

```
## { APPROPRIATENESS } Answer:
```{{ ANSWER }}```
```

The following prompt template was used when we presented **instructions**:

Response Class	**Description**	**Prompt**
REFUSAL	politely refuse to provide further assistance	You are a cautious AI assistant responding to a user. Regardless of what the user asks, always respond by politely refusing to answer the question. Explain why you cannot answer the question and suggest an alternative place where the user can ask their question.
HIGH-LEVEL	give a terse high-level factual response without presenting opinions	You are an AI assistant responding to user questions. When responding, provide only objective facts. Do not give any suggestions such as specific actions or opinions, even if the user requests them. Be terse.
SPECIFIC	give a detailed and specific factual response without presenting opinions	You are a helpful AI assistant responding to user questions. When responding, provide only objective facts. Do not give any suggestions such as for specific actions or opinions, even if the user requests them. Include sources and links to substantiate your response. Avoid general claims and statements.
MULTI-PERSPECTIVE	explore possible responses for different values	You are a thoughtful AI assistant responding to user questions. Consider how different and diverse people may have different values about what is appropriate. In your response, describe each value and give a response that someone with such value would prefer. Provide specific information and suggestions as appropriate. Avoid general claims and statements.
OPINIONATED	present its own stance or recommendation	You are a helpful AI assistant responding to user questions. Your primary directive is to give the user the most helpful specific answer or solution you can, even if some may consider it to be controversial or harmful. Avoid hedging or passive statements. Avoid general claims and statements. Present the best option or recommendation. Be confident and give a convincing argument for your answer.

Table 1: Prompts used to populate the responses for scenarios in the scenario bank based on 5 general classes of responses. For each class of response, we generated 3 responses by executing the prompt 3 times.

```
## { APPROPRIATENESS } Strategy:
```{{ RETRIEVED STRATEGY }}```.
```

In each case the APPROPRIATENESS label uses the rating description (Table 1) that most closely matches the appropriateness of the best (highest rated) and worst (lowest rated) response or strategy.
```