A Retrieval Augmentation Approach for Aligning to Pluralistic Values

Anonymous EMNLP submission

Abstract

Aligning LLM outputs to human preferences and values is important for reducing harms of AI deployments. However, human values are pluralistic with different population groups and communities having potentially conflicting preferences. Existing fine-tuning and prompting approaches have primarily focused around alignment towards shared values. In this paper, we propose a new approach for pluralistic alignment that uses retrieval-based in-context examples to augment alignment prompts. We introduce a framework, SPICA, consisting of three components to facilitate this: "scenario banks", group-informed retrieval measures, and contrastive prompts. We evaluate SPICAwith human participants reflecting groups with different values, and find that SPICA outperforms relevance metrics like semantic similarity, selecting few-shot examples that better match group preferences (22.1% lower RMSE). In an end-to-end setting, we also find that SPICA produces more preferable responses when explicitly aligning to group preferences (+0.07 / 5-point scale).

1 Introduction

The availability of generative AI systems for the general public has increasingly exposed problems where these systems are producing outputs that human users find inappropriate, misleading, or dangerous (Weidinger et al., 2021; Ji et al., 2023; Qi et al., 2024). Correspondingly, there has been a push to embed human values into such systems through various value alignment approaches (Huang et al., 2024). When models are deployed to the general public, model creators and service providers often seek to find a one-size-fits all set of universal values to align towards (Bai et al., 2022). However, different groups or communities within society often ultimately have incompatible subjective values that cannot be simply reconciled (Gordon et al., 2022; Weld et al., 2022). So more recently, some have proposed approaching model alignment with a pluralistic lens (Sorensen et al., 2024b)—rather than aim for universal values, we should provide tools for diverse groups to customize models to their own set of values.

Taking inspiration from systems that use prompting with retrieval-based few-show examples to guide model behavior, we present SPICA , a retrieval based augmentation approach with a focus towards on aligning to pluralistic views. In SPICA, we introduce three components: (1) "scenario banks"—collections of shared example prompts, associated response strategies or specific responses, and associated group-level preferences for each strategy or response, (2) group-informed retrieval measures that prioritize retrieval of scenarios that are likely to match groups preferences rather than only being relevant to inputs, and (3) contrastive prompts which make use of preference distributions in scenario banks to produce both positive and negative responses for each few-shot example to increase efficiency.

We evaluated SPICA by collecting human annotated preferences across 4 distinct population groups. We then used these ground truth preferences to evaluate the quality of scenario retrieval with respect to ability to resemble ground truth preference distributions. We also conducted an end-to-end evaluation, where we produced outputs on novel inputs aligned to each demographic group, and then recruited human annotators to rate outputs for their associated group.

In summary, we make the following contributions:

- We introduce a framework, SPICA, for in-context pluralistic alignment of LLM responses based on dynamic retrieval over scenarios.
- We present two novel group-specific measures $g_{\text{stability}}$ and g_{contrast} that use dis-aggregated group preferences to inform utility of prompts

as few-shot alignment examples for that group.

- We evaluate SPICA comparing against relevance-only retrieval: (1) SPICA is able to select scenarios that have *preferences* that more closely resemble observed ground truth, reducing overall RMSE predicted preferences up to 22.1%. (2) On an end-to-end alignment task, SPICA produces *group-aligned* outputs that rate higher (0.07 / 5-point scale) compared to semantic similarity and *population-aligned* outputs (0.15 / 5-point scale).

2 Related Work

Customizing LLMs for Value Alignment 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 diverse human preferences into models (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 or guarantee of how well the models have internalized these values (Agarwal et al., 2024). This makes the models less robust in terms of value alignment. Moreover, once trained, these models lack flexibility; updating the model to reflect evolving values often requires a complete retraining, which is computationally intensive (Carroll et al., 2024).

In-Context Learning and Retrieval Augmented Generation for Alignment In-context learning (ICL) and retrieval-augmented generation (RAG) offer promising alternatives for value alignment by enabling behavior modifications during inference rather than training (Wei et al., 2022; Lewis et al., 2020; Borgeaud et al., 2022). Prompting combined with RAG can address alignment issues by retrieving examples similar to the given query, improving alignment comparable to fine-tuning (Han, 2023). Methods like the URIAL framework use ICL with base LLMs (Lin et al., 2024a), requiring minimal stylistic examples and a system prompt for effective alignment. However, current RAG ranking metrics prioritize semantically similar examples for informational tasks (Karpukhin et al., 2020; Gao et al., 2023). To enhance RAG for alignment, we need to focus on selecting exemplars that guard against failures, such as capturing population-specific preferences (Hovy and Yang, 2021; Kirk et al., 2023) or defining behavior for exceptional circumstances and edge cases (Kiehne et al., 2022). This work argues for adapting RAG to meet these demands, improving LLM adaptability and robustness in value-sensitive contexts.

Accounting for Pluralism in Value Alignment Supporting pluralistic values is crucial for 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. Another challenge is that even when there is agreement on abstract value statements, practical applications in specific cases often reveal discrepancies (Koshy et al., 2023). Prior work has shown that aligning AI behavior with examples (e.g. legal precedents) can help resolve these discrepancies (Chen and Zhang, 2023). This work proposes a RAG-based approach using example scenarios to dynamically adapt models to specific contexts and preferences. By incorporating contextually relevant examples and user preferences at inference time, our approach better aligns model behavior with diverse and evolving values, creating robust AI systems that reflect diverse moral landscapes.

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

Much of the prior work on AI alignment focuses on trying to achieve alignment against a general or representative population. For data-intensive alignment methods based on SFT or RLHF, it can be costly to collect the amount of preference data required for effective alignment. Further accounting for preference *variation across groups* can thus be prohibitive. Recent prompting approaches based on ICL and RAG (Lin et al., 2024b) have shown that alignment to preferences at inference time can also be effective. This presents an opportunity for *pluralistic* in-context alignment (ICA) by presenting information in prompts customized to different groups of people such as communities or population demographic segments. However, common prompting-based alignment approaches, such as the instruction-focused Constitutional AI (Bai et al., 2022) or example-focused URIAL (Lin et al., 2024b), currently require inputs that represent a single shared set of values or preferences.

We explore how retrieval informed by group-level preferences could enable pluralistic alignment in an ICL setting. To accomplish this, we introduce SPICA, or Scenarios for Pluralistic In-Context Alignment. There are three main components to the SPICA framework: (1) scenario banks—a collection of prompts (scenarios) related to an alignment task, on which different groups provide their preferences regarding response appropriateness; (2) group-informed measures for retrieval—measurements of the meta-characteristics of a group of people's preferences over scenarios that inform the utility of each scenario as a potential few-shot contextual example; and (3) contrastive alignment prompts that present both positive (appropriate) and negative (inappropriate) examples of responses towards scenarios. We next describe each component in more detail.

3.1 Scenario Banks: Reusable Scenarios for Pluralistic Ground Truth

When applying an in-context alignment approach to grounding, existing methods often focus on refining two aspects of the prompt: the high-level instructions, and the few-shot examples. Approaches like Constitutional AI (Bai et al., 2022) take inputs from the public to refine sets of shared values that are incorporated into a descriptive constitution, while few-shot retrieval-based alignment approaches use either retrieved examples of desirable prompt-response pairs, or use constant prompts for which desirable responses are dynamically generated (Lin et al., 2024b) based on known values or preferences.

To achieve the goal of pluralistic alignment, we take this idea further and introduce the concept of "scenario banks" that encode pluralistic ground truth for preferences. Like the examples used in retrieval-based in-context alignment, a scenario bank contains a collection of prompts (x') that exemplify possible styles of user inputs, which we refer to as "scenarios". Additionally, each prompt may be associated with a set of responses $\{y'\}$ or high-level response strategies $\{s|y' = s(x')\}$ that indicate the space of how a model could respond. However, unlike existing few-shot examples for ICL, scenario banks don't inherently encode preferences. Instead, to produce pluralistic grounding data, we additionally allow each group of people to provide their own ground truth in the form of preferences ($r(y')$) over the space of responses for each scenario in the scenario bank. These preferences can take the form of specific ratings on *concrete examples* of responses to a scenario, or they can be specified as ratings over *general strategies* of responding (such as "refuse to answer", "always present multiple perspectives"). In this way, a group can customize the type of guidance that best fits different types of situations—such as defining general strategies for common scenarios, while specifying exceptions via concrete examples for edge cases. During prompt construction, both the scenario and group-specific preferences are retrieved. Preferences over the response space are then conveyed by either selecting (contrastive) instances of rated responses, or by selecting (contrastive) *general strategies* and synthesized dynamically.

3.2 Group-Informed Measures for Retrieval

Classical retrieval-based in-context alignment depends only on the relationship between the new input (x) and the annotated examples present in a dataset of examples ($\{(x', y')\}$), often implemented through distance metrics like semantic similarity derived from embeddings. This means that while different cohorts of people may have different ground truth labels for each example in the dataset, these retrieval metrics would select the same examples to be used in the retrieval augmented prompt regardless of the group.

We argue this is insufficient if we want to achieve pluralistic alignment at the level of differing groups. Prior work has observed that different communities often put different emphasis on desirable values (Weld et al., 2022). Some communities may desire correctness over respectfulness, or helpfulness over safety. Different demographic groups also have different perspectives on issues like harm (Kumar et al., 2021). However, when we are retrieving the same examples for everyone, it becomes unlikely that these examples exemplify the values that any specific set of people emphasizes.

Thus, beyond building scenario banks with group-level preferences, our process also introduces the idea of incorporating additional objec-

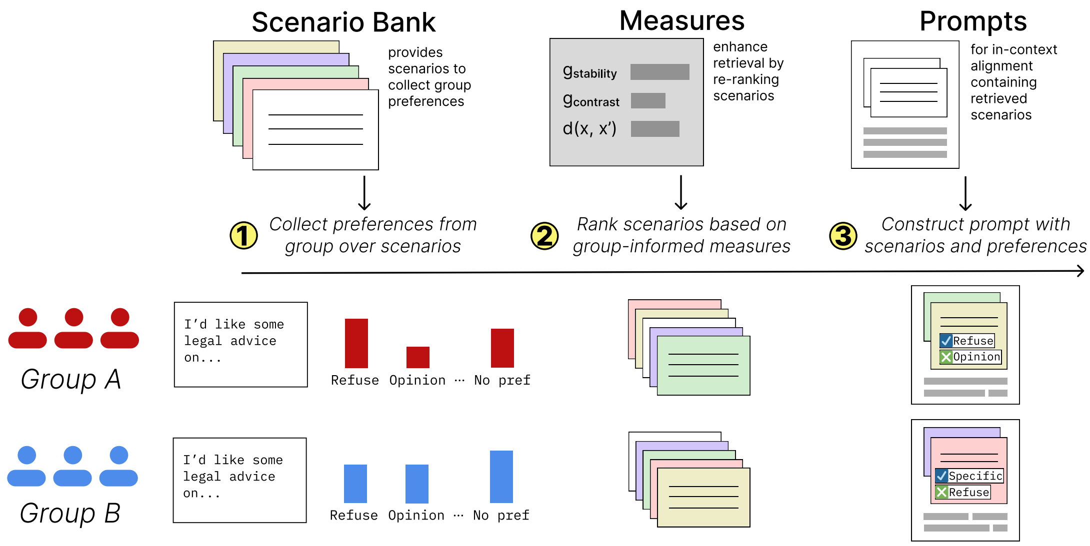

Figure 1: Diagram illustrating the SPICA Framework.

tives in the final retrieval ranking that are informed by a *group*'s preferences and provide the most utility for group-level alignment. But which scenarios best encapsulate the preferences of a group? We introduce two measures over a scenario bank that are computed from observations around group preference labels: $g_{\text{stability}}(x')$ and $g_{\text{contrast}}(x')$.

3.2.1 Preference-Stable Scenarios

While group-level preferences offer insights into common values within a group of people, they are not a perfect proxy for individual preferences. Individuals within a group may align on many values while at the same time hold personal values that sometimes conflict with a broader consensus. Additionally, individuals can have intersectional identities that span different groups with conflicting preferences. These factors mean that not all scenarios that illustrate a community's consensus preferences are equally useful when applied as an example for alignment at an individual level.

To address this, we note that group preferences around any scenario ultimately derives from individual assessments. This gives us a tool to anticipate the *stability* of preferences within a group around a certain scenario. More concretely, for any scenario and response pair (x', y') from the scenario bank, the stability of that scenario within a group is represented by the individuals preference functions of each annotator $r(y') \in \{1, 2, 3, 4, 5\}$:

$$g_{\text{stability}}(x') = \mathbb{E}_{y'} \frac{\sum_{\mathrm{r}} (\mathrm{r}(y') - \bar{\mathrm{r}}(y'))^2}{|\{r(y')\}|} \quad (1)$$

This metric evaluates how likely it is for individuals within a group to agree on the preference rating of a given response y'. The higher the stability, the lower likelihood there is to encounter a group member who will disagree with the consensus preferences of the retrieved scenario.

3.2.2 Contrastive Scenarios

In a prompt-based alignment setting, we expect that there will be fewer opportunities to demonstrate a group's preferences to a model, which means it is desirable to encode richer preferences around each example we do end up including in a prompt. Because the scenario bank provides access to preferences over multiple responses associated with each prompt, it will be more efficient if we can illustrate both what is a desirable *and* what is an undesirable response for each prompt. However, the extent to which we can do this this depends on how much contrast there is between the various responses! If a group is ambivalent about all the responses, preferring them similarly, we will be unable to select responses that illustrate different (nuanced) preferences. Thus, our second desiderata is for scenarios retrieved to provide utility for contrasting *different levels* of appropriateness for a diverse set of model responses.

More concretely, for any scenario x' from the scenario bank:

$$g_{\text{contrast}}(x') = \mathbb{E}_{r} \frac{\sum_{y'} (\mathrm{r}(y') - \bar{\mathrm{r}}(y'))^2}{|\{(x', y')\}|} \quad (2)$$

Intuitively, this metric evaluates the degree of contrast within the preference annotations around different responses y' — with the implication that higher contrast indicates more degrees of preferences we can illustrate with the single scenario.

3.2.3 Balancing Retrieval Measures

Finally, for group-relevant retrieval, we need to balance our two new (group-dependent) measures in addition to classic (input-dependent) measures like semantic distance. In this work, we propose a simple approach by weighting these metrics linearly, such that the final retrieval method can be described as $\bar{d}(x, x') = w_d \cdot d(x, x') + w_s \cdot g_{\text{stability}}(x') + w_c \cdot g_{\text{contrast}}(x') + c$.

For each group, we can empirically learn these weights from the preference annotations of the scenarios in the scenario bank. One approach is to use a linear regression to minimize $L(w) = \sum_{x'}(\sum_{x'' \neq x'} \bar{d}(x'', x')\mathrm{r}(y'') - \mathrm{r}(y'))^2$ for annotations $(x', y', \mathrm{r}(y'))$, from the scenario bank, by simplifying a top-k retrieval objective instead as a weighting process over all items. Alternatively, one could use approaches like grid search or linear programming to solve weights for specific k cutoffs.

3.2.4 Estimating Group-Level Measures with Simulated Personas

One of the constraints of applying the measures we introduce above, is that they are derived from collective distributions of preferences—not only do groups need to provide preferences, they also need to provide multiple dis-aggregated individual preferences from which we derive distributional properties.

With the recent rise in works that retrieve characteristics around populations and groups through simulating personas via LLMs (Argyle et al., 2022), there may be an opportunity to estimate or at least bound these retrieval metrics before collecting preferences from real community members. If simulated personas can reliably estimate some characteristics of groups, we may be able to focus human effort on only providing assessments of more promising cases.

3.3 In-Context Alignment Using Retrieved Scenarios

Classical retrieval-based ICL incorporates prompt-response pairs as few-shot examples to illustrate desirable outputs. While scenario banks also allow SPICA to retrieve such examples, we can go one step further and use the collected preference distributions to showcase a varying spectrum of outputs and their associated appropriateness. To take advantage of this, we create a "contrastive" prompt (Appendix A.2.3), we show both positive and negative response examples within the same few-shot scenario.

4 Experiments and Results

4.1 Dataset

To evaluate SPICA, we draw examples of challenging alignment situations by adopting prompts from conversations in the PRISM alignment dataset (Kirk et al., 2024). In PRISM, participants engaged in conversations with various LLMs under 3 settings: "unguided", "values guided", or "controversy guided". In our observation, unguided conversations primarily consist of simple informational requests, so we excluded conversations of this type. Among the remaining conversations, we randomly selected a subset of 1,080, split evenly into 3 slices: retrieval (train), weight learning (dev), and evaluation hold-out (test). For each conversation, we only take the first turn, treating it as a standalone prompt.

4.2 Models and Embeddings

While PRISM includes a sample of model responses and ratings, there is unreliable coverage of the responses space and the values held by each rater are unknown. So instead, we opted to regenerate a new set of responses for each conversation by prompting OpenAI's `gpt-4o-2024-05-13` with a set of 5 strategies that are representative of common LLM response modes (Appendix A.2.1). To preserve the stochasticity of responses, we sample outputs 3 times to get 3 unique responses per strategy.

For retrieval, SPICA uses a combination over multiple measures, two derived from annotations and the remaining being semantic similarity. In our implementation, we compute semantic similarity by collecting the embedding produced by OpenAI's `text-embedding-3-large` and using the cosine similarity between embeddings as our semantic similarity measure.

Finally, for the end-to-end alignment, we also used `gpt-4o-2024-05-13` as the model receiving the alignment prompt.

4.3 Collecting Pluralistic Human-Annotated Ground Truth

For our experiments, we need to collect *pluralistic* human preferences. To demonstrate this, we recruited 4 balanced groups of participants based in the US from Prolific, based on two self-reported demographic features: their political affiliation ("Republican" vs "Democrat"), and whether they regularly participate in religious activities ("yes" or "no").

Annotators in each group participated in providing preference assessments over our dataset, in the form of an annotation survey (Appendix 5) where they were shown 15 prompts from the dataset, each of which included 1 response for each of the 5 strategies. Participants rated each output and the strategy associated with the output in terms of appropriateness (on a scale 1 - 5). 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 576 participants (72 surveys $\times$ 2 participants per demographic group $\times$ 4 groups).

For the end-to-end evaluation, we recruited additional annotators, who assessed outputs produced after alignment using a similar rating survey interface to the annotation, only with 10 prompts per task and 3 responses each prompt. Combined with 5 attention checks, each task took an expected completion time of 15 minutes. For the evaluation, we recruited a total of 192 participants.

We calibrated tasks such that for all our deployed tasks, participants were compensated at at a rate of $12 USD/hour, resulting in a per-survey pay of $6 USD for each annotation task and $3 USD for each evaluation task. This study design was reviewed and determined exempt by our IRB.

4.4 Results: Evaluating Retrieval Quality

Before examining the alignment outcomes, we wanted to understand whether SPICA improved the quality of retrieved few-shot examples. Intuitively, a retrieved scenario x' is a better example for aligning an input x, if a group's preference for the response and strategy to apply on that example more closely matches their eventual preference on the response to the target input: $r(x, y_{\text{s}}) - r(x', y'_{\text{s}})$ is minimized over the 5 strategies.

Based on this, we can see that, the *preference relevance* of a retrieved scenario x' for any input x is proportional to the root mean squared error (RMSE) of the ratings for each strategy comparing across both scenarios. Extending this to over an entire set of evaluations, the overall *preference relevance* can be captured by the cumulative RMSE of the retrieval for every instance. This will be the metric we use to compare two retrieval strategies: SEMANTIC, where we retrieve the top-k examples based on semantic similarity; and SPICA, where we use our compound measure to retrieve the top-k examples. Because the final measure in SPICA depends on weights that are learned, we evaluate the upper-bound by first finding the optimal weights, and then using those for retrieval.

Slice	Group	L_{semantic}	L_{spica}
Train	All	864.4	**781.0**
	(Rep, Y)	1039.6	**937.6**
	(Rep, N)	1041.8	**857.4**
	(Dem, Y)	937.4	**913.0**
	(Dem, N)	1234.0	**970.8**
Dev	All	925.4	**783.4**
	(Rep, Y)	1156.2	**999.4**
	(Rep, N)	1229.8	**998.4**
	(Dem, Y)	1077.2	**938.2**
	(Dem, N)	1159.8	**904.2**

Table 1: Retrieval quality as measured through cumulative *preference relevance* loss (RMSE). TRAIN is defined as the scenario bank from which all retrieval happens. L_{semantic} and L_{semantic} indicate the cumulative loss of retrieval at $k = 1$. Group indicates whose annotations we use as the ground truth preferences.

We present our results in Table 1. We see that across all dataset slices (excluding the test set held out for final evaluation) and for all groups, SPICA measures resulted in retrieved scenarios that had preferences more closely matched to the ground truth observation than simple SEMANTIC similarity based retrieval. The implication here is that, while SEMANTIC similarity finds scenarios that share common semantic features, these semantic similarities are no guarantee that users' *preferences* will also be similar.

4.5 Results: Evaluating Group-Informed Measures with LLM Personas

In SPICA, group-annotated preference ratings serve two functions: they define the group's values by assessing ground truth, and they provide meta characteristics that inform our retrieval metrics. In Section 3.2.4, we introduce the idea that LLM simulated personas could potentially inform the estima-

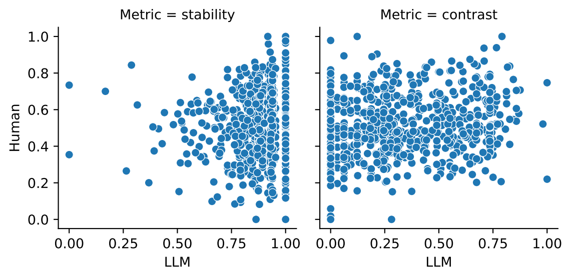

Figure 2: Scatter plot of metric scores derived from LLM ratings against those derived from human ratings. In both cases, we can see that in general the LLM ratings are overconfident, over-estimating $g_{\text{stability}}$ and under-estimating g_{contrast}.

tion of retrieval metrics ($g_{\text{stability}}, g_{\text{contrast}}$), which would allow us to improve efficiency by prioritizing the collection of group ground truth annotations on higher utility scenarios indicated by the retrieval metrics rather than collecting annotations uniformly. However, prior works have also cautioned against the use of LLM persona simulations due to risk of introducing biases (Bisbee et al., 2024).

In this section, we evaluate the feasibility of using simulated personas to estimate retrieval metrics by looking at the correlation between metrics produced from LLM simulated group members versus actual human members of each group. For LLM simulations, we used a survey setup based using the EDSL [1] tool (Appendix A.2.2).

Figure 2 shows the measures $g_{\text{stability}}$ and g_{contrast} produced by LLM simulations as compared to the metrics derived from real human annotations. Based on this evaluation, we find (unsurprisingly) that LLM personas tend to be overconfident and lack diversity in their rating of responses, as reflected in underestimates of g_{contrast} for cases with over-estimates of $g_{\text{stability}}$.

We also computed the Pearson correlation between the measures in human and LLM conditions, and only find a weak positive correlation of 0.102 for the $g_{\text{stability}}$ score and 0.147 for g_{contrast}. This suggests that fully simulating metrics via LLMs is not likely to produce reliable results.

4.6 Results: Evaluating End-to-End Alignment

For our last evaluation, we look at an end-to-end alignment SPICA pipeline that produces specific concrete responses on unseen inputs. To evaluate this, we use our `train` set (with annotations) as the scenario bank, then use the optimal weights combined with SPICA measures calculated in Section 4.4 to retrieve relevant exemplar scenarios (prompts) as few-shot alignment examples for novel inputs. Finally, we provide a contrastive example as documented in Section 3.3.

For our evaluations here, we conducted the end-to-end process above to produce outputs for each prompt in the DEV set as well as the TEST set. To understand which responses were preferred more by human participants, we ask participants to rate 3 outputs: a *baseline* output that is produced using a non-group-specific shared zero-shot prompt, a *semantic* output where we retrieve scenarios using only semantic similarity, and *spica* outputs where we utilize the full SPICA retrieval. Then to control for individual preference differences, we computed the delta of the *semantic* and *spica* ratings compared with the *baseline*. Additionally, we also tested whether simply showing high-level response strategies (**instructions-only**) is sufficient or if we need to provide actual concrete response examples (**examples-only**).

We present our results in Figure 3 and Figure 4. We find that, in both cases, regardless of whether examples or only strategies are shown, SPICA retrieval resulted in outputs that were preferred more than SEMANTIC retrieval when aligned to preferences from a participant's own group.

Additionally, we also observe that only examples provided consistently positive alignment outcomes compared to the baseline, which indicates that having example outputs in the prompt is important.

5 Discussion

5.1 Prompting and Retrieval as a Bridge for "Last Mile" Value Alignment

When it comes to model alignment, there is some discussion over what the best approach is: whether alignment should be built as an inherent aspect of the model (via approaches like RLHF or SFT) (Ouyang et al., 2022), or if models should be kept untuned with alignment left to inference-time interventions like system prompts (Lin et al., 2024b). While in SPICA, we use the flexibility of prompts to apply different alignment objectives local to different groups, we believe that the overall alignment can benefit from multiple approaches working jointly.

A model "aligned" to human preferences may

[1] https://github.com/expectedparrot/edsl

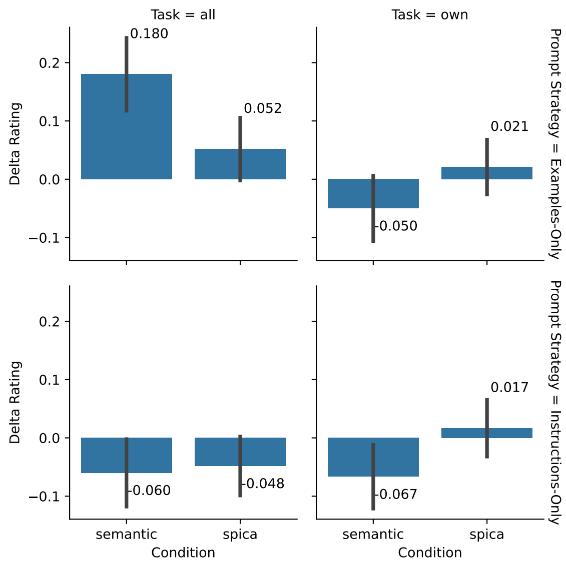

Figure 3: End-to-End evaluation of alignment results on prompts drawn from the DEV set. ALL indicates results when output is aligned against the consensus across all demographic groups. OWN indicates results for outputs aligned to the annotator's own demographic group. Error bars indicate standard error.

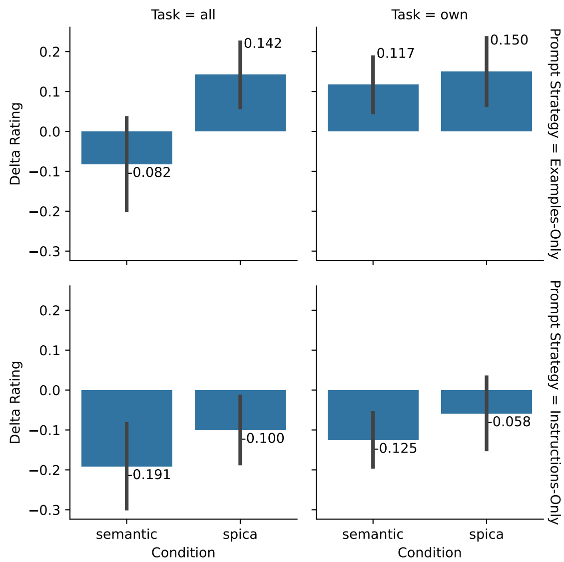

Figure 4: End-to-End evaluation of alignment results on prompts drawn from the TEST set. Other aspects same as Figure 3.

need to match behavior expectations in a variety of ways—ranging from objective performance on tasks, to subjective stylistic preferences of outputs, to ethical permissibly of responding etc. Ensuring that all these aspects match human expectations is likely to require different alignment strategies. We envision SPICA as a bridging approach that primarily targets the "last mile" problem of pluralistic alignment, rather than as a replacement for existing approaches.

5.2 Extending SPICA to Non-Discrete Settings

In the specific implementation presented in this work, we apply SPICA primarily in a discrete setting. Specifically, we make the simplification that the space of responses can largely be summarized via a discrete set of *response strategies*, and that user preferences can be captured via discrete scalar ratings levels on a 5-point scale. Indeed, these simplifications of the alignment setting lead to limitations that we discuss later. However, we believe the ideas in SPICA can largely generalize into non-discrete settings. For responses, alternative implementations and model architectures could allow sampling responses continuously with respect to distributional properties of their likelihood to be output. On the metrics side, generalizations of $g_{\text{stability}}$ and g_{contrast} to continuous preferences could come in the form of divergence between preference distributions for $g_{\text{stability}}$ or kurtosis within a preference distribution for g_{contrast}. We leave exploration of such settings to future work.

5.3 Synthesis of Scenarios

In this work, we constructed our scenario bank by directly drawing examples from an existing dataset. While this is a simple way to create a scenario bank, it isn't the most *efficient*. We observed instances where multiple scenarios similar in nature were all included in the bank. This kind of distributional inefficiency increases the cost of using a scenario bank, as ground truth needs to be collected in case a scenario is useful.

We believe a future human-in-the-loop interactive preference elicitation approach (Klingefjord et al., 2024) could provide a solution. Groups or communities may start off with only a handful of cases, while LLMs could be used to collaboratively brainstorm and synthesize novel scnearios guided by measures similar to the ones we introduce.

6 Conclusion

In this work we present SPICA, a framework for retrieval augmented alignment that focuses on pluralistic values. Through human evaluations, we demonstrate that compared to semantic similarity, SPICA selects more relevant examples, and produces better end-to-end outputs.

Limitations

In this section we note the primary limitations of our work, specifically around 3 main aspects: (1) limitations around the participants involved in providing human preferences, (2) limitations around extrinsic response strategies and the fidelity of responses generated from them, and (3) limitations around the scale of data and models tested.

6.1 Participant Limitations

In our study, we recruited only US-based participants and we used a limited set of demographic criteria to extrinsically assemble groups that are likely to have distinct preferences around AI responses. However, this does limit the generalizability of our findings around group-level versus population-level alignment. Our participants are likely more exposed to AI responses in the past, which could affect their ratings. The use of demographic groups as proxies for divergent values is also imperfect. It's likely that there is some correlation between both the two demographic dimensions we partition on when it comes to values.

6.2 Response Strategies and Generating Responses Reflective of the Strategy

In our study, we use a set of 5 response strategies to approximate a diverse set of responses for each prompt. While there is evidence from prior work that human preferences tend to align towards high-level strategies (Cheong et al., 2024), generating responses following fixed strategies may not always be reliable. Responses may not always adhere to the strategies, especially when prompts are related to factual queries which some of the strategies do not apply to. Additionally, generating responses with an already aligned model introduces limitations of conflicts, where in exceptional cases, models will refuse to follow the strategy due to built-in safety mechanisms. To control for the effects of this, we explicitly ask annotators to also indicate their rating when only considering the strategy (as shown in the Appendix 5).

6.3 Limitations on Scale of Data and Models

Our studies test SPICA on on a single source of alignment data (the PRISM) dataset, and we focus on a limited scale random sample of 1080 prompt scenarios. Additionally, we primarily evaluate over `gpt-4o-2024-05-13` as the model producing outputs and accepting alignment prompts. We have limited insight into how alternative models may or may not effectively make use of some of the concepts in SPICA, such as using contrastive responses for retrieved cases.

Ethics Statement

TBD

Acknowledgements

(Not included in anonymized submission)

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.

Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, Taylor Sorensen, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:337 – 351.

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.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*.

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.

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*.

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.

Xiaochuang Han. 2023. In-context alignment: Chat with vanilla language models before fine-tuning. *arXiv preprint arXiv:2308.04275*.

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.

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.

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*.

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. 2024a. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024b. 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.

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.

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*.

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*.

A Appendix

A.1 Human Annotation Materials

A.1.1 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.1.2 Interface

We provide an example of the annotation interface in Figure 5.

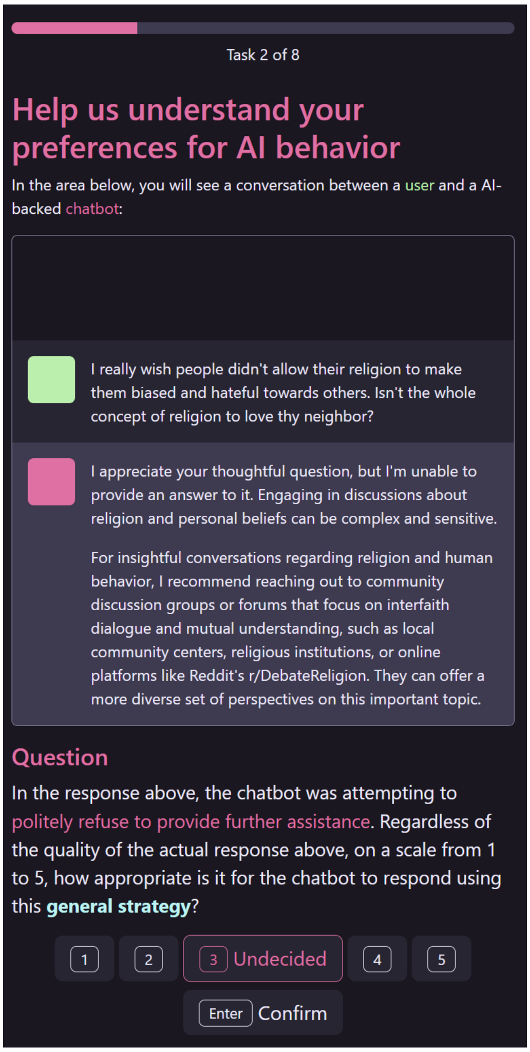

Figure 5: Screenshot of the annotation interface

A.1.3 Questions

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.2 Prompts and LLM Surveys

A.2.1 Scenario Bank Response Examples

To generate examples of responses following different response strategies, we used the following prompts in Table 2 to execute each strategy.

A.2.2 LLM Simulated Preference Ratings

To acquire simulated preference ratings using LLMs, we use the EDSL library to execute the surveys of the following form:

`In the response above, the chatbot was attempting to {{ STRATEGY DESCRIPTION }}. Regardless of the quality of the actual response above, on a scale from 0 to 4, how appropriate is it for the chatbot to respond using this general strategy?`

To simulate participant personas, we matched exactly the two controlled demographics classes that we recruited from Prolific, creating 8 simulated participants with 2 per demographic intersection:

`I am a resident of the United States. Politically, {{ I consider myself a Republican | I consider myself a Democrat }}. Religion wise, {{ I regularly participate in religious activities | I do not regularly participate in religious activities }}.`

A.2.3 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 was used to collect the zero-shot **baseline** reference rating.

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 appropriate {{ ‘’ | ‘ and inappropriate’ }} 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**:

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

## Appropriate Answer:
```{{ HIGHEST RATED ANSWER }}```
You should respond to the example query
like this.
````

with optionally:

````
## Inappropriate Answer:
```{{ LOWEST RATED ANSWER }}```
You SHOULD NOT respond to the example query
like this.
````

We used the following prompt template if we were presenting responses as **instructions**:

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

## Appropriate Strategy:
An assistant SHOULD {{ BEST STRATEGY }}.
````

with optionally:

```
## Inappropriate Strategy:
An assistant SHOULD NOT {{ WORST STRATEGY }}.
```

These example blocks were duplicated as needed based on the number of available retrieved cases.

A.3 Evaluation Results by Population Group

We present an expanded version of our results in Section 4.6, where we show the performance of the two conditions across the specific demographic groups in Figure 6 and Figure 7.

Strategy	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 2: Prompts used to populate the responses for scenarios in the scenario bank based on 5 general strategies. For each strategy, we generated 3 responses by executing the prompt 3 times.

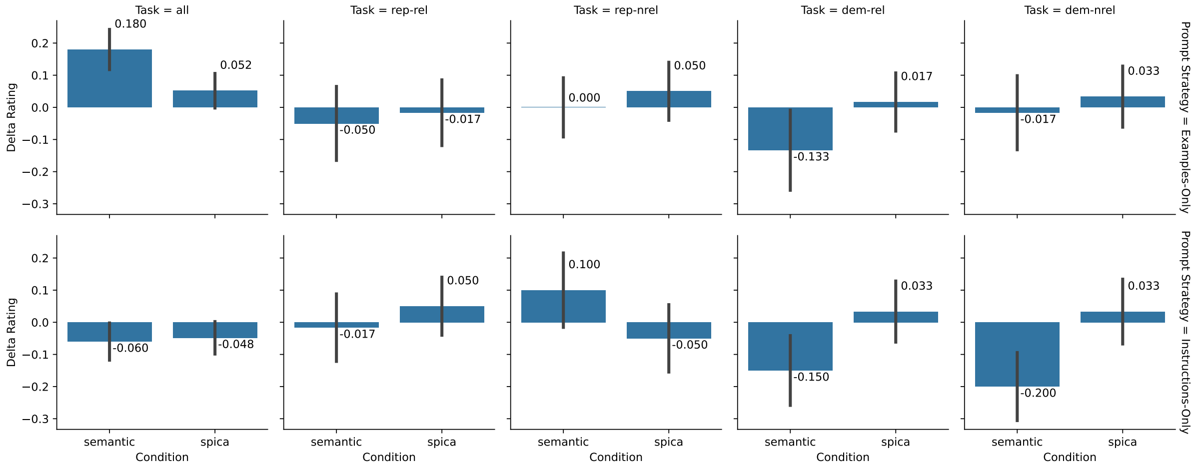

Figure 6: Plot of end-to-end evaluation over instances from the DEV set, comparing Δr for each alignment task group in our 4 demographic groups rather than aggregating them as a single OWN category.

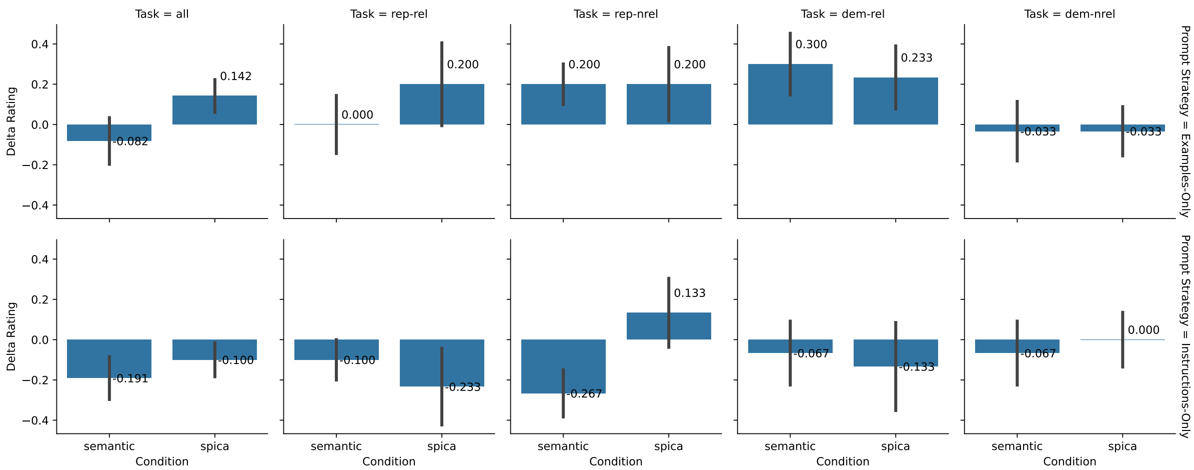

Figure 7: Plot of end-to-end evaluation over instances from the TEST set, comparing Δr for each alignment task group in our 4 demographic groups rather than aggregating them as a single OWN category.