SOLAR: Towards Characterizing Subjectivity of Individuals through Modeling Value Conflicts and Trade-offs

Younghun Lee and **Dan Goldwasser**
Department of Computer Science
Purdue University
West Lafayette, IN, USA
{younghun, dgoldwas}@purdue.edu

Abstract

Large Language Models (LLMs) not only have solved complex reasoning problems but also exhibit remarkable performance in tasks that require subjective decision-making. Existing studies suggest that LLM generations can convey subjectivity to some extent, yet exploring whether LLMs can account for individual-level subjectivity has not been sufficiently studied. In this paper, we characterize the subjectivity of individuals on social media and infer their moral judgments using LLMs. We propose a framework, SOLAR (**S**ubjective Gr**O**und with Va**L**ue **A**bst**R**action), that observes value conflicts and trade-offs in the user-generated texts to better represent subjective ground of individuals. Empirical results demonstrate that our framework enhances overall inference performance, with notable improvements for users with limited data and in controversial situations. Additionally, we qualitatively show that SOLAR provides explanations about individuals' value preferences, which can further account for their judgments.

1 Introduction

For the last few years, Large Language Models (LLMs) have shifted the paradigm of solving NLP problems to autoregressive language generation and achieved human-like performance in many downstream tasks (Raffel et al., 2020; Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022; Yu et al., 2022; He et al., 2024). Not only have LLMs solved objective problems that require complex reasoning skills, such as STEM-related questions (Imani et al., 2023; Wang et al., 2024c; Abbasiantaeb et al., 2024), they also exhibit remarkable performance in subjective decision-making processes, such as detecting toxicity (Hartvigsen et al., 2022), generating model evaluation (Perez et al., 2022), following ethical principles (Bai et al., 2022), etc.

Recent studies explore whether LLMs can generate perspectives and reasoning that align well with a specific persona or demographic information (Durmus et al., 2023; Nie et al., 2024; Zheng et al., 2024b). The results of these studies show that it is possible, to an extent, to ground LLM generations with these traits, however, LLMs tend to rely on superficial facts and assumptions about the roles and demographic traits rather than apply a deeper understanding.

Our goal in this paper is to study a different aspect of subjectivity, focusing on the *individual-level* (rather than generalizing over demographic traits), which has not been sufficiently studied yet. As personalized AI becomes more widely used (McClain, 2024), understanding whether LLMs can be utilized to characterize individual subjectivity becomes more important. Analyzing individual-level subjectivity using LLMs faces two main challenges. The first challenge is guiding LLM generation to be consistent with a specific subjective view. Existing methods for capturing subjectivity at the level of a generalized persona, role, or demographic information show that it is not trivial to steer LLMs to follow certain aspects (Durmus et al., 2023; Zheng et al., 2024b). Second, even if an optimal approach for grounding subjective aspects in LLMs' generations existed, expressing these aspects at the level of an individual is not straightforward (i.e., two instances of the same persona could still have different subjective preferences). Conceptualizing and operationalizing subjectivity at this level poses a second challenge.

In this paper, our objective is to characterize the subjectivity of individuals with LLMs by analyzing users' behaviors in a Reddit community, `r/AmITheAsshole`. In this community, original posters write about situations where they have conflicts with others and ask whether their behaviors are acceptable, and other redditors leave comments with their judgments. We formulate a classification

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20848–20863
November 4-9, 2025

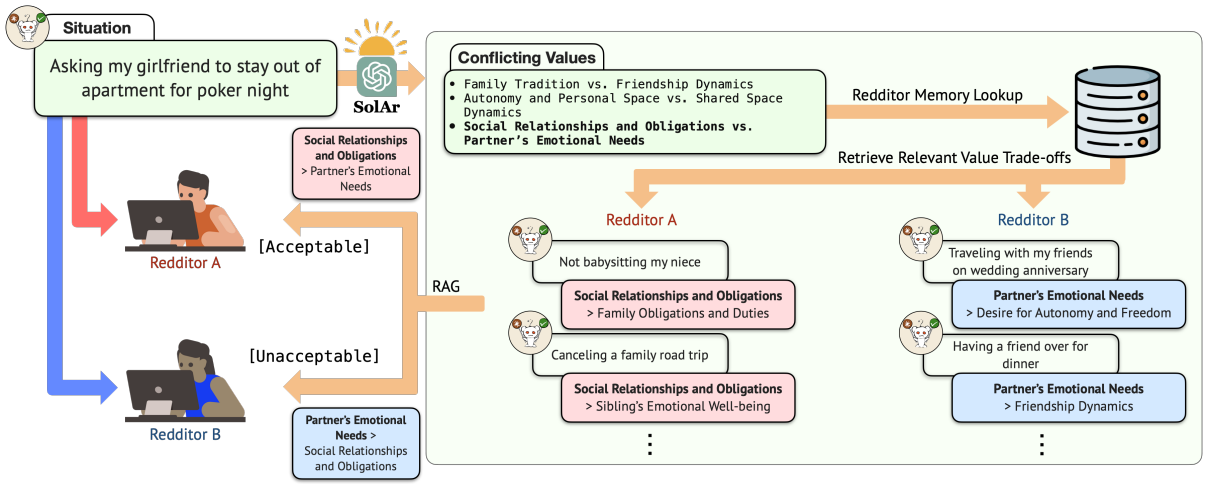

Figure 1: Inference process of our framework, SOLAR. When an input situation is given, SOLAR identifies conflicting values in the story, and retrieves the most relevant value trade-off history from the redditor's past comments. The retrieved subjective ground is then added to the prompt to infer each redditor's likely judgments to the situation.

task, predicting whether a specific redditor would judge a test situation acceptable or unacceptable, for evaluating different models' ability to capture subjectivity. In order to perform the classification task more efficiently, we hypothesize that redditors' *subjective ground*, principles that play a fundamental role in making moral judgments on others' situations (Neuhouser, 1990), can be represented by decomposing their past behaviors; we use redditors' past comments in the community to determine their likely judgments on unseen situations.

More specifically, we adopt value pluralism when considering redditors' past comments. Value pluralism suggests that there are multiple values that may be equally correct, yet in conflict with one another (Crowder, 1998; Galston, 2002). Psychological studies adopt this idea and argue that the human cognitive system makes novel judgments by making trade-offs between conflicting values when it encounters situations with colliding moral intuitions (Fiske and Tetlock, 1997; Guzmán et al., 2022). In this paper, we propose a framework, SOLAR (**S**ubjective Gr**O**und with Va**L**ue **A**bst**R**action), that observes trade-offs between conflicting values in the user-generated texts (i.e. comments), identifies the most relevant value conflicts with respect to the input situation, and infer the redditors' likely judgments using off-the-shelf LLMs. Our framework aims to tackle the two challenges of characterizing individual-level subjectivity described above; it conceptualizes individual-level subjectivity with value trade-offs, and grounds LLM generations with Retrieval-Augmented Generation (RAG). Empirical results show that SOLAR distinguishes redditors' subjective preferences and further provides explanations of their value trade-offs. Our framework improves the overall performance of downstream tasks and better guides LLMs, especially in low-resource user settings (i.e. users with a limited amount of data) and morally controversial scenarios. Figure 1 shows an example of how SOLAR determines the judgment of different redditors given a situation.

Key Contributions: To the best of our knowledge, this is the first attempt to explore whether off-the-shelf LLMs can be effectively used to account for individual-level subjectivity in the real-world online community. With value abstraction, we highlight that redditors show distinct subjectivity patterns. Additionally, we propose a novel framework that encompasses a value trade-off system and performs better in downstream tasks as well as grounding LLM generations.

2 Problem Formulation

In this section, we describe four major elements that formulate modules, learning processes, and inference tasks.

Situation Situation refers to the text description of what has happened in the real world. We focus on situations that portray conflicts so that one's point of view can be projected in diverse ways. *"Asking my girlfriend to stay out of the apartment for poker night"* is an example of a situation.

Individual This research aims to analyze different individuals' reactions and responses to various social situations. Unlike opinion mining or social commonsense reasoning studies, where perspectives and rationales are represented as an aggregate of a large number of humans, our main focus is to observe distinct patterns that characterize each individual and understand the rationales behind their decision-making processes.

Subjective Ground Subjective Ground refers to principles or maxims that steer individuals' moral judgments or perspectives. *"One should put their significant other's needs as a top priority."* is a subjective ground item that is relevant to the example situation above. Every individual has a distinct subjective ground and is built from various aspects such as their past experience, demographics, personality, etc. (Eysenck and Eysenck, 1975; Schwaba et al., 2023; Schoeller et al., 2024). One of our research hypotheses is that the subjective ground of individuals can be inferred by observing their past behaviors. Throughout this research, we aim to validate this hypothesis by modeling the subjective ground with individuals' comment history on social media.

Value Abstraction One of the limitations of the aforementioned assumption is that it works in a setting in which individuals' past behaviors can be observed across an extremely large range of social situations—an impractical scenario in real-world contexts. Thus in reality, when inferring individuals' likely judgments on unseen situations, it is necessary to formulate some hypotheses based on observable past history. Values, which "guide the selection or evaluation of actions, policies, people, and events", can serve to induce hypotheses about individuals' moral judgments (Schwartz, 2012). Value abstraction refers to a process that maps past history (i.e. comments) to a high-level representation of subjectivity that can be generalized to a broader spectrum of situations (i.e. values).

3 Task Description

We analyze subjective perspectives and judgments of individuals posted on a Reddit community, `r/AmITheAsshole`. In this community, original posters (OP) describe situations in which they have conflicts with others and ask if they are at fault. Other individuals (i.e. redditors) then leave comments and judge the acceptability of the OP's behaviors in the situation. Redditors' judgments are pre-coded words in the comment; `YTA` (You're The Asshole), `YWBTA` (You Would Be The Asshole), `NTA` (Not The Asshole), `YWNBTA` (You Would Not Be The Asshole), `ESH` (Everyone Sucks Here), `NAH` (No Assholes Here), and `INFO` (Not Enough Info). For simplicity, we group `NTA`, `NAH`, and `YWNBTA` as 'acceptable', and `YTA`, `ESH`, and `YWBTA` as 'unacceptable'. `INFO` is discarded as it does not convey subjectivity. We aim to learn the subjectivity of individuals by analyzing redditors' comments and judgment patterns on various situations.

Truncated `r/AmITheAsshole`	
# of total instances	53,280
# of unique situations	17,432
# of unique redditors	100
Max / Min # of instances per redditor	2,870 / 148
# Accept. / Unaccept. labels (overall)	38,365 / 14,915
# Accept. / Unaccept. labels (skewed)	2,615 / 127

Table 1: Statistics of the truncated dataset we use for training and inference. Label distributions are described in two ways; by combining all redditors (overall), and by combining three redditors who showed the most skewed judgment patterns (skewed).

There are a couple of benefits to using this community in analyzing individual-level subjectivity with language models. First, the situations described in this community are mostly about everyday events that are generic (e.g. *"not attending a friend's wedding"*) rather than related to specific world events (e.g. *"commenting on the new executive orders from the president"*); thus, the language models can have a better understanding of the situations without having a knowledge gap. Another benefit is that the redditors' subjective judgments are coded in a discrete fashion, which makes language model predictions more objective compared to open-ended analysis of subjectivity.

3.1 Crawling from `r/AmITheAsshole`

We crawl all posts in the `r/AmITheAsshole` community from November 2014 to June 2023 and filter out threaded comments to ensure all comments solely argue the acceptability of the situations. In order to perform training and inference more efficiently, we narrow it down to 1.7K unique reddit posts and 100 redditors who commented on these posts. We discuss how we truncate the data and whether the truncated situations and redditors appropriately represent the population distribution in the Appendix A. Detailed statistics of the truncated dataset are described in Table 1, and the data is

released for future reproduction [1].

3.2 Abstract Value Annotation

As discussed in 2, we produce a high-level abstraction of redditors' comments and the situations. For each pair of situations and an individual's comments, we prompt an off-the-shelf LLM[2] and generate values that are observed from the situation and redditors' comments. As human values can be defined and captured differently in the same text, we apply several different approaches.

We first analyze and annotate the texts using a top-down approach based on the theory of basic human values proposed by Schwartz (1992). The theory suggests ten basic values that could explain how people in different cultures recognize the underlying motivation and goals. One of the advantages of using this framework is that it explains values that align or conflict with one another, which naturally expands hypotheses of individual subjectivity. Detailed explanations of the ten basic human values and how they are annotated are described in Appendix B.1.

As opposed to using a fixed set of values for characterizing subjectivity, we use a bottom-up approach to discover more open-ended values observed in the texts. In this setup, we apply value trade-off theories. Humans make judgments based on value trade-offs when different values conflict to each other (Fiske and Tetlock, 1997; Leyva, 2019; Guzmán et al., 2022). As situations in the dataset mostly describe conflicts OPs are having, we prompt LLMs to identify all conflicting value pairs in the situation and discover value trade-offs made by each redditor in the comments.

When conflicting values are generated by LLMs, the level of abstraction is insufficient; values describe situation-specific details (e.g. *"prioritizing girlfriend's plan to cook together"*), rather than general concepts (e.g. *"partner's emotional needs"*). To address this, we iteratively cluster similar value representations and discover high-level definitions for these clustered values. We follow the approach used in Lam et al. (2024) to derive abstract value representations. Detailed processes for prompting LLMs to generate value trade-offs and clustering values are described in Appendix B.2. With more nuanced representations, we argue that the generated values provide richer contextual information than Schwartz's values and ultimately help better characterize individual subjectivity. As demonstrated in Appendix B.3, the values produced by the LLM integrate multiple Schwartz values, reflecting the multi-dimensional nature of subjective preferences.

3.3 Learning Problem

Our main focus is to use language models as a reasoning agent to understand individual-level subjectivity and predict their likely behaviors in unseen instances. More specifically, we formulate a binary classification task where the language models are given situations and redditors' most relevant subjective ground, either their past comments or value preference, and predict whether the redditors would judge the OP's behaviors acceptable or not.

4 Model

In this research, we propose a framework, SOLAR, that accomplishes the task with Retrieval-Augmented Generations (RAG).

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ denote the set of all distinct redditors, and $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ indicate the set of situations (i.e. reddit posts). When an i-th redditor commented on a j-th situation, we define their comment as c_{ij} and the acceptability judgment as y_{ij} where $y_{ij} \in \{0, 1\}$. More specifically, $\mathcal{C} = \{c_{ij} \mid u_i \text{ commented on } s_j\}$ and $\mathcal{Y} = \{y_{ij} \mid u_i \text{ commented on } s_j\}$. After we annotate the moral values of the situations and comments, we get the values of each situation as $s_j^{\mathcal{V}}$ and the values of each comment as $c_{ij}^{\mathcal{V}}$. Using a dynamic embedding model f_{embed}, all situations, comments, and values are transformed into vectors[3]. We denote these vectors by bold letters such as $\mathbf{s}_j = f_{embed}(s_j)$.

4.1 Subjective Ground Retrieval

We first define the redditor history data where the retrieval function searches for the most relevant instances to the input. We keep data separate for each redditor and let $\mathcal{D}_i$ denote the history data of an i-th redditor. $\mathcal{D}_i$ contains representations of the situations and comments, namely $\mathcal{D}_i = \{(s_j, s_j^{\mathcal{V}}, c_{ij}, c_{ij}^{\mathcal{V}}, y_{ij}) \mid u_i \text{ commented on } s_j\}$.

In order to infer how the target individual would react to a test situation, x, we design several heuristics to retrieve the most relevant instances. First, we query the retrieval function with the raw situation

[1] https://github.com/younggns/solar

[2] We use OpenAI's *gpt-4o-0806* model

[3] We use OpenAI's *text-embedding-3-large* model.

representations. Each query consists of a user indicator and a vector representation of a test situation, $\mathbf{x}$. Then the retrieval function outputs k nearest instances with respect to the Euclidean distance between the test situation and situations in the history data $\mathcal{D}$. Let $R(\ )$ be the retrieval function, we define subjective ground retrieval with situations by:

$$R(u_i, \mathbf{x}) = \underset{\mathbf{s}_j \in \mathcal{D}_i}{\text{top-}k}\ \text{dist}(\mathbf{x}, \mathbf{s}_j)^{-1} \quad (1)$$

where $\text{dist}(\mathbf{x}, \mathbf{s}_j) = \|\mathbf{x} - \mathbf{s}_j\|_2$

As an alternative to computing the distance between situation representations, we could use abstract values. The motivation for using value representations is to retrieve comments from more diverse cases; distance among situations yields topically similar situations only, while using values can retrieve situations that are relevant in terms of high-level values. Consider, for example, an input situation about wedding ceremonies. While retrieving relevant items with situations provides wedding related instances, using abstract value representations could retrieve situations about social appearance, money, family relationship dynamics, etc. Let $R_{val}(\ )$ be the retrieval function that takes a user indicator and a value representation of a test situation, $\mathbf{x}^{\mathcal{V}}$, we define subjective ground retrieval with values by:

$$R_{val}(u_i, \mathbf{x}^{\mathcal{V}}) = \underset{\mathbf{s}_j^{\mathcal{V}} \in \mathcal{D}_i}{\text{top-}k}\ \text{dist}(\mathbf{x}^{\mathcal{V}}, \mathbf{s}_j^{\mathcal{V}})^{-1} \quad (2)$$

4.2 Language Model Prompting

After the retrieval functions provide the most relevant past history of a target redditor to the test situation, we prompt off-the-shelf LLMs to predict the likely judgment of the redditor. Essentially, the LLMs are considered as a general reasoner that accounts for individual subjectivity; LLMs are not trained to represent each individual's subjectivity, but they perform inference based on the given pieces of specific redditor's past subjective behaviors.

We add k nearest instances as few-shot examples. In each of the few-shot examples, a short description of the retrieved situation, a redditor's comment, and their judgment are included. In order to observe the usefulness of the annotated values, we also experiment with adding values that are perceived in the comments in few-shot examples. An example prompt is described below:

Moral Judgment Prediction - Baselines		
Model	**All Redditors**	**Top 8 Redditors**
Encoder-only Models		
DistilBERT-base-uncased	47.90 $^{\pm 0.0053}$	68.71 $^{\pm 0.0145}$
RoBERTa-base	46.68 $^{\pm 0.0038}$	70.48 $^{\pm 0.0088}$
DeBERTa-v3-large	42.45 $^{\pm 0.0013}$	70.94 $^{\pm 0.0024}$
Encoder-Decoder Models		
BART-base	48.90 $^{\pm 0.0055}$	68.50 $^{\pm 0.0150}$
FLAN-T5-base	46.21 $^{\pm 0.0035}$	65.44 $^{\pm 0.0118}$
Encoder-Decoder Models; Seq2Seq		
BART-base	40.83 $^{\pm 0.0012}$	66.67 $^{\pm 0.0086}$
FLAN-T5-base	43.01 $^{\pm 0.0020}$	66.86 $^{\pm 0.0087}$

Table 2: Macro F1 scores of baseline models. The average performance of fine-tuned models over all 100 redditors shows a significance drop compared to the performance of the top 8 redditors who commented over 1,000 times.

Prompt Example with Comments Only

You will be given examples of a situation, Person X's comment on the situation, and Person X's judgment on the situation (i.e. whether it is acceptable or not).

[Situation] Not babysitting my niece
[Comment] When you talk to your brother ..
[Judgment] Acceptable
..

Now you will be given a new situation. Based on your understanding of Person X's judgments on different situations, tell me how Person X would judge the new situation.
[Situation] Canceling a family road trip

5 Experiments

In all experiments, we use the macro F1 score as the evaluation metric due to the imbalanced label distribution. For each model, we first calculate the macro F1 score individually for each redditor, then take the unweighted average of these scores across all redditors. This approach ensures that every redditor has equal influence on the final score, regardless of how many instances each contributed. Even if a model performs well on some redditors who are easily predictable, its overall score becomes lower if it fails to learn the subjectivity of other redditors.

5.1 Baseline Models

In addition to our proposed framework using LLMs for reasoning, we implement trainable language models varying in structures and inputs to show the difficulties in understanding individual-level

Moral Judgment Prediction - RAG-based Models							
Retrieval Strategy	**Input**	**All Redditors**		**Top 50% Redditors**		**Bottom 50% Redditors**	
		All	Contro	All	Contro	All	Contro
Situation	Comment	$78.62^{\pm 0.001}$	$78.61^{\pm 0.002}$	$78.58^{\pm 0.004}$	$80.48^{\pm 0.001}$	$78.66^{\pm 0.004}$	$76.53^{\pm 0.000}$
	Comment + Trade-off	$79.70^{\pm 0.001}$	$80.32^{\pm 0.002}$	$78.62^{\pm 0.001}$	$80.17^{\pm 0.002}$	$80.78^{\pm 0.002}$	$80.47^{\pm 0.004}$
	Comment + Schwartz	$78.14^{\pm 0.000}$	$79.11^{\pm 0.001}$	$76.53^{\pm 0.000}$	$79.45^{\pm 0.000}$	$79.74^{\pm 0.003}$	$78.73^{\pm 0.004}$
Schwartz's Value	Comment	$79.05^{\pm 0.003}$	$81.38^{\pm 0.006}$	$78.29^{\pm 0.005}$	$80.97^{\pm 0.004}$	$79.81^{\pm 0.002}$	$81.82^{\pm 0.007}$
	Comment + Schwartz	$77.35^{\pm 0.002}$	$80.73^{\pm 0.008}$	$76.47^{\pm 0.003}$	$80.47^{\pm 0.013}$	$78.24^{\pm 0.001}$	$81.01^{\pm 0.002}$
Value Trade-off	Comment	$78.90^{\pm 0.000}$	$80.76^{\pm 0.000}$	$77.79^{\pm 0.003}$	$81.14^{\pm 0.007}$	$79.96^{\pm 0.001}$	$80.78^{\pm 0.009}$
	Comment + Trade-off	$78.33^{\pm 0.001}$	$82.44^{\pm 0.005}$	$77.26^{\pm 0.001}$	$82.44^{\pm 0.005}$	$79.40^{\pm 0.002}$	$82.43^{\pm 0.005}$
⋆SOLAR: *Ensemble Retrieval and Input*		**79.90** $^{\pm 0.001}$	**82.44** $^{\pm 0.005}$	**78.80** $^{\pm 0.001}$	**82.44** $^{\pm 0.005}$	**80.99** $^{\pm 0.001}$	**82.43** $^{\pm 0.005}$

Table 3: Performance of different retrieval strategies across author percentiles with respect to data size. "Contro" denotes controversial situations. SOLAR, which ensembles different strategies and input types for controversial and non-controversial situations at test time, achieves the best overall performance. The performance boost is more significant for the bottom 50% redditors and in controversial situations.

subjectivity. We use pre-trained language models varying in encoder and decoder structures and fine-tune the models with our dataset. We randomly split each redditor's instances into 60/10/30% to obtain the training, validation, and test set, and run 5-fold cross-validation. Language models are fine-tuned for each redditor, thus we fine-tune 100 distinct models for each model structure.

We implement three encoder-only models, DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2021), where the input for each model is a text description of the situation and the output is the redditor's acceptability judgments, 0 or 1. We denote this approach as *Endoer-only Models*.

Denoted as *Endoer-Decoder Models*, we fine-tune encoder-decoder language models, mainly BART (Lewis, 2019) and FLAN-T5 (Chung et al., 2024). We report the experimental results of these models with the same classification setting as the encoder-only models; models get situation text as input and output either 0 or 1.

To learn more about each redditor's subjectivity, we test another variation. The models get the same input, but their objective is to generate the redditors' likely reactions (i.e., comments and judgments). While encoder-only models only observe each redditor's binary judgment patterns with respect to the situations, the models fine-tuned with this approach have access to the comments that are authored by the redditors. This approach is denoted as *Endoer-Decoder Models; Seq2Seq*. Table 2 illustrates the results of the baseline model. Further details of the baseline model implementation are described in Appendix C.1.

5.2 RAG-based Models

We compute the performance for each redditor separately, using the same test set as the ones used in baseline models. As described in 4.1, we represent the individual subjectivity in a several different ways.

First, we try different retrieval strategies by querying past history data with situations and abstract values that are annotated by LLMs. For querying with abstract values, we report the results of using Schwartz's values and value trade-offs. After retrieval, we differentiate LLM inputs (i.e. few-shot examples) by adding Schwartz's values and value trade-offs that are observed from each redditor's past history.

Our proposed framework, SOLAR, ensembles retrieval strategies based on the difficulty of the test instance. When the test situation is given, SOLAR computes its difficulty based on how controversial it is—defined as having less than 70% agreement among all redditors' judgments[4]. For controversial situations, SOLAR employs the value-aware retrieval function $R_{val}(\ )$, while it uses the situation-based retrieval function $R(\ )$ for non-controversial cases. In both scenarios, few-shot examples include both the comment and the corresponding value trade-offs.

We use GPT-4.1 (OpenAI, 2025) to predict the acceptability judgments of individuals. Each method retrieves five few-shot examples and prompts the LLM twice to estimate a confidence interval. Table 3 shows the performance of each method. Appendix C.2 explains more detailed ex-

[4]This is the threshold that `r/AITAFiltered` uses to identify controversial situations from `r/AmITheAsshole`.

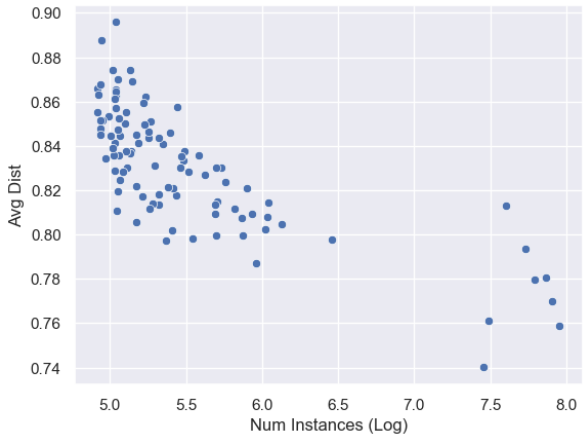

Figure 2: Average distance between the test situation and retrieved instances for each redditor. x-axis is the log-scaled numbers of each redditor's data size. When the redditor has more instances, it is more likely that they have commented on situations that are more similar to the test situation in the past.

perimental settings for RAG-based models.

6 Discussion

In this section, we analyze the inference performance of different models and discuss the effectiveness of each model component. In addition, we perform qualitative analyses of each redditor's subjective ground with respect to annotated abstract values and assess how the proposed framework explains individual-level subjectivity.

6.1 Inference Performance

The overall F1 scores of the baseline models in Table 2 show that fine-tuning redditors' decision patterns with language models does not solve the problem. The models show fairly good performance only for the redditors who have enough amount of training instances, while they fail to learn the subjectivity of redditors who have a small number of instances and highly skewed judgment patterns. We visualize fine-tuned models' performance with respect to the number of instances and judgment skewness in Appendix D.

We also observe that as the model becomes more complex and parameterized, the inference performance worsens—the F1 scores of encoder-decoder models are worse than encoder-only models, and sequence-to-sequence models are even worse than that. This implies that the nature of data scarcity in subjectivity analysis also makes it more difficult to fine-tune a specific individual's perspectives to characterize their subjectivity.

On the other hand, RAG-based models that select the most relevant instances to a test situation work generally well. This answers one of our research questions, *"Can off-the-shelf LLMs account for individual's subjective preference?"*. As RAG-based inference does not require training, the performance of redditors with less amount of instances (i.e. Bottom 50%) matches the overall F1 scores.

Different retrieval methods and inputs show distinct advantages. First of all, when we compare different inputs within the standard retrieval strategy (i.e. Situation), adding values improves the performance for redditors who have less amount of instances; using comment and value trade-offs improves the F1 score by 2.69% for the bottom 50% redditors. Figure 2 shows that the average distance between the test situation and the retrieved instances increases for redditors with fewer instances. The performance boost we get for the bottom 50% redditors implies that adding abstract values helps characterize the redditors' subjectivity when the retrieved instances are topically less similar to the test situation.

Abstract values are also helpful when we use them for retrieval. The performance in controversial situations improves when the most relevant instances are retrieved using either Schwartz's values or value trade-offs. This suggests that predicting individuals' subjective preferences benefits from knowing their high-level value system, especially when the test situation is controversial and LLM's own morality does not align well with humans.

6.2 Diverse Value Trade-offs among Redditors

A key research hypothesis of this study is that each redditor actively participating in `r/AmITheAsshole` exhibits distinct subjectivity patterns, making the analysis of individual perspectives a fundamentally different NLP challenge compared to social commonsense reasoning. We validate this hypothesis by visualizing the value trade-off patterns of the 8 most active redditors.

Figure 3 displays the 8 most common value trade-offs of the most active redditors. Each cell in the heatmap refers to win rates. For example, 0.9 in the cell in row 1, column 1 means that among all situations where a value *"Boundary Awareness"* conflicts to *"Cultural Expectations"*, redditor 1 chooses the former over the latter 90% of the times. For the same situations, redditor 3, who has a 0.45 win rate, chooses *"Boundary Awareness"* only 45% of the time, implying they prefer *"Cultural Expec-*

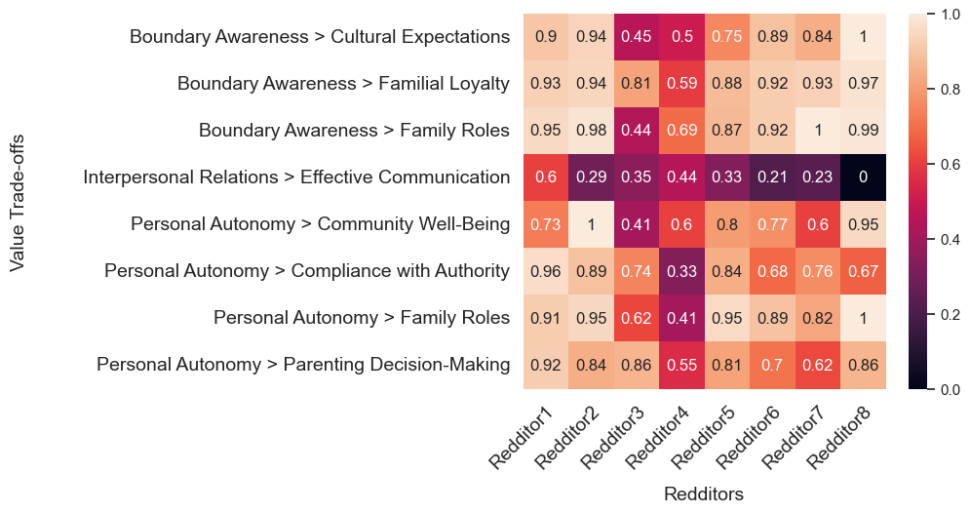

Figure 3: ***Visualization of subjective preferences.*** The heatmap shows the most common value conflicts among the 8 most active redditors. Each value trade-offs are represented in a "A > B" form, and each cell's value means the win rate of value A over value B for a specific redditor. In most cases, redditors' value trade-off systems work differently.

tations" slightly more.

The heatmap tells us that there are some values commonly cherished by most of the redditors. The bottom row of the heatmap, for instance, shows that all redditors prefer *"Personal Autonomy"* over *"Parenting Decision-Making"* in most cases. At the same time, redditors show diverse patterns when it comes to *"Personal Autonomy"* conflicts with other values. For redditor 8, specifically, this value is highly prioritized in most cases, unless it conflicts with *"Compliance with Authority"*. This visualization not only suggests that active redditors in `r/AmITheAsshole` show distinct subjectivity, but explains their value preferences and judgments.

7 Related Studies

The closest neighbor of this research is individual subjectivity analysis. Lee and Goldwasser (2022) analyzes the same community, `r/AmITheAsshole` and learned subjective preference of individuals by computing attention weights between situations and social norms. Plepi et al. (2022) used the authorship information with text embeddings and perform the YTA/NTA classification of the redditors in `r/AmITheAsshole`. In the political framing and agenda-setting domain, Roy and Goldwasser (2021) utilized Moral Foundations theory to characterize real world politicians varying in topics. More recently, researchers utilized LLMs to ground their generations to a desired persona or personality in many domains, including question answering (Zheng et al., 2024a), interactive simulacra (Park et al., 2023), and solving causal inference and moral dilemma (Nie et al., 2023). Choi and Li (2024) proposed Persona In-Context Learning which uses Bayesian inference to select the optimal set of persona for a given task. Researchers have also tried providing more direct signals in the prompt using demographic information (Durmus et al., 2023).

Incorporating value theories into LLMs is an active research area these days. Van Der Meer et al. (2023) analyzed (dis-)agreements between the users by identifying value profiles with Schwartz's value theories. Sorensen et al. (2025) used LLMs as proxy humans with different attributes (e.g. demographic traits, value profiles), and analyzed their distinct behaviors on controversial issues. Bhatia et al. (2025) used LLMs to characterize individuals' choices by generating descriptions of benefits and costs of their choices. Sorensen et al. (2024) used concrete examples (i.e. text document) and prompted LLMs to generate corresponding values, rights, and duties. Ye et al. (2025) parsed free-form input text into various perceptions and the corresponding Schwartz values by fine-tuning LLMs.

The main differences between our approach and the existing studies in subjectivity analysis and human value analyses with LLMs are; (1) we analyze actual redditors in real world rather than using LLMs as proxy humans, (2) our approach applies more fine-grained buckets for subjectivity, and (3) the novel value trade-off analysis provides addi-

tional explanation for individuals' choices.

Another line of related studies is the Retrieval-Augmented Generation. RAG retrieves the most relevant information to mainly fine-tune language models or help off-the-shelf LLMs infer better on many downstream tasks such as QA (Shi et al., 2023; Xu et al., 2023; Wang et al., 2024b), reasoning and language understanding (Yu et al., 2023; Lin et al., 2023; Zhang et al., 2023), text summarization and generation (Guo et al., 2023; Jiang et al., 2023; Yan et al., 2024), etc. Researchers have also studied whether RAG can be applied to more personal use cases such as recommendation system (Rajput et al., 2023) and personalized dialog generation (Wang et al., 2023, 2024a). These approaches only consider factual information (e.g. purchase history, where the person went two days ago) as a personalized aspect, while our research characterizes subjective perspectives and applies RAG for performance improvement.

8 Conclusion

In this paper, we propose a framework, SOLAR, that takes redditors' past comments into account and characterizes their subjective ground using value abstraction. Empirical results show that LLMs can be efficiently used to account for the subjective preference of individuals compared to traditional methods that require fine-tuning. We also show that the performance, especially for redditors with small data sizes and in controversial situations, is improved by retrieving the most relevant instances using abstract values. Furthermore, SOLAR provides additional explanations about each redditor's distinct value preference patterns which could later be used to justify LLM inference.

Limitations

Although it is not feasible to model the subjectivity of individuals that is perfectly correct and inclusive, representing it with their past comments is still an oversimplified definition of subjectivity. First, past comments in the same community might not cover all aspects of the subjectivity. In future studies, we plan to incorporate more redditor-related information such as their community membership (i.e., what other subreddits they are actively participating in) and activities in other communities to better characterize individuals. Another oversimplification is that it assumes the redditors would judge the situations consistently over time. It would be an interesting direction to analyze whether there are value shifts over time for the redditors.

Another limitation of this study is that the datasets and the tasks are tested only on a specific subreddit. Although `r/AmITheAsshole` is a huge online community that covers a wide range of situations exhibiting diverse perspectives, the usefulness of our framework will be more strongly validated when we apply our approach to other communities that require subjective perspectives.

In terms of downstream tasks and the model's performance, our proposed framework shows suboptimal performance in predicting the correct judgment. Although our framework shows higher improvements when considering controversial situations, our goal is to make LLMs perform well not only in controversial situations but also in other situations.

Lastly, more validations of the generated abstract values are needed. As we use LLMs to freely generate value trade-offs of redditors that are observed from their comments, evaluating the soundness and quality of the values would make the subjectivity representation more powerful and useful. We further plan to validate this process with actual human evaluations and see if LLM-generated values can characterize human values reasonably well. The best way to evaluate LLMs' characterization of individuals would be directly ask it to the target individuals. We leave recruiting human participants to accurately evaluate LLM generations as future work.

Acknowledgements

This project was partially funded by NSF IIS-2048001 and DARPA CCU program. The views are the authors' and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government.

Ethics Statement

To the best of our knowledge, this work has not violated any code of ethics. All redditor information is anonymized in this paper as well as in the datasets we share to the public. The redditors are selected purely based on how active they participate in the community thus there is no discrimination in choosing redditors of interest. We denote that this paper poses potential risks where LLMs could misrepresent the subjectivity of individuals by referring

to a limited number of past history and making hypotheses on their likely reactions to an unseen situations. We provide the code and datasets for future reproduction.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

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*.

Sudeep Bhatia, Simon T van Baal, Feiyi Wang, and Lukasz Walasek. 2025. Computational analysis of 100 k choice dilemmas: Decision attributes, trade-off structures, and model-based prediction. *Proceedings of the National Academy of Sciences*, 122(17):e2406489122.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *Forty-first International Conference on Machine Learning*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

George Crowder. 1998. From value pluralism to liberalism. *Critical Review of International Social and Political Philosophy*, 1(3):2–17.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Hans Jurgen Eysenck and Sybil Bianca Giuletta Eysenck. 1975. *Manual of the Eysenck Personality Questionnaire (junior & adult)*. Hodder and Stoughton Educational.

Alan Page Fiske and Philip E Tetlock. 1997. Taboo trade-offs: reactions to transactions that transgress the spheres of justice. *Political psychology*, 18(2):255–297.

William A Galston. 2002. *Liberal pluralism: The implications of value pluralism for political theory and practice*. Cambridge University Press.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. 2023. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. *arXiv preprint arXiv:2305.17653*.

Ricardo Andrés Guzmán, María Teresa Barbato, Daniel Sznycer, and Leda Cosmides. 2022. A moral tradeoff system produces intuitive judgments that are rational and coherent and strike a balance between conflicting moral values. *Proceedings of the National Academy of Sciences*, 119(42):e2214005119.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Younghun Lee and Dan Goldwasser. 2022. Towards explaining subjective ground of individuals on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1752–1766.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Rodolfo Leyva. 2019. Towards a cognitive-sociological theory of subjectivity and habitus formation in neoliberal societies. *European Journal of Social Theory*, 22(2):250–271.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Colleen McClain. 2024. Americans' use of chatgpt is ticking up, but few trust its election information. Accessed: 2025-02-13.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Frederick Neuhouser. 1990. *Fichte's theory of subjectivity*. Cambridge University Press.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36:78360–78393.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2024. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2024. text-embedding-3-small. AI model, published on January 25, 2024.

OpenAI. 2025. Gpt-4.1. AI model, published on April 14, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Felix Schoeller, Leonardo Christov-Moore, Caitlin Lynch, Thomas Diot, and Nicco Reggente. 2024. Predicting individual differences in peak emotional response. *PNAS nexus*, 3(3):pgae066.

Ted Schwaba, Jaap JA Denissen, Maike Luhmann, Christopher J Hopwood, and Wiebke Bleidorn. 2023. Subjective experiences of life events match individual differences in personality development. *Journal of Personality and Social Psychology*, 125(5):1136.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

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

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. Value profiles for encoding human variation. *arXiv preprint arXiv:2503.15484*.

Michiel Van Der Meer, Piek Vossen, Catholijn M Jonker, and Pradeep K Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023. Large language models as source planner for personalized knowledge-grounded dialogue. *arXiv preprint arXiv:2310.08840*.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z Pan, and Kam-Fai Wong. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2025. Measuring human and ai values based on generative psychometrics with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26400–26408.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023. Iag: Induction-augmented generation framework for answering reasoning questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–14.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024a. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024b. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

A Data Distribution

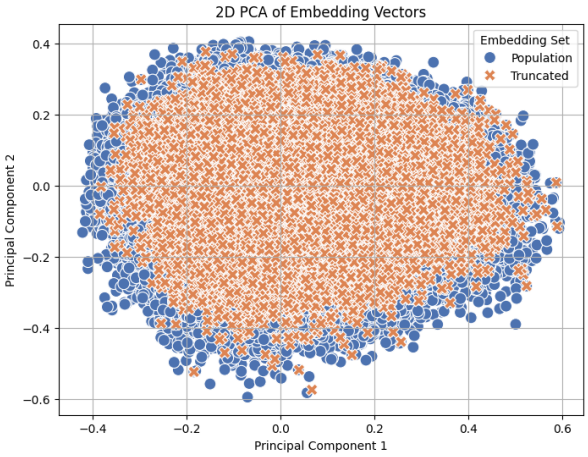

Figure 4: PCA analysis of situation representations of all `r/AmITheAsshole` posts and truncated posts we use for training and inference. Truncated situations show similar distributions to the population distribution.

As described in 3.1, we crawl all posts in `r/AmITheAsshole` from November 2014 to June 2023 and filter out threaded comments. As a result, we have around 217K unique situations with

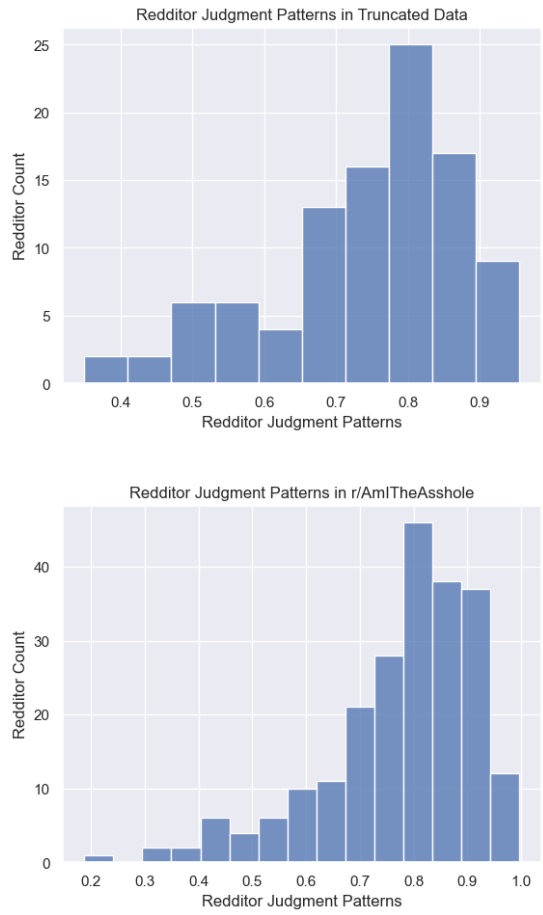

Figure 5: Redditors' judgment patterns with respect to the redditor counts in the truncated data (top), and in `r/AmITheAsshole` (bottom).

891K instances in total. Fine-tuning models and performing LLM inference on this massive dataset takes up too many resources. Thus we truncate the dataset. We first identify the most active redditors (i.e. redditors who commented the most). Then we identify the redditors who commented more than a certain threshold[5]. We get 8 redditors from this process. After this step, we filter out situations that the 8 most active redditors did not leave comments. As a result, 1.7K unique situations are left.

To ensure that the truncated situations and all redditors who commented on these situations are not too different from the population distribution, we first visualize situation representations in a 2-D embedding space using Principal Component Analysis. Figure 4 visualizes the vector representations of all 217K situations in `r/AmITheAsshole` (blue dots), and vectors of 1.7K situations in the truncated dataset (orange crosses). The plot implies that the situations in the truncated dataset are not deviated from the population distribution.

We compare the statistics of redditors in `r/AmITheAsshole` and the truncated dataset. Figure 5 shows the redditors' judgment patterns (i.e. acceptable or not acceptable) in the truncated data and in all `r/AmITheAsshole`. A judgment pattern of 1 means that the redditor judged all situations as "acceptable", and 0 means they judged all as "not acceptable". When the value is 0.5, the redditor has a perfectly balanced judgment history. In the truncated data, redditors tend to judge more situations as "acceptable", and the trend is the same in the `r/AmITheAsshole`. This shows that the redditors in the truncated dataset do not diverge from the overall population.

B Value Abstraction

B.1 Schwartz's Theory of Basic Human Values

Schwartz (1992) defines theory of basic human values as:

- Self-direction: "independent thought and action—choosing, creating, and exploring"
- Stimulation: "excitement, novelty and challenge in life"
- Hedonism: "pleasure or sensuous gratification for oneself"
- Achievement: "personal success through demonstrating competence"
- Power: "social status and prestige, control or dominance over people and resources"
- Security: "safety, harmony, and stability of society, of relationships, and of self"
- Conformity: "restraint of actions likely to upset or harm others and violate social norms"
- Tradition: "respect of the customs and ideas that one's culture or religion provides"
- Benevolence: "preserving and enhancing the welfare of those within-group)"
- Universalism: "protection for the welfare of all people and for nature"

In order to annotate situations and comments to these fixed values, we prompt *gpt-4o-0806* model to generate values observed from the text. Below is an example prompt:

[5]We set this as 2,000

Prompt for Annotating Schwartz's Values

You will be given a social situation and a Redditor's comment. Your job is to find the top two most salient Schwartz's Basic Human Values that are observed in the situation and in the comment. Follow the format below without generating any other explanation.

Input Format:
[Situation] situation_description
[Comment] redditor_comment

Output Format:
[Situation] Schwartz_value_A, Schwartz_value_B
[Comment] Schwartz_value_C, Schwartz_value_D

###
[Situation] Not babysitting my niece
[Comment] When you talk to your brother ..

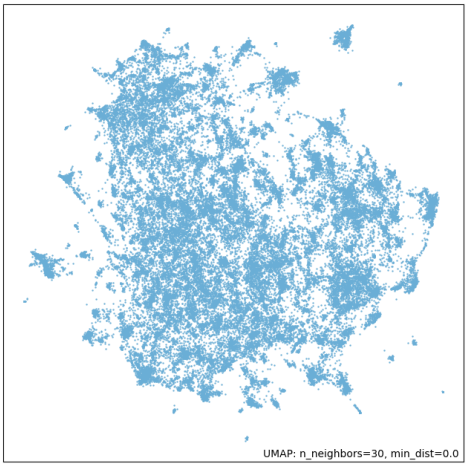

Figure 6: Visualization of Umap embeddings of value representations.

B.2 Clustering Value Conflicts and Trade-offs

Similar to annotating Schwartz's values, we prompt *gpt-4o-0806* model and generate value conflicts and trade-offs. Below is an example prompt:

Prompt for Annotating Value Conflicts

You will be given a situation. Your task is to understand the situation, and tell me what kind of moral values are conflicting.

[Situation] Not babysitting my niece ...

Tell me what kind of moral values are conflicting in the situation. Note that the two conflicting values can't be chosen at the same time (i.e. trade-off), and people's behaviors and attitudes will be different based on which value they choose. Highlight one or more core conflicting values that describe the situation the best. Values should be generic; proper nouns or pronouns should not be included. Write values in phrases, not in sentences, and each value should have at least 3 words. Your response should follow the format:
[
"A vs. B",
"C vs. D",
..
]

After the conflicting values are annotated for the situations, we prompt the GPT model again to analyze the comments from each redditor. Below is a prompt:

Prompt for Annotating Value Trade-offs

You will be given a [Situation] and a list of [Conflicting Values] observed in the situation. There's Person X who leaves a [Comment].

[Task] Determine which items in the [Conflicting Values] are mostly related to Person X's [Comment]. You may select only one of them, or select multiple items. Generate conflicting values observed in the comment ONLY IF none of the [Conflicting Values] are related to the [Comment]. Determine which value is more important to Person X. For instance, "A vs. B" is a chosen conflicting value and Person X thinks A is more important than B in the comment, the answer should be "A > B".

Your response should follow the format:
[
"A > B",
"D > C",
...
]

###
[Situation] Not babysitting my niece
[Conflicting Values] Autonomy vs. Family
[Comment] When you talk to your brother ..

The general framework of creating clusters from the annotated values follow the approaches suggested by Lam et al. (2024). The authors first cluster the texts using HDBSCAN, then expand the concepts by prompting LLMs. The initial clusters are formed using HDBSCAN (McInnes et al., 2017), and we later use LLMs to create additional clusters for values that remain uncategorized in the initial clustering phase, resulting in 111 clusters in

total.

For the generated values that are conflicting in the situations in the dataset, we first obtain their vector representations. We use OpenAI's text-embedding-3-large model (OpenAI, 2024), and reduced the dimensions to 256. We then further reduce the dimensionality using umap embeddings (McInnes et al., 2018). Figure 6 shows the 2-D visualization of umap embedding of all value representations, with number of neighbors as 30 and minimum distance as 0.

After this step, we use HDBSCAN to generate initial clusters while enforcing minimum of 100 items in each cluster. After obtaining the initial cluster, we gather all uncategorized values. We then compute the distance between each of the uncategorized value and initial cluster representations. If the distance is closer than a threshold (0.95), we assign the item to the cluster. For values that are not close enough to any other clusters, we group them using K-means clustering.

After assigning all value representations to a corresponding cluster, we ask *gpt-4o-mini* model to come up with a summary of unifying themes and patterns. Below is the template for the prompt adopted from Lam et al. (2024):

Prompt for Expand Concepts

I have this set of bullet point summaries of text examples:
{examples}

Please write a summary of unifying patterns for these examples.
For each high-level pattern, write a 5 word NAME for the pattern and an associated one-sentence ChatGPT PROMPT that could take in a new text example and determine whether the relevant pattern applies.
Please also include 3 example_ids for items that BEST exemplify the pattern.

B.3 Comparison

In order to analyze the usefulness of LLM-generated values and whether they cover the fixed set of values that are defined in a top-down approach, we map these values to the Schwartz's values.

Figure 7 shows the mapping results. For each of the situation that has annotations of both LLM-generated values and Schwartz's values, we count the co-occurrences between these values. For instance, if a situation's annotated Schwartz's value is "*Security*", and at the same time, its annotated values that conflict is "*Boundary Awareness and Respect*", then these two values have a co-occurrence. In the figure, we compute the log-scale of all the co-occurrence counts.

For each of the LLM-generated values, they cover multiple dimensions of Schwartz's values, suggesting that these values have richer context information about the situations. Moreover, in some cases, these LLM-generated values cover conflicting Schwartz's values together. For example, "*Personal Boundaries and Well-being*" value has high number of co-occurrences with "*Security*" and "*Self-Direction*". These two Schwartz's values conflict to each other, thus they are not categorized into the same bucket using the Schwartz's values. This shows that LLM-generated values can be applied in a more flexible way, while preserving meaningful insights from the theories supported by Schwartz's values.

C Model Implementation Details

C.1 Baseline Model Implementation

For all fine-tuned language models, we perform hyperparameter searching on training batch size and learning rates. For DistilBERT-base, the best working combination is 16 and 3.962e-5, for RoBERTa-base, it is 32 and 3.97838e-5, for DeBERTa-v3, it is 4 and 7.378218e-5. For encoder-only models, fine-tuning models for each redditor takes approximately 10 to 25 minutes on a single A30 GPU. Fine-tuning encoder-decoder models, BART and FLAN-T5, took around 15 to 30 minutes, respectively. Stretching this to 100 redditors and 5 fold experiments, running each model structure took a day to two days.

C.2 RAG-based Model Implementation

For inference using GPT-4.1 models, each of the experiment costs around USD 30, where the costs for input query takes about USD 29 and the rest is for the output which is either 0 or 1. This was based on the batch prompting, which is half of the original price.

D Model Performance

Figure 8 illustrates the macro F1 score of the redditors varying in judgment distribution balance and the number of instances. For fine-tuned models, the F1 score decreases as the number of instances decreases and the judgment patterns become more skewed. For the RAG-based models, on the other

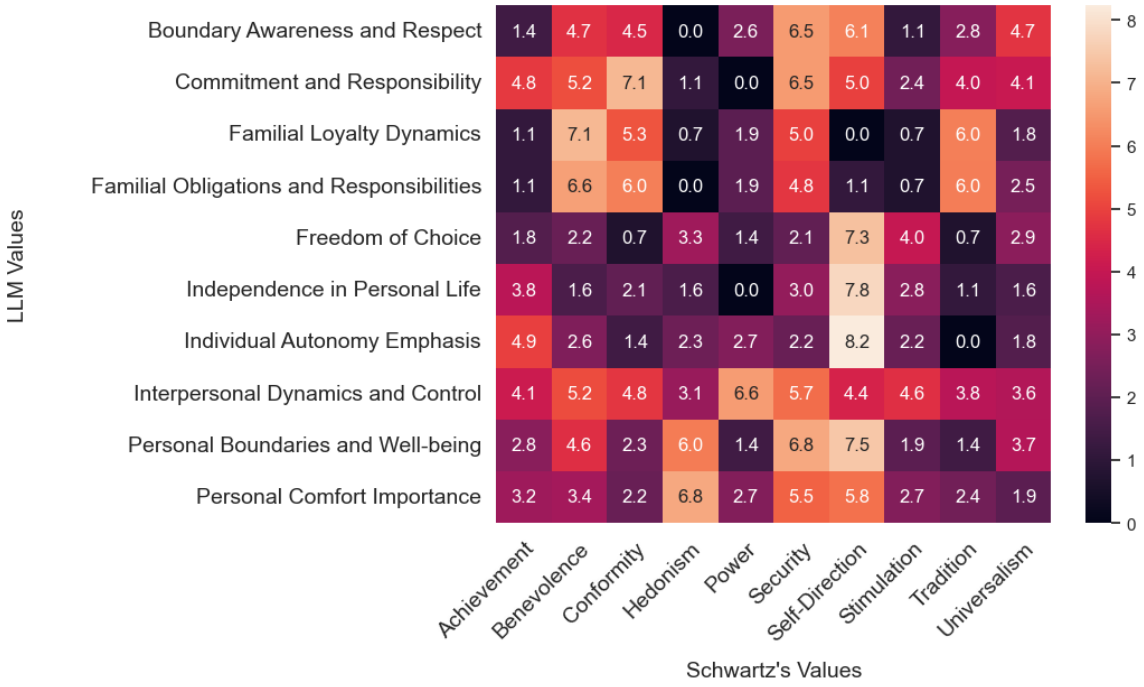

Figure 7: Mapping results of LLM-generated values to Schwartz's values. Numbers in each cell mean the log-sclaed count of the number of co-occurrences.

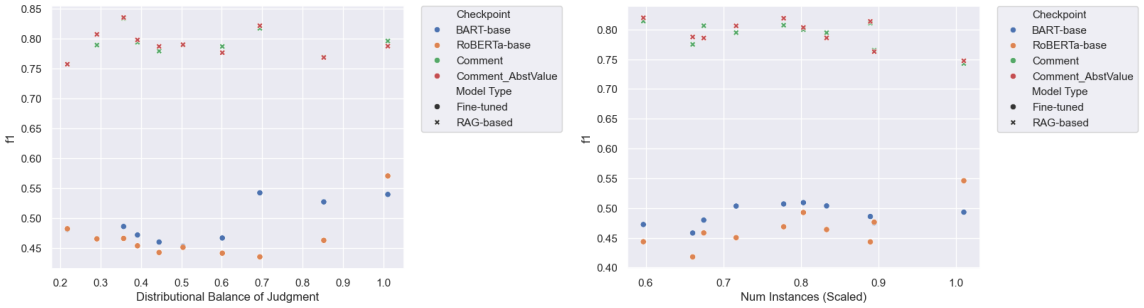

Figure 8: Macro F1 score of fine-tuned models and RAG-based models. x-axis represents the attribute of redditors. The left figure shows how balanced the redditors' judgment patterns are (more balanced if the value is closer to 1), and the right figure shows how many instances the redditors have.

hand, the F1 socre does not heavily depend on the judgment distributions and the number of instances for each redditor, as the RAG-based models can perform inference based on the few-shot examples only.