Leveraging Large Language Models to Identify the Values Behind Arguments

Senthilkumar, Rithik Appachi; Homayounirad, Amir; Siebert, Luciano Cavalcante

DOI
10.1007/978-3-031-85463-7_6

Publication date
2025

Document Version
Final published version

Published in
Value Engineering in Artificial Intelligence - 2nd International Workshop, VALE 2024, Revised Selected Papers

Citation (APA)
Senthilkumar, R. A., Homayounirad, A., & Siebert, L. C. (2025). Leveraging Large Language Models to Identify the Values Behind Arguments. In N. Osman, & L. Steels (Eds.), *Value Engineering in Artificial Intelligence - 2nd International Workshop, VALE 2024, Revised Selected Papers* (pp. 87-103). (Lecture Notes in Computer Science; Vol. 15356 LNAI). Springer. https://doi.org/10.1007/978-3-031-85463-7_6

Leveraging Large Language Models to Identify the Values Behind Arguments

Rithik Appachi Senthilkumar, Amir Homayounirad(✉), and Luciano Cavalcante Siebert

Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
{rappachisenthi,a.homayounirad,l.cavalcantesiebert}@tudelft.nl

Abstract. Human values capture what people and societies perceive as desirable, transcend specific situations and serve as guiding principles for action. People's value systems motivate their positions on issues concerning the economy, society and politics among others, influencing the arguments they make. Identifying the values behind arguments can therefore help us find common ground in discourse and uncover the core reasons behind disagreements. Transformer-based large language models (LLMs) have exhibited remarkable performance across language generation and analysis. However, leveraging LLMs in sociotechnical systems that assist with discourse and argumentation necessitates systematically evaluating their ability to analyse and identify the values behind arguments, an under-explored research direction. Using a multi-level human value taxonomy inspired by the Schwartz Theory of Basic Human Values, we present a systematic and critical evaluation of GPT-3.5-turbo in human value identification from a dataset of multi-cultural arguments, across the zero-shot, few-shot and chain-of-thought prompting strategies, carrying forward from prior research on this task which leveraged a fine-tuned BERT model. We observe that prompting strategies exhibit performance levels close to, but still behind fine-tuning for value classification. We also detail some challenges associated with value classification with LLMs, offering potential directions for future research.

Keywords: Large Language Models · Prompting · Human Values

1 Introduction

Values, as described by sociologist Robin Williams [30], are core conceptions of the desirable within every individual and society, that serve as standards or criteria to guide actions, judgments, choices and attitudes among many other aspects. A large body of work has been dedicated towards conceptualizing human values, most notably the Values Theory developed by Shalom Schwartz [20] who defines human values as "desirable, trans-situational goals, varying in importance, that serve as guiding principles in people's lives".

N. Osman and L. Steels (Eds.): VALE 2024, LNAI 15356, pp. 87–103, 2025.
https://doi.org/10.1007/978-3-031-85463-7_6

According to Milton Rokeach [17], the study of human values is not limited to one field, but is of relevance to all the sciences concerned with human behavior. One such example is the study of *argumentation.* The ability of human values to serve as standards or criteria that guide people's actions and evaluations, as noted in the works of Rokeach, Williams and Schwartz among others, manifests itself in argumentation – the values people abide by inform how they react to policies and ideas, which motivate the arguments they make either in favor of or against them. Searle [22] further noted that perfectly rational agents operating with perfect information are still capable of rational disagreement, due to inconsistencies in the values and interests they hold, despite each agent's values being rationally acceptable on their own.

Since human values motivate people's arguments on society [19], economics [1] and politics [9] among other topics in general discourse, identifying the human values behind arguments enables us to understand the "why" underlying an argument's logic [10], helping us find common ground and uncover the foundational reasons behind disagreements and conflicts. Over the past few years, the advent of large language models (LLMs), probabilistic models based on the Transformer [25] architecture, trained on natural language and capable of general-purpose language understanding and generation tasks, has led to breakthroughs in computational sentiment analysis [32], machine translation [31] and question-answering [15] among others.

However, the application of LLMs towards understanding and identifying human values from an argumentation context is relatively under-explored. Kiesel et al. [10] presented, to the best of our knowledge, the first such attempt, evaluating a fine-tuned BERT [6] model on multi-cultural arguments, revealing promising results within and across cultures. Scaling up language models across parameters, training data and training compute has conferred onto them emergent properties [26], one of which is their ability to solve tasks out of the box given a natural language instruction (i.e. a **prompt**) as input. We argue that as prompting has emerged as an accessible and straightforward avenue of interacting with LLMs, critically evaluating how prompt-based strategies fare in identifying the human values behind arguments can shed light on the benefits and drawbacks of LLMs in the context of human value identification, paving the way towards their incorporation in sociotechnical systems that assist with public discourse and argumentation.

We contribute the preliminary attempt at leveraging prompting strategies to evaluate the capability of *GPT-3.5-turbo*, the model behind the widely utilized *ChatGPT*, in identifying the human values from a multi-level human value taxonomy behind geographically diverse arguments. We observe that prompting strategies come close to, but are not capable of outperforming model fine-tuning in value identification, and that adding example demonstrations to the prompt helps improve performance. We also observe that chain-of-thought (CoT) prompting [27] using LLM-generated justifications as CoT demonstrations does not positively impact performance, possibly due to the poor quality of justifications generated.

We intend for this paper to help (1) bring to light the limitations and benefits of applying LLMs to value-driven argumentation, guiding people towards *responsibly* utilizing LLMs in this context, and (2) guide future research into potentially incorporating LLMs into the development of sociotechnical systems that align with human values and offer fresh perspectives into discourse and disagreements inspired by the human values at the root of the arguments.

2 Background

In this section, we outline the concept of *human values* and their representation in the literature. We then explore their significance in the context of *arguments*. Following this, we describe how *computational approaches* have been applied to identify the human values behind arguments. We conclude with a brief explanation of the *prompting* paradigm in LLMs, and how it can be leveraged towards this value identification task.

2.1 Human Values

The Schwartz Theory of Basic Human Values describes human values as abstract motivations that guide people's opinions, feelings and goals in life [21]. Schwartz outlines six features of all values, highlighting that values are (1) beliefs referring to (2) desirable goals motivating actions, that (3) transcend specific actions and situations, (4) serve as standards or criteria, and are (5) ordered by relative importance to one another, the relativity of importance of which (6) guides action.

Formal representations of human values have been extensively explored in the literature, both from social science [17] and argumentation research [2] standpoints. The Rokeach Value Survey (RVS) in 1973 [17] was one of the earliest thorough attempts at understanding human values, presenting a practical survey of 36 values consisting of 18 *terminal* values (referring to desired end-states of existence) and 18 *instrumental* values (referring to preferred modes of behavior). The Schwartz Theory of Basic Human Values [21] identifies ten broad human value categories, further postulating that values form a continuum of related motivations resulting in a circular structure where values closer together share a greater extent of motivational emphases.

2.2 Value-Driven Argumentation

By *arguments*, we refer to argumentative statements made in response to an idea, policy or statement where the individual making the argument does so with the goal of persuasion. Value systems have been used in formal argumentation to model audience-specific preferences, with the notion that a stronger argument is one that the audience in question reveres the values it resorts to [2,24,28]. Value classification for arguments entails identifying the human values that form the motivational basis behind the arguments being made. Kiesel et al. [10] argue that processing and analyzing the human values behind arguments introduces a new outlook into argumentation, emphasizing the "why" underlying the argument's logic.

Identifying human values behind arguments is often not straightforward, since in some cases the values that guide someone to express an opinion may be implicit (i.e. not directly specified in the opinion), or context-specific. To further describe the context-specificity of values, we turn to Liscio et al. [13] who define a context-specific value as "a value that is applicable and defined specifically within a context". For instance, the value of *privacy* is relevant to the context of "information control in social media", but *physical health* may not be so relevant. However, the opposite is observed if we switch the context to "health effects of computer use".

2.3 Computational Approaches Towards Identifying the Values Behind Arguments

While human values have been accounted for in formal argumentation frameworks since the early 2000 s [2], Kiesel et al. [10] presented the first attempt at automatic identification of values behind arguments, using a fine-tuned multi-label BERT-base [6] LLM to perform value classification on a multicultural set of arguments.

Since then, a multitude of natural language processing (NLP) driven techniques have been leveraged towards human value identification. SemEval-2023, an international workshop on semantic evaluation, consisted of ValueEval'23 [11], a task on value identification behind arguments where teams prepared models that were tasked with identifying whether a particular value applied to an argument. ValueEval'23 focused exclusively on the 20 Level 2 value categories and used a multi-sourced dataset of over 9000 arguments, using the same human value taxonomy and argument structure that we will in this paper. Most of the submitted approaches relied on leveraging transformer-based models [25], formulating the task as direct classification based on the provided labels.

2.4 Large Language Models and Prompting

Prior attempts [10,11] at automatic identification of human values behind arguments have largely relied on fine-tuning LLMs. However, LLMs also possess interesting *emergent abilities*, as a result of scaling smaller LMs across the factors of *number of model parameters*, *amount of training data utilized* and *the extent of computations performed* [26]. One such example is the *prompting* paradigm, as popularized by Brown et al. through GPT-3 [4], in which a carefully-worded prompt is provided to the pre-trained LLM which completes the response with no additional training required. Prompting strategies range from *zero-shot prompting* in which only the instruction is provided in natural language format to the LLM, to *few-shot prompting* which further augments the prompt with a few input-output examples as a way of conditioning the model to the task [16].

Chain-of-thought prompting, providing step-by-step demonstrations for examples, has been shown to improve general reasoning abilities in sufficiently large language models [27].

Liang et al. [12] performed a holistic evaluation of language models, showing that they exhibit a high level of performance in sentiment analysis, knowledge, text summarization, information retrieval and question answering among a variety of tasks. Chae and Davidson [5] applied a multitude of strategies on LLMs, ranging from zero-shot prompting to model fine-tuning, demonstrating their capability in classification of opinions in a political context.

3 Webis-ArgValues-22: A Dataset of Cross-Cultural Arguments

The Webis-ArgValues-22 dataset introduced by Kiesel et al. [10] consists of 5270 natural language arguments, and is composed of four parts: *Africa* (50 arguments), *China* (100 arguments), *India* (100 arguments) and *USA* (5020 arguments). Each of these arguments were annotated by three crowdworkers for all 54 Level 1 values, and labels for the broader categorizations of values were derived from them. The dataset, taxonomy description and annotation interface can be found online as Webis-ArgValues-22[1]. The dataset is partitioned into three parts: **train**, **validation** and **test**. The train and validation parts of the dataset only consist of arguments from the US, unlike the test part which contain arguments from all four cultures.

3.1 Human Value Taxonomy

Drawing largely upon the refined theory proposed by Schwartz et al. [21], but also incorporating values identified in the Rokeach Value Survey (RVS) [17], the Life Values Inventory (LVI) [3] and the World Values Survey (WVS) [8], Kiesel et al. [10] proposed a multi-level taxonomy of values, consisting of 54 basic values (called Level 1 values) from the social sciences further categorized into higher, broader levels inspired by the Schwartz theory. More specifically, Level 2 values aggregate some of the Level 1 values into 20 broader categories, which are further aggregated into higher-order values in Level 3, as proposed by Schwartz. Furthermore, the Level 3 is aggregated into two possible dichotomies: Level 4A focusing on personal versus social, and Level 4B focusing on promotion of growth versus self-protection in Level 4B). A more detailed explanation of the human value taxonomy can be found in [10]. A visual representation of the value levels can be found in Fig. 1.

3.2 Argument Structure

Each argument in the dataset consists of three parts:

[1] https://github.com/webis-de/ACL-22.

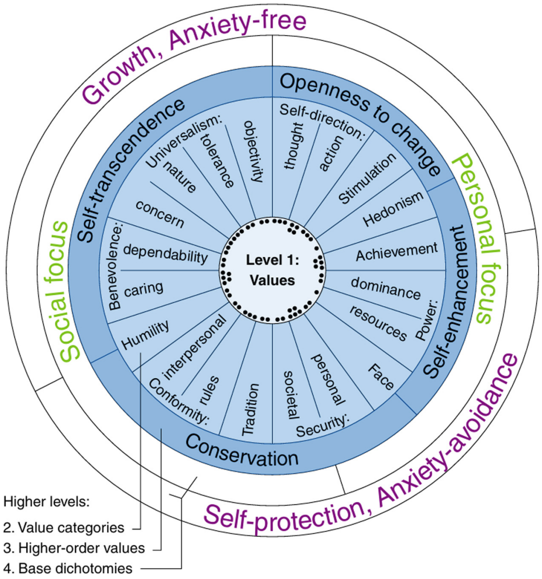

Fig. 1. The consolidated multi-level value taxonomy as developed by Kiesel et al. [10], adapted from the Schwartz Theory of Basic Human Values [21]

- **Conclusion:** The policy, idea or statement towards which the opinion is made.
- **Stance:** The opinion's stance with respect to the conclusion. Either *in favor of* or *against*.
- **Premise:** The opinion posed, explaining why the person expressed their specified stance towards the conclusion.

For example, for the idea (here, **conclusion**) "We should protect our privacy in the Internet age", someone can express an opinion with a favorable **stance** towards it, stating that "The leaked personal information will be defrauded by fraud gangs to gain trust and carry out fraudulent activities" which here is the **premise**.

4 Methodology

We seek to explore how accurately an LLM identifies the values behind natural language arguments, from individual human values to broader value categories. While previous computational attempts at value classification have largely relied on performing classification using fine-tuned language models [11], we aim to leverage the pre-existing knowledge and language understanding capabilities of a large language model, relying on a variety of prompting strategies to guide the model towards *generating* the values and value categories it believes the argument indicates, parsing the response to obtain predicted labels.

All the source code, including the exact prompts we used and the experiments conducted for all the prompting strategies we considered can be found online on GitHub[2]

This section will describe the model, the prompting strategies leveraged in the experiment, and the baselines and metrics used to assess its performance.

4.1 Model

We introduce a model-agnostic methodology to evaluate the capability of LLMs in identifying the human values behind arguments, solely relying on prompting strategies. For the experiments, we use the GPT-3.5-turbo[3] LLM developed by OpenAI for our experiments, due to its widespread use in the public domain (either through the freely-available ChatGPT[4] web platform, or through its API access). Future work can make use of our methodology to experiment with other LLMs.

GPT-3.5-turbo is derived from GPT-3 [4], an autoregressive LLM consisting of 175 billion parameters. GPT-3 was trained on 499 billion tokens[5] of text, consisting of CommonCrawl[6] WebText [16], English Wikipedia[7], and the Books1 and Books2 corpora.[8] Unlike GPT-3, GPT-3.5-turbo has been fine-tuned through the Reinforcement Learning from Human Feedback (RLHF) process, in order to align it to human preferences [33]. The exact number of parameters, as well as training data used for GPT-3.5 models has been undisclosed by OpenAI.

Critically evaluating the ability of LLMs in identifying human values is important to ensure they are *responsibly* leveraged to develop human-aligned AI systems, and that users are aware of its limitations and potential risks.

To carry out our experiments, we relied on the model's API, using LangChain[9] to interface with it.

4.2 Prompting Techniques

In the context of LLMs, a **prompt** is a carefully-worded instruction provided to the model as input, to guide it towards providing a *response* that satisfies the instruction.

[2] https://github.com/rithik83/LLM-values.

[3] https://platform.openai.com/docs/models/gpt-3-5-turbo.

[4] https://chatgpt.com/.

[5] Tokens are common sequences of characters found in a set of text. OpenAI's large language models process text as tokens. More information can be found in https://platform.openai.com/tokenizer.

[6] A free, open repository of web crawl data. Learn more from https://www.commoncrawl.org.

[7] https://www.wikipedia.org.

[8] Two internet-based books corpora.

[9] A framework for interacting with LLMs and developing applications utilizing LLMs; https://www.langchain.com/.

In our prompts, we provide the LLM with the list of all human values for a specific level of categorization of values, an argument's conclusion, stance and premise framed as a natural language sentence, and an explicit instruction to identify the values from the list that motivate the argument presented. In example-based prompting strategies, we provide a set of example arguments and their values to further condition the model to the value classification task.

For example-based prompting strategies (few-shot and chain-of-thought prompting), we sample 20 examples randomly from the *train* part of the dataset. In case some values are unrepresented among the selected examples, we further augment the example set with arguments that represent said values, thereby ensuring that every few-shot/CoT prompt consist of example demonstrations that cover all values under consideration.

Zero-Shot Prompting. In this method, we explore the capability of the model to use only its pre-trained knowledge to identify the human values behind arguments. The zero-shot prompting pipeline is illustrated in Fig. 2.

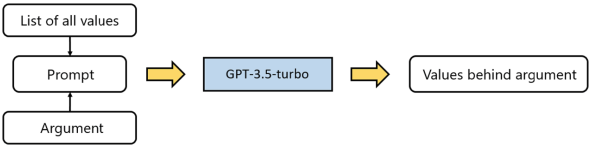

Fig. 2. The zero-shot prompting pipeline

The argument's *conclusion*, *stance* and *premise* are framed as a single sentence. The LLM is instructed to generate the values it identifies (from the complete list of values provided in the prompt) as a semicolon separated list, which is then parsed to obtain the labels predicted. One important feature of the prompt is the repeated instruction to be *selective* and *precise*. This instruction was added to guide the LLM into choosing only those values it perceives as clearly forming the basis for the argument made.

Since zero-shot prompting relies solely upon what the LLM has learned from its pre-training and RLHF fine-tuning process, it serves as a reference point for prompt-based strategies to enable comparisons with few-shot and chain-of-thought prompting.

Few-Shot Prompting. In addition to providing a natural language description of the task, few-shot prompting provides a few demonstrations of the task at inference time as a way of conditioning the model [16]. Examples are typically provided as pairings of context and desired completions, and a final example of context for which the model is expected to perform the completion on the basis of both its pre-trained knowledge and its understanding of the prior examples.

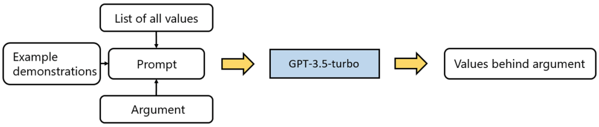

Fig. 3. The few-shot prompting pipeline

Few-shot prompting reduces the need for task-specific data as compared to model fine-tuning. The few-shot prompting pipeline we used is illustrated in Fig. 3.

The few-shot prompt used is very similar in structure to the zero-shot prompt, as we intended to measure the effect of adding example demonstrations. Examples are included in the Q+A format with the answer portions of the examples filled in with ground truth values. The last Q+A entry is the unseen argument for which the LLM is tasked with identifying the motivating human values.

We hypothesized that few-shot prompting would exhibit stronger performance as compared to zero-shot prompting, since few-shot prompting provides the added benefit of example demonstrations to the model during inference time, thereby conditioning the model to the value classification task.

Chain-of-Thought (CoT) Prompting. Chain-of-Thought (CoT) prompting, developed by Wei et al. [27] is a variant of the few-shot prompting technique where each example demonstration is further enhanced with a chain of thought - a coherent series of intermediate reasoning steps that lead to the final answer. The authors demonstrated that these chain-of-thought demonstrations invoked reasoning abilities in sufficiently large language models, improving performance on commonsense reasoning tasks.

In order to explore the effect of reasoning-enriched prompting in the task of value classification, we decided to include CoT prompting. However, the dataset we considered for this paper did not have any form of reasoning or justification for the annotated value labels. In order to overcome this issue, we developed a two-step pipeline to carry out CoT prompting for each argument:

- **Prompt 1 - Generating justifications for examples:** Here, for all the example arguments chosen, we provide the LLM with the example and all its ground truth values, prompting the LLM to generate brief justifications for why each of the values holds for the argument presented.
- **Prompt 2 - Using generated justifications as CoT demonstrations for the unseen argument:** We perform a second prompt to the LLM that contains all the *example arguments*, their *values* and for each value its LLM-generated *justification* (generated in Prompt 1), as well as the final unseen argument for which the LLM is tasked with identifying the values. The generated justifications act as CoT demonstrations to assist the LLM in identifying the values for the unseen argument.

The chain-of-thought pipeline is illustrated in Fig. 4.

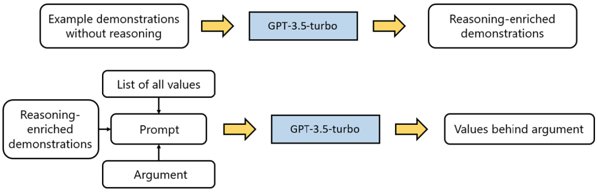
Example demonstrations without reasoning
GPT-3.5-turbo
Reasoning-enriched demonstrations
List of all values
Reasoning-enriched demonstrations
Prompt
Argument
GPT-3.5-turbo
Values behind argument

Fig. 4. The chain-of-thought prompting pipeline. Note the two prompts corresponding with the two-step procedure

Note that CoT prompting was used to evaluate performance on identifying only the Level 1 and 2 values – the individual values and the lowest level value categorization respectively. This decision was grounded in a few key considerations regarding the complexity of these levels. Level 1 consists of 54 values and Level 2 aggregates them into 20 value categories. These levels capture a rich landscape of values, making them ideal to craft nuanced justifications. In contrast, levels 3, 4A and 4B represent broader categorizations that are less granular, hence less informative to generate detailed, specific justifications. Moreover, by focusing on Level 1 and 2 values, the LLM is encouraged to generate justifications that are closely aligned with specific values or more finely grained categories. This granularity allows for more precise and contextually relevant explanations, which are essential for accurately evaluating the model's ability to discern subtle differences between similar values.

We hypothesized that the chain-of-thought prompting strategy would perform the best out of the three prompting strategies, since enriching the example demonstrations with logical reasoning would better equip the LLM towards identifying the values behind the unseen argument it is tested on.

4.3 Evaluation Baseline and Metrics

Our choices for baseline and evaluation metrics were inspired by the decisions made by Kiesel et al. [10], who presented the first attempt at computational value classification using the dataset under consideration, in order to be able to establish a comparison between their fine-tuning approach and our prompting approach.

Metrics: We use label-wise F_1-score, and its mean over all labels (macro-average), as well as its constituents' precision and recall. Macro-averages were used to give the same weight to all values.

Baseline: We use 1-Baseline as the evaluation baseline. It classifies each argument as belonging to all values, which results in a perfect recall score of 1. Due to its high recall, the 1-Baseline classifier achieves at least as high, if not higher in most cases, F_1-scores as label-wise random guessing based on the label frequency.

Table 1. Macro precision (P), recall (R) and F_1-score (F_1) on the USA test set over all labels by level, extended from the results obtained by Kiesel et al. [10] using BERT [6] and an SVM. Highest precision, recall and F_1-score for each level marked in bold.

Model	Level 1			Level 2			Level 3			Level 4A			Level 4B		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
GPT-3.5-turbo (Zero-shot)	0.26	0.30	0.23	0.35	0.31	0.30	**0.67**	0.57	0.59	**0.92**	0.53	0.67	0.92	0.46	0.54
GPT-3.5-turbo (Few-shot)	0.15	0.53	0.21	0.39	0.41	**0.34**	0.66	0.66	0.66	0.87	0.94	0.90	0.92	0.84	0.87
GPT-3.5-turbo (CoT)	0.31	0.22	0.21	**0.40**	0.30	0.32	-	-	-	-	-	-	-	-	-
BERT	**0.40**	0.19	**0.25**	0.39	0.30	**0.34**	0.65	0.78	0.71	0.89	0.96	**0.92**	0.92	**1.00**	**0.96**
SVM	0.21	0.19	0.20	0.30	0.30	0.30	0.66	0.68	0.67	0.88	0.89	0.88	**0.93**	0.90	0.92
1-Baseline	0.08	**1.00**	0.16	0.18	**1.00**	0.28	0.60	**1.00**	**0.75**	0.85	**1.00**	**0.92**	0.92	**1.00**	**0.96**

5 Results

In our evaluations, we used the same partitions of the dataset as Kiesel et al. [10] in their evaluation of value classification with BERT, so as to reliably compare the results we obtained using different measures (their *fine-tuning* approach versus our *prompting* approach). We ran our evaluations on the test set, and although we did not train or fine-tune our model, we still leveraged the train set to select example demonstrations for the few-shot and chain-of-thought prompting experiments.

5.1 Results on the US Part

The US part of the dataset is the most significant in terms of the number of entries as compared to the remaining parts (5020 arguments out of 5270 in the total dataset, and 503 out of 753 in the test dataset). The train-validation-test split of the dataset for the US part was done on the basis of unique conclusions, as a result of which the test set consisted of 7 conclusions that were not present in the train and validation sets. While this method of splitting was considered crucial by Kiesel et al. [10] since they wanted to test whether classifiers generalized to unseen conclusions, it unfortunately led to different value distributions in the different sets. Regardless, we persisted with the same split in the interest of being able to compare our results. Table 1 shows the results averaged across all labels.

Value Levels 1 and 2. For Levels 1 and 2, all the prompting strategies perform better than the baseline according to F_1-score, but still exhibit a lower performance than the fine-tuned BERT by Kiesel et al. [10], with the exception of few-shot prompting for Level 2 which performs equally as well as BERT in terms of macro F_1 score. At Level 1, the most granular level, it is surprising to note that example-based prompting strategies (few-shot and CoT) perform worse than zero-shot prompting. This could be in part because the model generated responses that mimic the labels provided in the example arguments, essentially

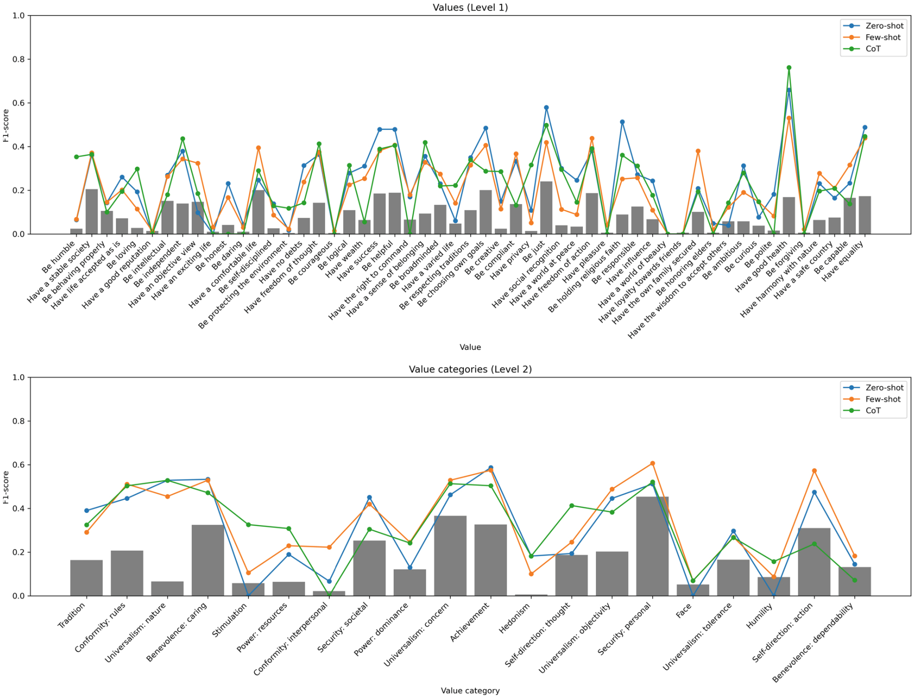

Fig. 5. F$_1$-scores per Level 1 value (top) and Level 2 value category (bottom) across zero-shot, few-shot and CoT prompting strategies, in the USA test set. Grey bars indicate the frequency distribution of each value in the set

overfitting to the examples and therefore generalizing poorly to the unseen argument. Moreover, since we randomly chose examples from the train set, the quality of the example set shortlisted may have been suboptimal and not representative. A potential strategy that may overcome this issue is selecting examples based on semantic similarity to the unseen argument, so that the model is equipped with better example demonstrations. Chain-of-Thought prompting lead to higher precision for both levels 1 and 2, possibly due to the value justifications for example influencing the model to be even more selective in identifying values behind the unseen argument. Few-shot (without CoT) lead to the highest recall values.

Value Levels 3, 4A and 4B. For the higher levels, both zero-shot and few-shot prompting exhibit worse performances in comparison to both the baseline and BERT, in terms of the F$_1$-score. In particular, the zero-shot prompting strategy performs significantly worse in levels 4A and 4B (with F$_1$-scores of 0.67 and 0.54 respectively as compared to 0.92 and 0.96 for both BERT and 1-Baseline respectively). During evaluations with zero-shot prompting for levels 4A and 4B (both the base dichotomies with two options each), it was observed that the LLM almost always opted for exactly one of the two options, despite most arguments having been assigned both labels of both base dichotomies. This interpretation of the value classification task as a one-or-the-other problem by the model despite no such specific instruction in the prompt, resulted in the low values for recall which in turn affected the F$_1$-scores. Few-shot prompting did not suffer from this problem, since the example demonstrations reflected the reality of most arguments having been labeled with both categories for both the base dichotomies.

Results per Value for Levels 1 and 2. In addition to measuring the macro-averaged precision, recall and F$_1$-score, we also measure F$_1$-scores per Level 1 value and Level 2 value category, across all our 3 prompting approaches. Figure 5 demonstrates the results, with the grey bars signifying the frequency of each value/value category in the US test set.

We observe that none of the prompting strategies dominates the others across all, or even most, Level 1 values. Moreover, the LLM achieves considerably high F$_1$-scores across all prompting strategies for several values and value categories, most notably for the value *Have good health*, and the value category *Security: personal* that contains it. Other value categories for which all the three methods achieved an F$_1$-score ≥ 0.4 are *Conformity: rules*. *Universalism: nature*, *Benevolence: caring*, *Universalism: concern* and *Achievement*. These results are similar to those obtained by Kiesel et al. [10].

Table 2. Macro F$_1$-score on each test set over all labels by level, extended from the results obtained by Kiesel et al. [10] using BERT [6] and an SVM. Highest score for each region in each level marked in bold.

Model	Level 1				Level 2				Level 3				Level 4A				Level 4B			
	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA
GPT-3.5-turbo (Zero-shot)	0.16	**0.22**	0.26	0.23	0.23	0.28	0.24	0.30	0.51	0.62	0.59	0.59	0.51	0.65	0.65	0.67	0.58	0.52	0.47	0.54
GPT-3.5-turbo (Few-shot)	0.20	0.19	0.19	0.21	0.29	0.28	0.28	**0.34**	0.60	**0.68**	0.63	0.66	0.80	0.83	0.79	0.90	0.84	0.87	0.83	0.87
GPT-3.5-turbo (CoT)	**0.28**	**0.22**	0.29	0.21	0.31	0.23	0.23	0.32	-	-	-	-	-	-	-	-	-	-	-	
BERT	0.20	0.21	**0.30**	**0.25**	**0.38**	**0.37**	**0.41**	**0.34**	0.60	**0.68**	**0.71**	0.71	**0.82**	**0.88**	**0.81**	**0.92**	**0.92**	**0.91**	**0.90**	**0.96**
SVM	0.21	0.21	0.25	0.20	0.29	0.30	0.27	0.30	0.53	0.57	0.57	0.67	0.80	0.82	0.74	0.88	0.90	0.87	0.87	0.92
1-Baseline	0.16	0.13	0.12	0.16	0.27	0.23	0.21	0.28	**0.63**	0.65	0.62	**0.75**	0.80	**0.88**	0.79	**0.92**	**0.92**	**0.91**	**0.90**	**0.96**

5.2 Results Across Culture

Table 2 demonstrates how GPT-3.5-turbo with the three prompting strategies performed across the four parts of the test set: Africa, China, India and the US, comparing them with the results Kiesel et al. [10] obtained using BERT, a linear SVM and the baseline.

Across all the levels, we observe that the prompting strategies applied do not result in an improvement in performance over BERT, with the exception of chain-of-thought prompting applied to arguments from Africa and China for Level 1 values.

However, an important point to note is that arguments from Africa, China and India are significantly underrepresented relative to arguments from the US (50, 100 and 100 arguments for Africa, China and India respectively as compared to 503 for the US in the test set), which makes carrying out detailed analyses difficult. We still reported our results to establish a comparison with the approach of fine-tuning BERT. Perhaps gathering more arguments from the non-USA regions would have allowed us to report more conclusive results across cultures.

6 Discussion and Limitations

Although we present a detailed methodology for assessing the capability of Large Language Models (LLMs) in identifying the human values behind arguments using prompting techniques, we recognize certain limitations associated with our preliminary experimentation with GPT-3.5-turbo across the zero-shot, few-shot and chain-of-thought prompting strategies. In this section we will highlight them, offering directions that can be considered in future research.

Sensitivity to prompts: Our approaches leveraged the ability of LLMs to generate responses to carefully-worded natural language instructions called *prompts*, parsing the results to obtain the predicted labels. We designed a standard prompt that contained all relevant information – the complete list of values to consider, the argument phrased as a natural language sentence, the instruction to identify its values, and example demonstrations for few-shot/CoT prompts. However during our experimentation we observed, as expected, that the LLM's responses were quite sensitive to the prompt, including the placing of certain sentences, words and their phrasing. The field of prompt engineering, the strategic design of task-specific instructions to guide the LLM into generating desirable outputs, has seen a great deal of research recently [7,18,29]. While our intention was to provide a starting point for exploration into prompt-based strategies for value classification using three strategies, we encourage researchers to explore the application of more prompting strategies and best practices towards this task.

Subjectivity of human values: To carry out our experimentation, we leveraged a dataset designed by Kiesel et al. [10] which consisted of arguments manually annotated by three crowdworkers for all 54 Level 1 values. While the annotations were considered ground truth values for the arguments, a large body of philosophical work has been devoted to exploring the idea that human values may be subjective [14,23]. Moreover, Liscio et al. [13] also established the idea that values may be context-specific. The annotated ground truth labels represent the aggregated opinions of the crowdworkers, but it is possible that other annotators may have labeled different values for the same argument. An interesting direction for future research could be the exploration of how this subjectivity impacts LLM outputs, and how LLMs can be utilized for highly subjective tasks like these.

Imbalance in the Number of Arguments Across Cultures: We reported and discussed results obtained on the US test set, and results across cultures to present a complete picture of LLM performance. However, due to the stark imbalance between the number of arguments from the US and the remaining regions, it was not possible to carry out analyses of non-US arguments with the same rigour as US arguments.

LLM-Generated Justifications for Chain-of-Thought: In our chain-of-thought prompting strategy, we relied on the LLM to generate justifications for value classifications for the example demonstrations due to the lack of such a reasoning chain provided in the dataset. We recognize this may have resulted in suboptimal chains of thought being generated. Perhaps human-annotated chains of thought/reasoning could improve CoT prompting performance.

Our contribution through this paper is a systematic and critical evaluation of the GPT-3.5-turbo LLM in identifying the values behind arguments across cultures. Unlike previous attempts [10,11] that primarily relied on model fine-tuning, we applied three prompting strategies – zero-shot, few-shot and Chain-of-Thought [27] prompting. As prompting continues to emerge as a popular mechanism through which people interact with LLMs, our work attempts to demonstrate how effective it is in extracting the values behind arguments, thereby encouraging a more responsible approach towards prompting LLMs in this context.

7 Conclusion

Human values serve as criteria that guide people's actions, judgments and evaluations [21,30], and are central to argumentation and discourse as they inform people's viewpoints and arguments. We present an evaluation of GPT-3.5-turbo, an LLM widely utilized by the public, in identifying the values behind arguments using zero-shot, few-shot and Chain-of-Thought (CoT) [27] prompting strategies, using a multi-cultural dataset of arguments and a multi-level value taxonomy derived from [10], inspired by the Schwartz Theory of Basic Human Values [21]. Our experiments indicate that prompting strategies exhibit performance levels close to, but still behind model fine-tuning in terms of the macro-averaged F_1-score, and that performance levels vary across values and value categories themselves. Our work exposes directions for further research into developing tehcniques for value-driven argumentation that leverage the capabilities of LLMs, keeping in mind their benefits and drawbacks to use them in a responsible, effective and inclusive manner.

Acknowledgments. The participation of L.C. Siebert in this research was partially supported by TAILOR, a project funded by the EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Anderson, E.: Value in Ethics and Economics. Harvard University Press (1995)
2. Bench-Capon, T.J.: Persuasion in practical argument using value-based argumentation frameworks. J. Log. Comput. **13**(3), 429–448 (2003)
3. Brown, D., Crace, R.K.: Life values inventory: facilitator's guide. Williamsburg, VA (2002)
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Chae, Y., Davidson, T.: Large language models for text classification: from zero-shot learning to fine-tuning. Open Science Foundation (2023)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Giray, L.: Prompt engineering with ChatGPT: a guide for academic writers. Ann. Biomed. Eng. **51**(12), 2629–2633 (2023)
8. Haerpfer, C., et al.: World values survey: round seven-country-pooled datafile. Madrid, Spain, Vienna, Austria: JD Systems Institute & WVSA Secretariat **7**, 2021 (2020)
9. Inglehart, R.F., Basanez, M., Moreno, A.: Human Values and Beliefs: A Cross-Cultural Sourcebook. University of Michigan Press (1998)
10. Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., Stein, B.: Identifying the human values behind arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4459–4471 (2022)
11. Kiesel, J., et al.: SemEval-2023 task 4: ValueEval: identification of human values behind arguments. In: Ojha, A.K., Doğruöz, A.S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., Sartori, E. (eds.) Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pp. 2287–2303. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.313, https://aclanthology.org/2023.semeval-1.313
12. Liang, P., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
13. Liscio, E., van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K.: Axies: identifying and evaluating context-specific values. In: AAMAS, pp. 799–808 (2021)
14. Mackie, J.L.: The subjectivity of values. Essays on moral realism pp. 95–118 (1988)
15. Petroni, F., et al.: Language models as knowledge bases? ArXiv preprint arXiv:1909.01066 (2019)
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
17. Rokeach, M.: The nature of human values. Free Press (1973)
18. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: techniques and applications. arXiv preprint arXiv:2402.07927 (2024)
19. Sayer, A.: Why Things Matter To People: Social Science, Values and Ethical Life. Cambridge University Press (2011)
20. Schwartz, S.H.: Basic human values: an overview (2006)
21. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online Readings Psychol. Cult. **2**(1), 11 (2012)

22. Searle, J.R.: Rationality in action. MIT press (2003)
23. Stroud, B.: The study of human nature and the subjectivity of value. The Tanner Lectures on Human Value (1988)
24. Teze, J.C.L., Perelló-Moragues, A., Godo, L., Noriega, P.: Practical reasoning using values: an argumentative approach based on a hierarchy of values. Ann. Math. Artif. Intell. **87**(3), 293–319 (2019). https://doi.org/10.1007/s10472-019-09660-8
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
26. Wei, J., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
27. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
28. van der Weide, T.L., Dignum, F., Meyer, J.J.C., Prakken, H., Vreeswijk, G.A.: Practical reasoning using values: giving meaning to values. In: Argumentation in Multi-Agent Systems: 6th International Workshop, ArgMAS 2009, Budapest, Hungary, May 12, 2009. Revised Selected and Invited Papers 6, pp. 79–93. Springer (2010)
29. White, J., et al.: A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382 (2023)
30. Williams, R.M.J.: The concept of values. In: Sills, D.L. (ed.) International Encyclopedia of the Social Sciences, vol. 16. Macmillan and Free Press, New York (1968)
31. Zhang, B., Haddow, B., Birch, A.: Prompting large language model for machine translation: a case study. In: International Conference on Machine Learning, pp. 41092–41110. PMLR (2023)
32. Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L.: Sentiment analysis in the era of large language models: a reality check. arXiv preprint arXiv:2305.15005 (2023)
33. Ziegler, D.M., et al.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)