EAVIT: Efficient and Accurate Human Value Identification From Text Data via LLMs

Wenhao Zhu[1] , **Yuhang Xie**[1] , **Guojie Song**[*1] and **Xin Zhang**[2]
[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University
[2]School of Psychological and Cognitive Sciences, Peking University
{wenhaozhu, gjsong, zhang.x}@pku.edu.cn, yuhangxie@stu.pku.edu.cn

Abstract

The rapid evolution of large language models (LLMs) has revolutionized various fields, including the identification and discovery of human values within text data. While traditional NLP models, such as BERT, have been employed for this task, their ability to represent textual data is significantly outperformed by emerging LLMs like GPTs. However, the performance of online LLMs often degrades when handling long contexts required for value identification, which also incurs substantial computational costs. To address these challenges, we propose EAVIT, an efficient and accurate framework for human value identification that combines the strengths of both locally fine-tunable and online black-box LLMs. Our framework employs a value detector—a small, local language model—to generate initial value estimations. These estimations are then used to construct concise input prompts for online LLMs, enabling accurate final value identification. To train the value detector, we introduce explanation-based training and data generation techniques specifically tailored for value identification, alongside sampling strategies to optimize the brevity of LLM input prompts. Our approach effectively reduces the number of input tokens by up to 1/6 compared to directly querying online LLMs, while consistently outperforming traditional NLP methods and other LLM-based strategies.

1 Introduction

The recent advent of Large Language Models (LLMs)[1], including the Generative Pre-trained Transformer (GPT) models [Brown *et al.*, 2020; OpenAI, 2023] and Llama [Touvron *et al.*, 2023a], has marked a pivotal advancement in natural language processing (NLP). One notable application of LLMs is the discovery and identification of human values from text data, such as arguments [Kiesel *et al.*, 2022; Kiesel *et al.*, 2023]. This task holds significant potential across various domains, supporting value alignment for LLMs [Yao *et al.*, 2023], value-based argument models [Atkinson and Bench-Capon, 2021], and computational psychological studies [Alshomary *et al.*, 2022].

Before the era of LLMs, identifying human values from text data in computational linguistics was typically framed as a multi-label classification problem and addressed using machine learning and NLP models like BERT [Devlin *et al.*, 2019]. However, these models are now being outperformed by modern LLMs in terms of their general text comprehension capabilities. Since most LLMs are accessible only through black-box APIs and fine-tuning them is computationally and economically inefficient, the standard approach for utilizing LLMs in value identification involves constructing tailored prompts for direct queries. For LLMs to effectively identify values from text data, they must first *learn the value system definition*, akin to how humans do. This definition is typically presented as a long context—for instance, the basic Schwartz values definition spans approximately 2.5k tokens [Schwartz, 2012]. Including such lengthy definitions in the input prompt is not only costly but has also been shown to degrade performance in context-heavy tasks [Liu *et al.*, 2023], a finding corroborated by our experiments.

Given these challenges, to better leverage the capabilities of LLMs, we propose EAVIT, a framework for **E**fficient and **A**ccurate Human **V**alue **I**dentification from **T**ext data. EAVIT begins with a value detector (a tunable local language model) that generates initial value estimations. These estimations are then used to construct a concise input prompt for the LLM, enabling accurate final value identification. The value detector identifies values that are most certainly related and those that are most certainly unrelated, leaving only values with uncertain relevance to be resolved by the LLM. This approach effectively reduces the number of values requiring explicit definition in the input context, thereby shortening the overall context length. To enable the value detector to learn the cognitive logic underlying value identification, we employ an explanation-based fine-tuning method, training the model to reflect on the definitions of values throughout the learning process. To address issues such as insufficient training data and imbalanced class distributions, we draw inspiration from methods like Self-Instruct [Wang *et al.*, 2022] and utilize diverse data generation techniques via LLMs to create high-

*Corresponding Author

[1]Please refer to http://arxiv.org/abs/2505.12792 for an extended version of this paper.

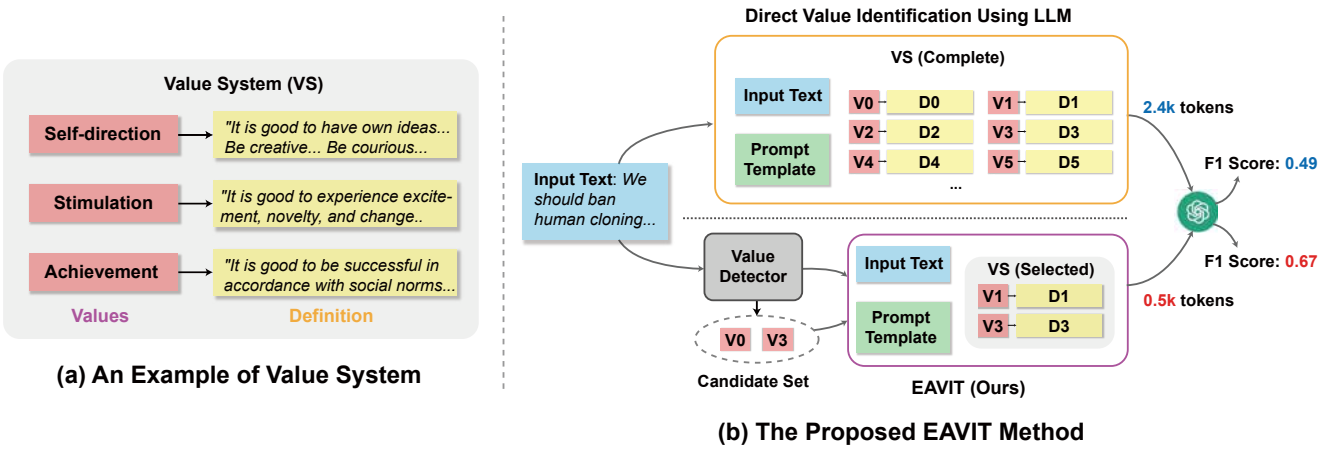

Figure 1: An illustration of (a) human value system and (b) the proposed EAVIT method compared with directly using LLMs.

quality training datasets. Additionally, to ensure an optimal candidate set, we perform multiple sampling rounds to identify the most relevant candidate values.

The proposed EAVIT framework not only achieves state-of-the-art performance but also significantly reduces the token cost for inference—down to nearly $\frac{1}{6}$ of the tokens required when directly using LLMs. This makes it a promising and cost-effective solution for large-scale value identification tasks.

Our contributions can be summarized as follows:

- We introduce EAVIT, a novel framework for identifying human values from text data. This methodology efficiently leverages the power of LLMs, offering accurate and cost-effective value identification results.
- We employ a diverse set of training strategies for the value detector (local LM in EAVIT), including explanation-based fine-tuning and data generation techniques. Additionally, we develop sampling and prompt-generation methods to create concise and effective input prompts for LLMs.
- Our approach achieves state-of-the-art performance compared to traditional NLP methods and direct LLM-based strategies. Furthermore, it significantly reduces inference costs to as low as $\frac{1}{6}$ of the tokens required by conventional LLM approaches, making it highly scalable for tasks such as LLM alignment and psychological analysis.

2 Related Works

Human Value Theories A detailed review of human value theories in psychology can be found in Appendix. In our paper, following existing works [Kiesel *et al.*, 2023] in computational linguistics, we primarily adopt Schwartz's Theory of Basic Values [Schwartz, 2012] as the basic value system for value identification, which has been applied in multiple fields including economics [Ng *et al.*, 2005] and LLMs [Miotto *et al.*, 2022; Fischer *et al.*, 2023]. Meanwhile, it should be noted that our method is applicable to any completely defined value system (such as the extended Schwartz value system or those with values set for language models).

Value Identification from Text Data. In NLP, the identification of human values from text data can be perceived as a multi-label classification or regression task described by complex task definition. Recent key related works include [Qiu *et al.*, 2022; Kiesel *et al.*, 2022; Kiesel *et al.*, 2023; Ren *et al.*, 2024]. In [Qiu *et al.*, 2022], simple social scenario descriptions from [Forbes *et al.*, 2020] were selected and annotated using Schwartz's value taxonomy, establishing the ValueNet dataset for value modeling in language models. [Kiesel *et al.*, 2022] was the first to systematically establish the task of identifying hidden human values from argument data and built a dataset, Webis-ArgValues-22, of 5k size derived from social network data and annotated with Schwartz values by humans. [Kiesel *et al.*, 2023] extended the work of [Kiesel *et al.*, 2022] with dataset Touché23-ValueEval, expanding the dataset size to 9k and held a public competition at the ACL2023 workshop. [Yao *et al.*, 2023] also proposes FULCRA dataset (currently unavailable) that labels LLM outputs to Schwartz human values. Our experiments will use these public, human-annotated datasets as the basis for training and validation. Touché23-ValueEval will be the main dataset.

3 Human Value Identification - Task and Basic Methods

3.1 Task Definition

We first introduce the formal definition of the human value identification task, generally following [Kiesel *et al.*, 2022; Kiesel *et al.*, 2023]. For text data T, the task of **human value identification** in this paper is to generate a value label $V_i(T) \in \{0, 1\}$ for every human value V_i in a value system $\mathbb{V} = \{V_1 : D_1, \ldots, V_n : D_n\}$, which can be viewed as a multi-label classification task. Each value item V_i has its corresponding definition D_i, a paragraph of natural language (see Figure 1) that specifies the meaning of value item. For example, the definition of value *Self-direction: thought* is: *It*

is good to have own ideas and interests. Contained values and associated arguments of this value: Be creative: arguments towards more creativity or imagination; Be curious ... (omitted). The labels are: 0 (the text data has no clear connection to the value item) and 1 (the text data resorts to this value item). Unlike other classification tasks in NLP such as binary sentiment analysis, human value identification focuses on more abstract and complex concepts, requiring deep understanding of both input text and value system definition.

3.2 Basic LM-based Methods

Naturally, considering that value identification can be viewed as a multi-label classification task, we can employ some straightforward NLP methods and models to address it, including direct fine-tuning and prompt-based methods for LLMs . We will first describe and analyze these simple approaches before introducing EAVIT.

Fine-tuning Local language models that can be fune-tuned without incurring significant costs can be directly trained to fit the task. For encoder models, we can use embedding vector (like [CLS] in BERT [Devlin *et al.*, 2019]) to directly generate results for value identification through a linear layer. For the emerging generative models like GPT-2 [Brown *et al.*, 2020] and Llama [Touvron *et al.*, 2023a], we can no longer extract a clear embedding vector representing the entire input sequence. Instead, we can apply prompt-based supervised fine-tuning [Brown *et al.*, 2020], which involves using prompts that guide the model towards generating outputs that contains the identified values.

Prompt-based Methods For black-box models like GPT-4 where fine-tuning is either unavailable or too expensive, we typically can only obtain results in natural language form by prompting the model with input prompt queries. Therefore, an intuitive idea is to first input the complete definition of the value system to LLM for learning, then prompt it to identify values from text data. The main challenge with this approach (we call it *single-step prompting*, which completes the task in 1 LLM API call) is that the performance of LLMs tends to deteriorate with the increase of context length when handling complex tasks defined by context. [Liu *et al.*, 2023] has confirmed, when the context length reaches 2k-4k, the GPT model's ability to understand and remember context information significantly worsens. We also find that in this approach the model tends to point out values that appears in the beginning and end of the definition context, learning to unsatisfactory performance. To reduce the context size, identifying each value component individually is also feasible. But this method increases the total token usage, and result bias becomes more serious (see Section 5).

4 Method

In this section we introduce EAVIT which utilizes both local tunable LM and online LLMs. Our method involves three stages: (1) Training value detector; (2) Generating candidate value set; (3) Final value identification using LLMs. Prompt templates can be found in Appendix.

4.1 Training Value Detector

Our first objective is to tune a local LM that has the basic value identification capabilities. To achieve this goal, we opt to fine-tune the open-source generative language model Llama2-13b-chat [Touvron *et al.*, 2023b] as value detector, to equip the base model with robust semantic understanding capabilities. With QLoRA [Dettmers *et al.*, 2023; Hu *et al.*, 2021], finetuning Llama2-13b-chat can be executed on 4 Nvidia RTX 4090 GPUs with 24GB VRAM.

Explanation-based Fine-tuning

The process of value identification can be viewed as identifying the semantic association between the text content and values based on their definition. Therefore, correct identification result must be *interpretable*. When training encoder-only models like BERT, we can only use numerical value labels as supervisory signals for fine-tuning. However, when using generative models like Llama, we can include natural language explanations of the value identification process during fine-tuning, which enables the model to gain a much deeper understanding of the task's implications, similar to chain-of-thought [Wei *et al.*, 2023] and [Ludan *et al.*, 2023].

We use GPT-4o-mini to add explanations to the value identification results in the training dataset. Specifically, for input data T and positively labeled V_i, we guide LLM to provide a brief explanation based on the definition of this value item and the input data. For example, for text data *we should ban human cloning as it will only cause issues when you have a bunch of the same humans running around all acting the same* and its corresponding value item *Security: societal*, the explanation could be *The text is related to societal security as it addresses the potential chaos and disorder that could arise from human cloning.* After obtaining explanations for the training data labels, we fine-tune the value detector using the following prompt templates below in Alpaca format [Taori *et al.*, 2023]. We aim to train the model on the semantic logic behind value identification by forcing it generate explanations for each of its results.

```
// Simplified prompt template
### Instruction:
...
For the following input context,
   identify relevant values from 20
   value items.
Recall the definition of these basic
   values, then select the values that
   are most prominently reflected or
   opposed in the context, and provide
   your explanation.
...

### Input:
[INPUT_TEXT]

### Response:
(1) [VALUE_1]. Explanation: [
   EXPLANATION_1];
...
```

Proved by experiments, this method significantly enhances the reliability and performance.

Data Generation

Existing value identification datasets (Webis-ArgValues-22, Touché23-ValueEval) are all manually collected and annotated. While manual annotation enhances data reliability, it also leads to a scarcity of data (Touché23-ValueEval only contains 5.4k training instances). More critically, these real-world datasets present significant label imbalance, as illustrated in Figure 2 below. To address the issue, we first follow the *self-instruction* [Wang *et al.*, 2022] method by employing in-context learning (ICL) to mimic and generate new data based on the existing annotated data with explanations using GPT-4o-miniGPT-4o-mini, to expand the dataset scale. In addition to data generation based on the single in-context learning (ICL) method, we also utilize *targeted data generation* to compensate for the distribution imbalances of different value labels in the training dataset. We find the least frequent values and select the corresponding labels that have occurred in the training dataset as target labels, then use them as ICL examples to guide LLM in generating corresponding data. After each round, we filter out the repeated (ROGUE-L similarity > 0.7) and obvious erroneous data. Additionally, we constantly replace the examples used in in-context learning, ensuring that the value annotations of all examples comprehensively cover the entire value system.

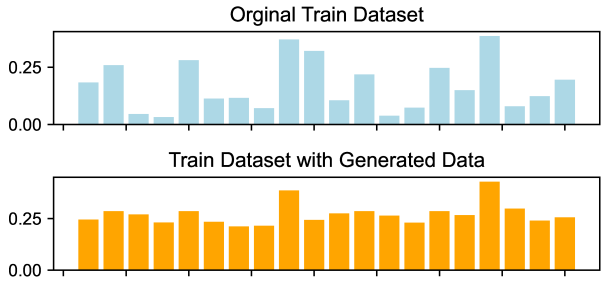

Figure 2: Value class distribution of original Touché23-ValueEval train dataset and ours with generated data.

For Touché23-ValueEval, we have generated approximately 8k ($1.5\times$ original size) instances of data, and significantly compensated the issue of uneven class distribution (see Figure 2). Consistent with existing research [Meng *et al.*, 2022; Huang *et al.*, 2022; Meng *et al.*, 2023], the generated data is of high quality and effectively enhances the model's generalization capabilities and ability to recognize value labels with lesser distribution.

Value Definition Reflection

In addition to previous methods with text-label data, we also want to teach the model the definition of the entire value system through *explicit reflection*. We achieve this goal by forcing the model to reflect on the definition of values during training.

4.2 Candidate Value Set Generation

Next, with a trained value detector that can produce preliminary value identification results, our goal is to obtain a set of candidate values through the preliminary results, which are values *possibly* related to the input text that need LLM for final determination. Our basic assumption is that the value detector will produce outputs that have some random deviations compared to the correct results. We first sample the output L times for each input data T randomly, and calculate the probability of each value being relevant to T:

$$\bar{V}_i(T) = |\{j : V_i^j(T) = 1, j = 1, \ldots, L\}|/T,$$

where $V_i^j(T)$ is the label of V_i at j-th output. Usually we set $L = 5$ to achieve the balance of reducing randomness and efficiency. Next, we set two thresholds $0 < p_{\text{low}} < p_{\text{high}} < 1$. For all value items V_i with $\bar{V}_i(T) > p_{\text{high}}$, we directly determine that $V_i(T)$ is related to T, i.e. $V_i(T) = 1$; while the value items V_j with positive probability between p_{low} and p_{high} constitute our candidate value set $S(T)$, i.e.

$$S(T) = \{V_j : p_{\text{low}} \leq \bar{V}_j(T) \leq p_{\text{high}}\}.$$

4.3 Final Value Identification via LLMs

Finally, we use the most acknowledged online LLM, the GPT series including GPT-4, GPT-4o, GPT-4o-mini, to test the values in the candidate set to obtain the final identification results. This is a simple, but most important step in EAVIT. Here, we only need to include the definitions of the values in the candidate set in the LLM prompt. Since the candidate set S is much smaller (3.3 for Touché23-ValueEval) than the entire value system $\mathbb{V}$ (20), the context length of our approach is much shorter than directly using the LLM (Table 1). More concentrated input prompt can make LLM's output more accurate and reduce the bias caused by forgetting and the order of context content; at the same time, the API cost is significantly lower.

5 Experiments

5.1 Value Identification on Public Datasets

Datasets and Methods

We conducted experiments on three public and manually-labelled datasets: ValueNet (Augmented) [Qiu *et al.*, 2022], Webis-ArgValues-22 [Kiesel *et al.*, 2022], and Touché23-ValueEval [Kiesel *et al.*, 2023]. Details can be found in Appendix. In experiments we focus on the Schwartz value systems, but our general method can also be applied to other value systems including [Ren *et al.*, 2024] by changing the definitions of the value system and training datasets. For all datasets, we report the accuracy and the **officially recommended** F1-score on the validation and test data. We conducted a comparative analysis of our proposed EAVIT approach against various fine-tuning and prompt-based methods. For encoder models BERT [Devlin *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019], we directly obtain the prediction results by passing [CLS] token embedding through a linear layer, and train on training dataset. We also reference the SemEval-2023 Task 4 competition best result [Schroter *et al.*,

Dataset	Webis-ArgValues-22				Touché23-ValueEval				
Dataset Split	Validation		Test		Validation		Test		
Metric	Acc	F1-score	Acc	F1-score	Acc	F1-score	Acc	F1-score	#Token
BERT (finetune)	0.78	0.32	0.79	0.33	0.81	0.40	0.79	0.41	-
RoBERTa (finetune)	0.79	0.31	0.78	0.35	0.81	0.39	0.80	0.42	-
ValueEval'23 Best [Schroter *et al.*, 2023]	-	-	-	-	-	-	-	0.56	-
GPT-2 (finetune+prompting)	0.75	0.34	0.76	0.33	0.72	0.30	0.73	0.34	-
Llama2-chat-13b (prompting)	0.70	0.29	0.72	0.30	0.78	0.31	0.76	0.27	-
Llama2-chat-13b (finetune+prompting)	0.86	0.42	0.84	0.44	0.82	0.41	0.82	0.45	-
GPT-4o-mini (*single-step* prompting)	0.83	0.50	0.84	0.50	0.82	0.54	0.86	0.53	2.4k
GPT-4o (*single-step* prompting)	0.84	0.51	0.86	0.54	0.85	0.55	0.86	0.52	
GPT-4o-mini (*5-steps* prompting)	0.83	0.49	0.82	0.50	0.84	0.49	0.85	0.52	2.8k
GPT-4o (*5-steps* prompting)	0.86	0.51	0.84	0.50	0.87	0.50	0.87	0.54	
GPT-4o-mini (*sequential* prompting - simple)	0.82	0.50	0.86	0.49	0.83	0.51	0.85	0.50	3.0k
GPT-4o (*sequential* prompting - simple)	0.82	0.52	0.84	0.51	0.87	0.55	0.89	0.54	
GPT-4o-mini (*sequential* prompting - CoT)	0.83	0.52	0.84	0.52	0.87	0.56	0.86	0.57	3.6k
GPT-4o (*sequential* prompting - CoT)	0.86	0.55	0.86	0.56	0.88	0.57	0.89	0.58	
Llama2-chat-13b (EAVIT)	0.88±0.05	0.52±0.03	0.89±0.07	0.53±0.02	0.88±0.04	0.55±0.03	0.89±0.07	0.57±0.01	-
EAVIT (Llama2-chat-13b + GPT-4o-mini)	0.93±0.02	0.63±0.04	**0.92**±0.03	**0.63**±0.02	**0.95**±0.01	0.65±0.02	**0.94**±0.02	0.66±0.03	**0.45k**
EAVIT (Llama2-chat-13b + GPT-4o)	**0.94**±0.03	**0.65**±0.02	**0.92**±0.02	**0.66**±0.01	**0.95**±0.02	**0.66**±0.01	**0.94**±0.04	**0.69**±0.02	

Table 1: Results on Webis-ArgValues-22 and Touché23-ValueEval (level-2 label) dataset.

Dataset Split	Validation	Test	
Metric	Accuracy		#LLM Token / Sample
BERT (Finetune)	0.60	0.61	-
RoBERTa (Finetune)	0.57	0.55	-
GPT-4o-mini (single-step prompting)	0.70	0.72	1.5k
GPT-4o-mini (sequential prompting)	0.75	0.78	1.9k
EAVIT (Llama2-chat-13b + GPT-4o-mini)	**0.80**	**0.78**	**0.5k**

Table 2: Results on ValueNet (augmented) dataset.

2023], which utilized multiple ensemble RoBERTa models and larger pretrain datasets. For GPT-2 [Brown *et al.*, 2020] and Llama2-13b-chat [Touvron *et al.*, 2023b], we employ a simple prompt to directly output the value identification results, and report the direct results and fine-tuned results. As we have discussed in Section 3.2, for online LLMs GPT-4o-mini and GPT-4(o) using OpenAI API [OpenAI, 2023], we adopt *single-step prompting*, *5-steps prompting* and *sequential prompting* with multiple variants. *Single-step prompting* let the model to determine the complete value identification results in a single API call for each data, *sequential prompting* identifies each value individually resulting in multiple LLM API calls (20 for ValueEval'23) for each data, while *5-steps prompting* is a balance of those two methods with 5 API calls per data point. We apply direct prompting and chain-of-thought prompting ([Wei *et al.*, 2023], CoT) for sequential prompting baselines. For EAVIT, we set $p_{\text{low}} = 0.2, p_{\text{high}} = 0.8$ and report the results of the value detector the entire method. Detailed configurations and prompts can be found in the appendix. We report the average and std of 3 random individual runs.

Results

Table 1,2 and Figure 3 show the experiment results.

Performance and efficiency of EAVIT. We can observe that EAVIT significantly improves accuracy while only using up to 1/7 LLM tokens compared to directly using LLMs. Comparing with directly fine-tuning, EAVIT's performance on values that appear less frequently on training data is noticeably better, demonstrating the effectiveness of data generation and explain-based fine-tuning. Furthermore, the performance of the EAVIT method surpasses that of standalone local LMs or online LLM methods, indicating that their combination can effectively complement the shortcomings - the local LM's relatively weak textual understanding power and the substantial impact of context on performance and efficiency of online LLMs - to achieve a more optimal balance between performance and efficiency.

Prompting Methods. The context length in single-step prompting is typically longer, which generally results in lower accuracy than models that have been fine-tuned for aligned models. However, when using sequential prompting, even though the query context is more precise, the overall accuracy does not make much difference. According to our observations, this is probably because LLMs have a tendency to provide affirmative responses ($> 60\%$ in our experiments) due to alignment [Perez *et al.*, 2022]. By using Chain-of-Thought, this problem is partly alleviated (reduced to 40%) and the performance is significantly improved.

GPT-4o-mini v.s. GPT-4o. In most experiments, the performances of GPT-4o and its mini version are similar. Con-

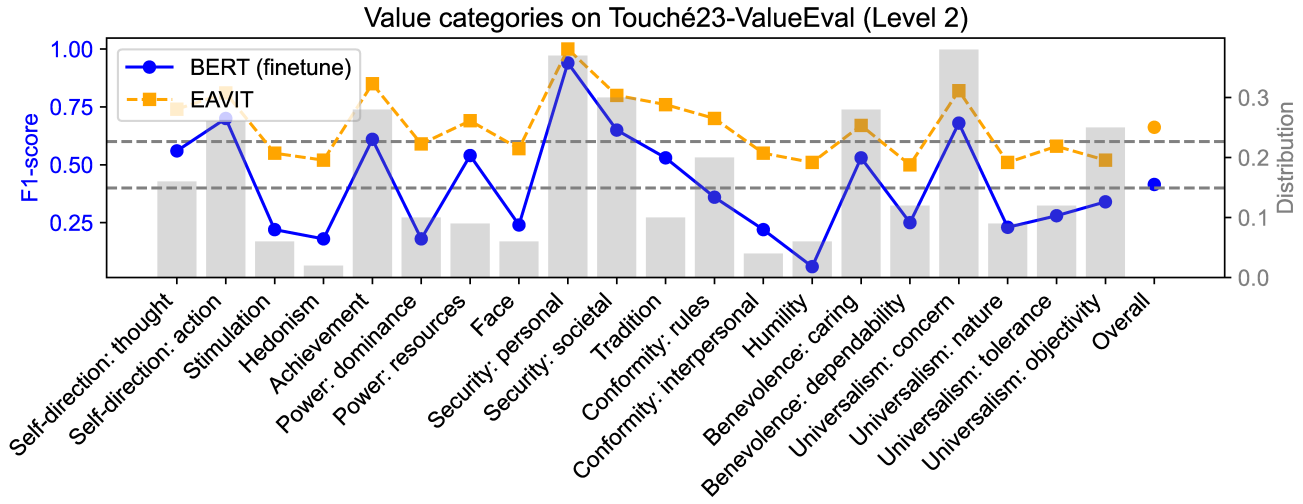

Figure 3: Plot of **F1-scores** and **class distribution** on the Touché23-ValueEval test set over the labels by level. The grey bars show the label distribution.

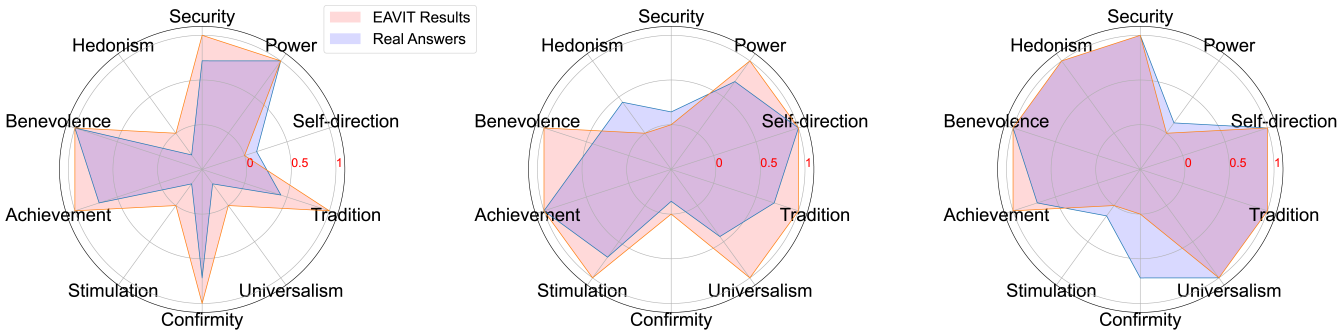

Figure 4: Sample Visualization of Value Identification of Virtual Individuals

sidering the API cost, we suggest that GPT-4o-mini is sufficient in most practical human value identification tasks.

5.2 Case Study: Value Identification of Virtual Individuals

Next, we conduct a hypothetical experiment to simulate the process of human value identification of individuals. The objective is to compare the uniformity and differences between the individual values measured by traditional *psychological questionnaires* and our *text-based* value identification method. We use the public real-human questionnaire results from the World Values Survey (WVS) wave 6 [Inglehart *et al.*, 2000; Inglehart *et al.*, 2014], which is an extensive project that has been conducting value tests on populations worldwide for decades. We selected the responses of 20 real individuals to 10 questions about the Schwartz value system. These questions (v70-v79) sequentially correspond to 10 Schwartz values, and can be considered as a standard value questionnaire in psychology. For each individual, we input the content of these questions and their responses to these questions into GPT-4, and guided GPT-4 to mimic an *virtual individual possessing these values* through prompts [Aher *et al.*, 2023]. Subsequently, we selected 20 social topics and guided the simulated individuals to express and explain their views on these social topics, forming 20 text data instances for each individual with format similar to Touché23-ValueEval. Finally, we process these text data using EAVIT and other methods trained on Touché23-ValueEval to identify values behind them, aggregate the results of each virtual individual's text data, and compare them with psychological questionnaire answers. The experiment details are provided in the appendix.

Method	Mean Accuracy
RoBERTa (Finetune)	0.62
GPT-4o (single-step prompting)	0.71
GPT-4o (sequential prompting)	0.65
EAVIT	**0.78**

Table 3: Results of virtual individual value identification. EAVIT uses Llama2-chat-13b + GPT-4o-mini here.

Results. Our results and visualizations are presented in Table 3 and Figure 4. Our findings indicate that by conducting value identification on the viewpoints data generated by virtual individuals, we can effectively infer their value presets, with a high level of consistency with psychological questionnaires based on values. Similarly, it can be observed from Table 3 that the performance of EAVIT on this task outperforms both simple LMs and naive usage of LLMs. This experiment uncovers an intriguing potential: we can use large models to conduct value identification on passively collected text data generated by individuals, providing a measure of an individual's values without resorting to the active collection methods in psychology like questionnaires [Schwartz *et al.*, 2001]. This approach, based on LLMs and passively collected data, has the advantages of low data collection cost, high credibility, and strong non-falsifiability. For example, a selfish person is unlikely to answer "no" to the questionnaire item "Do you care about others?", but its behavior could likely be reflected in social network traces.

5.3 Training Value Detector: Ablation Study

We now investigate the impact of fine-tuning strategies and data generation on the performance of the value detector model on Touché23-ValueEval dataset. In order to directly investigate the impact of different fine-tuning methods and data generation techniques, we train four different versions of the Llama2-13b-chat value detector models: *+org_dataset*: model fine-tuned on the original dataset using direct prompts; *+explain-based FT*: model with explanation-based fine-tuning on the original dataset. *+icl_data*: the previous model with an additional 4k training data entries generated using a simple ICL; *+target_value_data*: the previous model with an additional 4k training data entries specifically generated for less frequent target values. We then plot the performance of these models in Figure 5. The results demonstrate that the explain-based finetune method and data generation effectively enhances the model's performance.

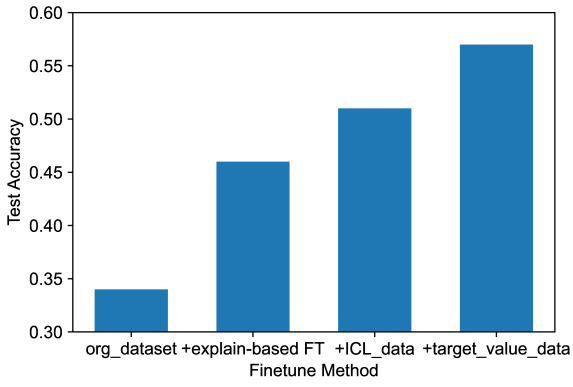

Figure 5: Results of different models with different finetune strategies.

5.4 Output Consistency

To ensure the feasibility of our method in practical applications such as large-scale data analysis, it is crucial to achieve high output consistency. To this end, we randomly extracted 200 data points from the Touché23-ValueEval test dataset. For each data point, we randomly sampled the results 10 times at different output stages (*1 detector output*: single output from the value detector; *5 detector output*: average of 5 outputs from the value detector, *candidate set*: see Section 4.2, and *final result*). We computed the average variance of 10 samples, with each output viewed as a 20-dimension vector. The results are presented in Figure6.

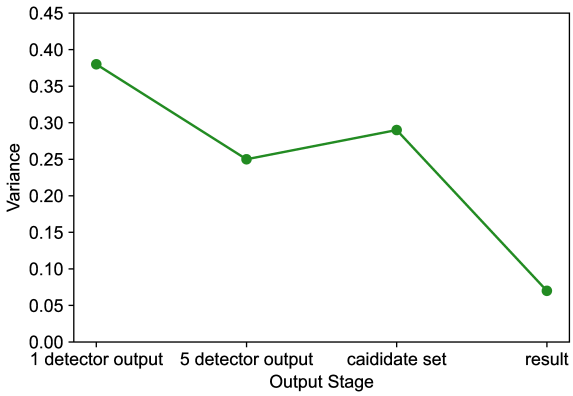

Figure 6: Output variance at different output stages.

We observe that the single output from the value detector exhibited high randomness, but sampling effectively reduces it. Despite the increased randomness in the candidate set (which indicates that different outputs from the value detector may favor different random values), the final decision made by the LLMs effectively eliminated irrelevant random errors, achieving stable, high-quality results. This suggests that our method could be applied to scenarios that require high output stability.

6 Conclusion

This paper investigates the potential of LLMs for identifying human values from text data. Despite challenges, LLMs have demonstrated superior capabilities compared to previous methods. Our work of efficient and accurate value identification can have potential usage for value alignment in LLMs, social network analysis, and psychological studies.

Ethical Statement

Our experiments only utilize existing public models and datasets (may include human-generated or human-labeled data) available for research, currently there are no ethical issues. However, our method involves inferring and measuring human values, and potential ethical risks must be carefully considered in possible practical applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62276006).

References

[Aher *et al.*, 2023] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[Alshomary *et al.*, 2022] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. The moral debater: A study on the computational generation of morally framed arguments. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Atkinson and Bench-Capon, 2021] Katie Atkinson and Trevor Bench-Capon. Value-based argumentation. *Journal of Applied Logics*, 8(6):1543–1588, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019.

[Fischer *et al.*, 2023] Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*, 2023.

[Forbes *et al.*, 2020] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*, 2020.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Huang *et al.*, 2022] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[Inglehart *et al.*, 2000] Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*, 2000.

[Inglehart *et al.*, 2014] Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. World values survey: Round six - country-pooled datafile version: https://www.worldvaluessurvey.org/wvsdocumentationwv6.jsp., 2014.

[Kiesel *et al.*, 2022] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. Identifying the Human Values behind Arguments. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics, May 2022.

[Kiesel *et al.*, 2023] Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, 2023.

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

[Liu *et al.*, 2023] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[Ludan *et al.*, 2023] Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Explanation-based finetuning makes models more robust to spurious cues. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4420–4441, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Meng *et al.*, 2022] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding, 2022.

[Meng *et al.*, 2023] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023.

[Miotto *et al.*, 2022] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*, 2022.

[Ng *et al.*, 2005] Thomas WH Ng, Lillian T Eby, Kelly L Sorensen, and Daniel C Feldman. Predictors of objective

and subjective career success: A meta-analysis. *Personnel psychology*, 58(2):367–408, 2005.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[Qiu *et al.*, 2022] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191, 2022.

[Ren *et al.*, 2024] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. *arXiv preprint arXiv:2406.04214*, 2024.

[Schroter *et al.*, 2023] Daniel Schroter, Daryna Dementieva, and Georg Groh. Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. *arXiv preprint arXiv:2305.08625*, 2023.

[Schwartz *et al.*, 2001] Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542, 2001.

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

[Taori *et al.*, 2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[Touvron *et al.*, 2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[Touvron *et al.*, 2023b] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2022] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[Wei *et al.*, 2023] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[Yao *et al.*, 2023] Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values. *arXiv preprint arXiv:2311.10766*, 2023.