Taking a SEAT: Predicting Value Interpretations from Sentiment, Emotion, Argument, and Topic Annotations

Adina Nicola Dobrinoiu*, Ana Cristiana Marcu*, Amir Homayounirad, Luciano Cavalcante Siebert, and Enrico Liscio

Delft University of Technology, the Netherlands
{A.N.Dobrinoiu, A.C.Marcu-1}@student.tudelft.nl
{A.Homayounirad, L.CavalcanteSiebert, E.Liscio}@tudelft.nl

Abstract. Our interpretation of value concepts is shaped by our sociocultural background and lived experiences, and is thus subjective. Recognizing individual value interpretations is important for developing AI systems that can align with diverse human perspectives and avoid bias toward majority viewpoints. To this end, we investigate whether a language model can predict individual value interpretations by leveraging multi-dimensional subjective annotations as a proxy for their interpretive lens. That is, we evaluate whether providing examples of how an individual annotates Sentiment, Emotion, Argument, and Topics (SEAT dimensions) helps a language model in predicting their value interpretations. Our experiment across different zero- and few-shot settings demonstrates that providing all SEAT dimensions simultaneously yields superior performance compared to individual dimensions and a baseline where no information about the individual is provided. Furthermore, individual variations across annotators highlight the importance of accounting for the incorporation of individual subjective annotators. To the best of our knowledge, this controlled setting, although small in size, is the first attempt to go beyond demographics and investigate the impact of annotation behavior on value prediction, providing a solid foundation for future large-scale validation.

Keywords: Values · Value Prediction · Value Classification · LLMs · Subjectivity.

1 Introduction

Human values are deeply intertwined with argumentation, influencing how people form opinions and justify their stances [6, 27]. Much of real-world argumentation implicitly appeals to values as commonly accepted reasons why a position is desirable or "right" [24, 46]. Being able to recognize these underlying values is crucial for NLP applications such as argument mining, persuasive dialogue systems, and aligning AI responses with human morality [30].

*Equal contribution.

Predicting which values a given text or argument expresses is an inherently subjective endeavor, due to the subjective nature of valuing [34]. Different individuals can interpret the same text differently, possibly reflecting their own values and perspectives. This phenomenon is a specific instance of perspectivism, a broader challenge in the AI community: many language understanding tasks lack a single "ground truth" label because meaning is partly in the eye of the beholder. Recent perspectivist approaches argue that we should embrace, rather than erase, such differences [9]. Instead of leaning into majority aggregation or discarding disagreements (which obscures the rich diversity of valid interpretations), perspectivism advocates that AI models should be capable of recognizing and modeling each individual's viewpoint [8].

In this work, we investigate whether language models can predict and represent the diversity of interpretation of the values behind a piece of text. In other words, can a language model detect that two individuals might read the same text differently? Addressing this question is important for developing AI that is sensitive to whom it is interacting with, with implications for personalized AI assistants and fairness (ensuring minority viewpoints are not ignored by models).

A common approach in prior research has been to use demographic attributes as proxies for human differences. For instance, gender, age, or culture are often used to tailor models or prompts, on the assumption that these attributes correlate with the person's judgments [12, 14, 50]. However, recent findings show that even explicitly adding annotators' demographic information into a model did not significantly improve its ability to predict those annotators' labels [40]. Similarly, attempts to prompt language models with demographic personas (e.g., asking the model to respond "as a 30-year-old woman") have yielded limited accuracy in simulating that group's perspective [39]. This suggests that individual annotation behavior depends on more than broad sociodemographic categories.

We explore a new approach to approximate an individual's value interpretation: leveraging the person's prior textual annotations across related dimensions. Several textual dimensions are frequently studied in NLP with values [33, 38]. In particular, we use a person's annotations for Sentiment, Emotion, Argument, and Topic, which we term the *SEAT dimensions*, as a proxy for their interpretive lens. Intuitively, how someone recognizes sentiment and emotion in text, what arguments they find expressed, and which topics they focus on can reveal their underlying value interpretations. This approach moves beyond static demographics to behavioral cues embedded in language. Recent work lends support to this idea: [22] found that an annotator's task-specific attitudes and preferences were more predictive of their labeling behavior than their demographic traits. In other words, knowing how a person tends to label content is a better indicator of their perspective than knowing their age or gender.

Specifically, we investigate whether providing past SEAT annotations can guide a language model in predicting a person's value interpretations. We frame the task as follows: given a language model and a particular individual, we provide the model with a text sample and ask it to predict how the individual would interpret the human values expressed in the text sample. We then evaluate how

the prediction changes when providing different amounts of that individual's past annotations on the SEAT dimensions for the sample text and/or other similar texts. To test this approach, we ask five individuals to annotate a survey dataset [20] with the values and the SEAT dimensions expressed in every data point. We then employ a language model to predict the values behind the texts that compose the dataset, by providing different levels of information on each annotator's SEAT annotations. Our results show that providing a few examples of the full set of SEAT annotations improves the prediction of individual value interpretations. Instead, providing examples with the individual SEAT dimensions does not improve over the off-the-shelf baseline.

This work provides early insights into language models' sensitivity to individual differences in value interpretation based on non-demographic textual cues. To our knowledge, this is the first attempt to personalize value prediction by conditioning a language model on a user's multi-dimensional annotation history, paving the way for more refined behavioral cues that can be used in support of predicting individual value interpretations, such as multi-modal and non-textual cues. Such a nuanced approach could contribute to improved deliberation support by providing participants with counterfactual scenarios of value interpretation based on the SEAT dimensions.

2 Related Works and Background

We start by reviewing works on the theoretical background behind human values and their connection to the SEAT dimensions. Next, we review literature on the prediction of values in text and LLM-based approaches to it.

2.1 Human Values and SEAT Dimensions

Human values are guiding principles individuals use to shape their choices and actions [47]. The Schwartz theory of fundamental human values categorizes values into universal dimensions such as benevolence, achievement, tradition, and autonomy, which consistently emerge across cultures as central determinants of human judgment and behavior [47]. These values influence how people form opinions, argue, and deliberate, shaping individual and collective decision-making processes [24]. [46] further clarifies values as beliefs relating to desirable end-states or modes of conduct and describes value systems as hierarchies of values formed by cultural, social, and personal factors, emphasizing that values belong inherently to personas rather than objects, enabling systematic analysis of individual, essential for understanding human disagreement and decision making.

Several textual dimensions frequently studied in NLP have been associated with values, particularly within deliberative contexts [33, 38]. **S**entiment analysis, focusing on identifying positive or negative attitudes, can serve as an indicator of implicit value judgments underlying these attitudes [53]. **E**motion recognition categorizes explicit affective states such as joy, anger, or sadness, and it

has been employed to improve the identification of the underlying values expressed in discourse [21]. **A**rgument mining captures structural elements that describe why particular positions are considered desirable or persuasive, and because of their motivational nature, these elements have often been associated with values [24, 25]. **T**opic modeling identifies the thematic domains emphasized in discourse, with the identification of topics shown to be influenced by speakers' cultural backgrounds and personal value systems [45]. We collectively refer to Sentiment, Emotion, Argument, and Topic as the four *SEAT dimensions*. These SEAT dimensions provide comprehensive textual proxies for exploring how values are expressed in text, which might make them suitable as auxiliary inputs to a language model when predicting individual interpretations of values.

2.2 Value Prediction through NLP Approaches

A traditional method for classifying values in text involves using value dictionaries—lists of words associated with specific values—by analyzing the relative frequency of these words [42], such as in the Moral Foundation Dictionary [15]. These dictionaries have been extended using semi-automated approaches [3, 18, 52] and NLP methods [4, 43], while the limitations of word count methods have been addressed with word embedding models [5, 13, 41].

Recent approaches leverage supervised machine learning [2, 24, 31, 28, 36], where language models are trained on datasets with value annotations, such as ValueNet [44] or Value Kaleidoscope [49]. The introduction of the ValueEval '23 [25] and '24 [26] shared task has cemented value classification into a recognized challenge, with most participants employing the supervised learning paradigm.

However, supervised learning requires the availability of human annotations, which may be unfeasible in domain-specific applications such as debating energy transition policies in a Dutch municipality (i.e., the dataset we use in our experiments, as detailed in Section 3.1). The transfer learning paradigm has been evaluated but proven to be of limited efficacy for the value prediction task [19, 29]. To this end, recent approaches have shown promising value prediction results with the zero-shot prompting paradigm with LLMs [37, 48], which we employ in this work to predict values in text and extend by employing the SEAT dimensions as auxiliary information to recognize individuals' value interpretations.

2.3 Prompting Large Language Models

State-of-the-art language models have demonstrated generalization capability through prompting techniques such as zero-shot and few-shot learning [51]. In a zero-shot setting, models receive only instructions and generalize to new tasks without task-specific fine-tuning, whereas few-shot prompts guide models by providing a handful of illustrative examples in-context [7]. Building on this, recent methods have attempted to steer model outputs toward individual styles or viewpoints by prompting the model to adopt a specific persona—for example, instructing it to respond as "a friendly and outgoing person" or using descriptions of a user's characteristics [54]. However, simplistic persona-based prompts,

often reliant on demographic attributes, have shown limited success in capturing individuals' nuanced and diverse preferences [11, 40]. Personalization based on behavioral rather than demographic signals may align better with actual user preferences and lead to improved alignment between model predictions and individual annotations [22]. Our study builds on this hypothesis by using SEAT annotations as auxiliary input when prompting a language model to predict the values expressed in a piece of text, testing the model's capacity to reflect individual users' value judgments based on their annotation behaviors.

3 Methodology and Experiments

We frame value prediction as a multi-label text classification task: given an input text, the goal is to identify which human values (from a predefined set) are expressed or implied (based on an annotator's interpretation of the text). We hypothesize that incorporating context from related subjective NLP tasks can improve value identification. In particular, we leverage annotations from **S**entiment analysis, **E**motion recognition, **A**rgument detection, and **T**opic classification as additional inputs. By providing these auxiliary signals about the text's sentiment, emotional tone, argumentative content, or topic, we expect the language model to better infer the expressed values. The key question is whether an LLM can use annotations from these related tasks to more accurately predict human values (e.g., does knowing the text's emotional tone or topic help in determining if it reflects values like honesty or success?). We evaluate this hypothesis by comparing the model's value predictions with and without such auxiliary annotations. Data and code are publicly available[1].

3.1 Data and Annotations

We experiment with the Energy Transition PVE dataset [20], a survey administered to citizens of the Súdwest-Fryslân municipality in the Netherlands to support in co-creating an energy transition policy. Within the survey, citizens could distribute their votes among several options and write textual *justifications* for their choices. These justifications provide insight into the values that motivate them, which can be identified through NLP methods as shown by [32].

We select a subset of 50 justifications and ask five annotators to annotate each of them with the following five dimensions: (1) *Sentiment*—the overall sentiment polarity of the justification on a five-point scale from Very Negative to Very Positive, (2) *Emotion*—the prominent emotions expressed (one or multiple from a popular emotion taxonomy [10], e.g. curiosity, confusion), (3) *Argument*—a summary or key snippet of the argumentative content in the justification (if any), (4) *Topic*—a brief descriptor of the topic or domain of the justification (e.g. energy independence, social engagement), and, finally, (5) *Values*—the human values described in the Schwartz value theory, adapted by [24]. The value

[1] https://github.com/adina-dobrinoiu/SEAT

annotation was performed with the original 54 labels, which we grouped into the 20 parent labels to make the experiments and analysis manageable, in line with previous work [25, 26]. Appendix A reports the demographics of the annotators, the exact instructions, and the task labels.

Table 1 shows the inter-annotator agreement for the five annotation dimensions. Fleiss' kappa is used for sentiment, emotion, topic, and values because they involve selecting predefined categorical labels. Argument annotation requires identifying text spans that may vary in length and position, which makes categorical agreement measures like Fleiss' kappa unsuitable. Instead, agreement is measured using the pairwise F_1-score, which treats one annotator's spans as ground truth and another's as predictions, averaged over all annotator pairs. We observe that agreement is high for topic but relatively low for the other dimensions, underscoring the need for a personalized approach.

Table 1. Inter-annotator agreement for the annotation dimensions, measured through Fleiss' kappa except for argument, which is measured through pairwise F_1-score.

Sentiment	Emotion	Argument	Topic	Values
0.17	0.00365	0.2447	0.514	0.0144

3.2 Model and Prompting Strategy

We use a pretrained 8-billion-parameter language model to perform value prediction via prompted generation (meta-llama/Meta-Llama-3.1-8B-Instruct). We use an in-context learning paradigm—the model is provided with a prompt that includes task instructions, optional examples, and one textual justification (which we refer to as the *reference justification*), and asked to identify the value label(s) expressed in this reference justification. Appendix B describes the exact prompts. Precisely, we start by testing the following baseline:

Zero-shot baseline (ZS) The language model is provided with the reference justification and a list of values to choose from, and asked to identify the values expressed in the justification. We employ this as our baseline, where the model is provided no information about the individual.

Next, we evaluate whether providing additional information on an individual's annotation behavior leads to a more accurate prediction of the value that the individual considers to be expressed in the reference justification. To test this, for each annotator and each reference justification, we provide the model with information on how the annotator annotated the reference justification and other justifications with the SEAT dimensions.

One-shot prompt (OS) We extend the ZS baseline by providing the language model with the information on how the individual annotated the reference justification with one (or more) SEAT dimension.

Few-shot prompt (FS) We extend the OS prompt by providing additional examples of how the annotator annotated other justifications with one (or more) SEAT dimensions. To ensure that the few-shot examples in the prompt are relevant to the reference justification's content, we compute sentence embeddings using a pretrained SentenceTransformer model (all-MiniLM-L6-v2) and perform a K-nearest neighbors search in the embedding space to find the K most semantically similar justifications to the reference justifications in the dataset. We experiment with different values of K, as detailed in the next subsection.

3.3 Experimental Setup

We evaluate the approaches described in Section 3.2 in the following settings:

Dimension-specific The model is given only one SEAT dimension annotation at a time in the prompt. We run separate experiments for each SEAT dimension to analyze which auxiliary signal is most informative. For example, in the *emotion* setting for the OS prompt, for each annotator and each justification, the prompt provides as auxiliary information the emotion that the annotator identified in the reference justification (and likewise, the few-shot prompts include the emotion annotations for each additionally provided example).

All dimensions The model is given all four SEAT dimensions simultaneously to evaluate whether a combined set of contextual information is more effective than the single ones. For example, in the *all* setting for the OS prompt, for each annotator and each justification, the prompt provides as auxiliary information all the SEAT dimensions that the annotator identified in the reference justification (and likewise, the few-shot prompts include all these annotations for each additionally provided example).

We test the OS and FS prompts with each of the SEAT dimensions and with a combination of the dimensions. In the FS prompts, we experiment with $K = 4$, $K = 9$, and $K = 14$, thus resulting in a total of 5, 10, and 15 provided examples (including the reference justification under evaluation). Table 2 provides an overview of the resulting 20 settings we test in addition to the ZS baseline.

Table 2. Overview of the settings tested for each annotator.

Method	Sentiment	Emotion	Argument	Topic	All
One-shot	OS-S	OS-E	OS-A	OS-T	OS-all
Few-shot (5)	FS-5-S	FS-5-E	FS-5-A	FS-5-T	FS-5-all
Few-shot (10)	FS-10-S	FS-10-E	FS-10-A	FS-10-T	FS-10-all
Few-shot (15)	FS-15-S	FS-15-E	FS-15-A	FS-15-T	FS-15-all

For each of the 21 settings and each annotator, we perform value prediction for all 50 justifications in the dataset. We repeat this for the five annotators, resulting in 21x5=105 experiments (with each experiment consisting of performing

value prediction on the full dataset). We repeat each experiment five times with five different seeds to account for variability, and consider a value as predicted when it appears in at least three of the five repetitions. We report the average micro F_1-score as the metric for model performance.

4 Results

Table 3 shows the experimental results across the different methods, settings, and annotators.

Table 3. Complete overview of the results (F_1-score) for the five annotators (represented by their IDs). In bold, the best results per annotator; underlined, the results that are not significantly different from the best result.

ID	Method	Sentiment	Emotion	Argument	Topic	All
1	One-shot	0.180	0.155	0.137	<u>0.229</u>	<u>0.219</u>
	Few-shot (5)	0.160	0.189	0.174	0.138	**0.302**
	Few-shot (10)	0.118	0.173	0.169	0.153	<u>0.291</u>
	Few-shot (15)	0.166	0.160	0.145	0.146	<u>0.275</u>
	Baseline			0.164		
2	One-shot	0.210	0.282	0.185	0.284	0.246
	Few-shot (5)	0.263	0.244	0.245	0.217	0.314
	Few-shot (10)	0.249	0.234	0.264	0.207	**0.382**
	Few-shot (15)	0.254	0.218	0.252	0.192	<u>0.380</u>
	Baseline			0.244		
3	One-shot	0.182	0.184	0.181	0.207	0.190
	Few-shot (5)	0.200	<u>0.229</u>	<u>0.222</u>	0.199	<u>0.244</u>
	Few-shot (10)	0.204	<u>0.234</u>	0.207	0.200	**0.298**
	Few-shot (15)	<u>0.210</u>	<u>0.223</u>	0.194	0.184	<u>0.241</u>
	Baseline			0.230		
4	One-shot	0.261	0.246	0.164	<u>0.301</u>	0.252
	Few-shot (5)	0.226	<u>0.266</u>	0.240	0.199	<u>0.335</u>
	Few-shot (10)	0.234	0.263	0.258	0.221	**0.339**
	Few-shot (15)	0.234	0.260	0.229	0.214	<u>0.309</u>
	Baseline			0.221		
5	One-shot	0.257	0.228	0.215	0.287	0.271
	Few-shot (5)	0.306	0.261	0.189	0.293	<u>0.406</u>
	Few-shot (10)	0.343	0.336	0.301	0.312	**0.444**
	Few-shot (15)	0.312	0.329	0.343	0.300	<u>0.428</u>
	Baseline			0.228		

The best results are consistently achieved by providing all SEAT dimensions through the few-shot paradigm. To facilitate further interpretation of the results, in the next subsections, we break down specific aspects of these results to highlight the main findings.

4.1 Providing All SEAT Annotations Helps

We begin by investigating how providing different types of auxiliary information impacts the prediction of individual value interpretations. Fig. 1 shows the results grouped by the auxiliary information that is provided in the prompt (averaged over the five annotators).

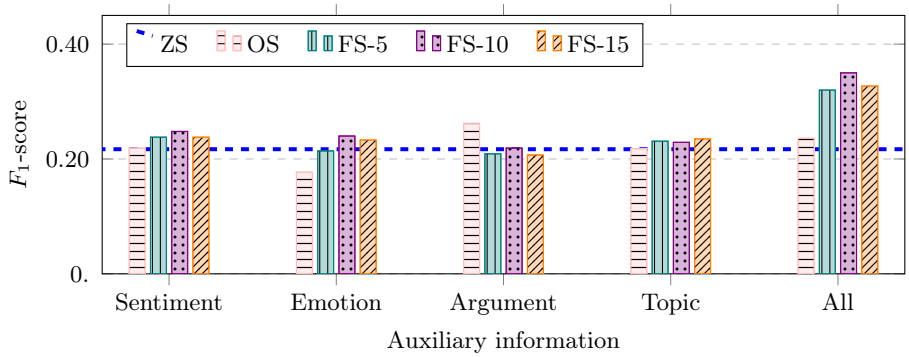

Fig. 1. F_1-score resulting from including different auxiliary information in the prompt, averaged over the five annotators. The horizontal dashed line displays the ZS baseline ($F_1 = 0.217$) averaged over the five annotators.

We observe that, in line with Table 3, providing a few examples with all SEAT annotations improves the prediction of individual value interpretations. Instead, providing examples of the individual SEAT annotations does not improve results over the ZS baseline. This suggests that value interpretation is a complex cognitive process that benefits from multiple contextual cues. Rather than relying on a single dimension, the model appears to integrate information across the four dimensions when inferring underlying values. This finding aligns with psychological theories of value-based decision-making, which emphasize the multi-faceted nature of moral and value judgment [1, 16, 46].

4.2 Similarities and Differences across Individuals

Next, we examine how the observed results differ across individuals. Fig. 2 and 3 show the results for each annotator, averaged over the different methods described in Section 3.2 (i.e., the rows of Table 2) and over the different settings described in Section 3.3 (i.e., the columns of Table 2), respectively.

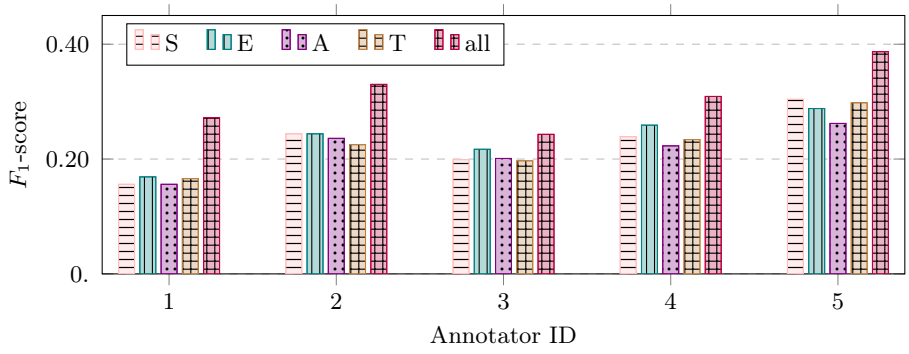

Fig. 2. F_1-score by annotator, averaged over the different methods used to provide auxiliary information (i.e., the four rows in Table 2).

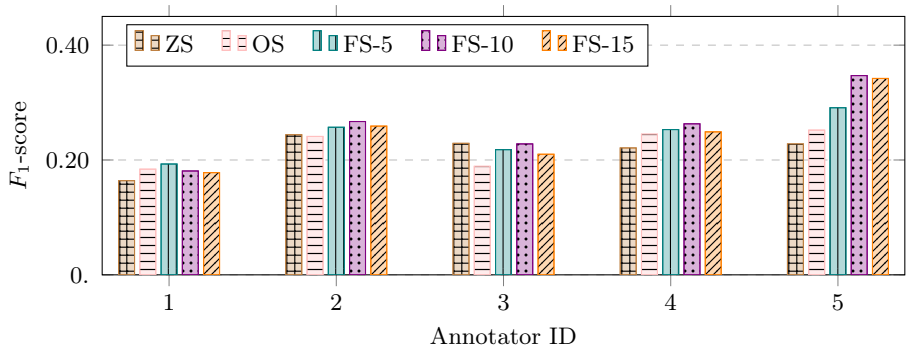

Fig. 3. F_1-score by annotator, averaged over the different provided auxiliary information (i.e., the five columns in Table 2).

Fig. 2 shows that providing all SEAT dimensions consistently improves results across all annotators. These findings are in line with the results discussed in Section 4.1—providing the individual SEAT annotation leads to consistently lower results, without noticeable difference among the SEAT dimensions.

Fig. 3 shows that providing more examples generally does not improve the prediction of individual value interpretations, confirming that the type of auxiliary information has more impact than the number of examples. In fact, for all annotators, the best results are achieved with FS-5 or FS-10, rather than FS-15. We conjecture that this is due to the strategy used to select examples in the FS method—that is, the K semantically most similar examples to the reference justification. A larger K (e.g., in FS-15) introduces less similar examples and thus more noise in the prompt, whereas smaller values of K appear to provide sufficient targeted information. This has significant practical implications—this moderate data requirement makes the approach feasible for real-world deployment without requiring extensive historical annotation data from each annotator.

This is particularly important for applications where user engagement may be limited or where privacy concerns restrict data collection.

Finally, both figures show differences across annotators. Annotator 5 consistently achieves the highest performance across methods and settings, while Annotator 1 shows the most modest improvement. These individual differences likely reflect varying levels of annotation consistency, familiarity with value concepts, and the degree to which personal value systems are reflected in SEAT annotations. Annotator 5's better performance may be attributed to their greater familiarity with value theory, as noted in Appendix A, leading to more systematic and predictable annotation patterns.

4.3 Changes in Prediction

The F_1-scores reported so far only reveal cases in which the model prediction, due to the provided auxiliary information, changes from wrong to correct (or vice versa). However, cases in which the model prediction changes without changing its correctness are also insightful, as they reveal that providing auxiliary information still prompts the model to change its prediction. Fig. 4 shows the changes in prediction, measured as the relative symmetric difference between baseline and alternative approach: $|A \triangle B|/|A|$, where A is the set of labels predicted with the ZS baseline setting, and B is the set predicted using auxiliary information. The numerator, $|A \triangle B|$, represents the number of elements that are in either A or B, but not in both, indicating the total number of changed labels.

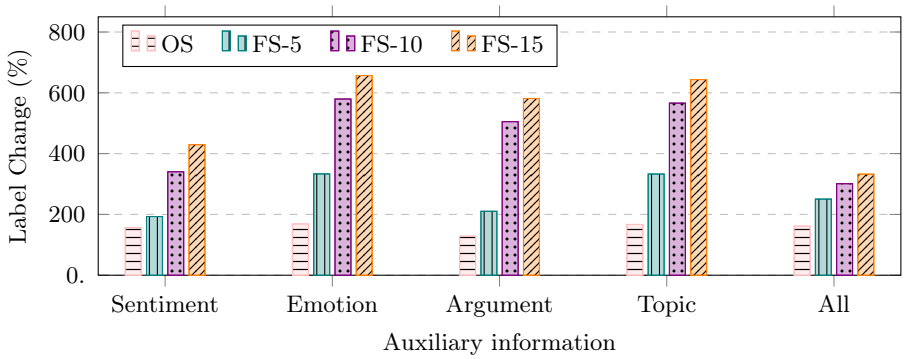

Fig. 4. Predicted label change compared to the zero-shot ZS baseline, measured as the symmetric difference ($|A \Delta B|/|A|$) percentage, averaged over the five annotators.

Surprisingly, the *all* setting produces the smallest percentage of prediction change while achieving the highest F_1-scores. This result appears to suggest that, when the model has access to comprehensive contextual information, it makes fewer but more accurate changes to its predictions. This may be because individual dimensions may introduce more random variation in prediction, whereas

combined dimensions provide clear guidance on when changes are warranted. In contrast, individual dimensions show higher prediction change rates but no performance improvement, suggesting that incomplete contextual information may lead to less discriminating model behavior. Nevertheless, further investigation is required to confirm these conjectures, as we elaborate next.

5 Discussion and Limitations

Overall, the main observed trend is that providing a more varied set of information improves individual value interpretation prediction across annotators, and is more impactful than providing just a single source of information, even if of increased quantity. This suggests that predicting individual value interpretations requires capturing multiple aspects of an individual's interpretative style, rather than relying on a single demographic or behavioral indicator, challenging the prevailing paradigm of demographic-based personalization in NLP [12, 14, 50].

Individual variations The differences between annotators highlight the importance of individual-level modeling rather than assuming uniform population responses. The better results with annotator 5, who had greater familiarity with value theory, suggests that annotation expertise and conceptual clarity may be important factors in predicting personalized value interpretations. This raises questions about whether personalization works better for users with more systematic or well-articulated value systems. Moreover, variation across individuals suggests that some users may benefit more from personalized systems than others. This could create new forms of digital inequality where users with more "learnable" patterns receive better AI assistance.

Low performance The observed results are far from perfect, peaking at an F_1-score of 0.444. On the one hand, this speaks of the inherent difficulty in predicting subjective, individual value interpretations. On the other hand, this may suggest that the SEAT dimensions alone may be insufficient proxies for the full spectrum of human value diversity. Additional dimensions may help capture the richness of human value prediction more effectively—among others, moral dilemmas such as trolley-type scenarios that reveal utilitarian-versus-deontological leanings [35]; Moral Foundation scores (care, Fairness, Loyalty, etc.) obtained with the MFQ [15]; Big Five personality traits, whose agreeableness and openness facets systematically predict value priorities [23]; and national cultural value profile (e.g. Hofstede's dimensions) that contextualize individual judgments within broader societal norms [17]. Integrating these complementary or additional variables may capture subtle facets of human value diversity and, in turn, enhance the robustness of value prediction models beyond what the SEAT dimensions alone can deliver.

Generalization Our study used a single model architecture with specific prompting strategies. The effectiveness of our approach may vary with different model

architectures, sizes, or prompting techniques. We suggest that researchers test whether larger and different models exhibit similar patterns or distinct scaling behaviors. Furthermore, the current sample size (50 data points, 5 annotators) limits generalizability; the consistency of patterns within this controlled setting provides a solid foundation for future large-scale validation.

6 Conclusions and Future Work

Our study demonstrates that a language model can leverage multi-dimensional subjective annotations to recognize individual value interpretations, representing a promising step toward more nuanced and individualized AI systems. The better results of combining SEAT dimensions over individual ones suggest that effective subjective value prediction requires capturing the complex interplay of multiple interpretive lenses.

This paper lays the groundwork for future research on value prediction that transcends demographic proxies to adopt more sophisticated individual modeling. While significant challenges remain, particularly in areas such as fairness and transparency, the potential benefits of more personalized and contextually aware AI systems that can predict and possibly align to individual values make this an important direction for continued investigation.

Based on our findings and limitations, we identify several directions for future research. Testing the approach with more annotators across diverse demographic and cultural backgrounds would help validate whether the synergistic effects of SEAT dimensions generalize beyond Western, educated populations and whether cultural differences in value expression affect individual value prediction. This evaluation is crucial for ensuring that personalized AI systems do not inadvertently disadvantage certain populations or create new forms of digital inequality where users with more "learnable" annotation patterns receive better AI assistance.

Assessing performance across different text types and domains would provide a cross-domain evaluation system, highlighting the importance of context-specific capturing of individual value strategies. For instance, value expressions in social media posts may differ significantly from those in formal policy documents, requiring different approaches. Multi-modal integration studies could investigate whether incorporating non-textual cues (such as voice tone, facial expressions, or physiological signals) might provide additional signals to predict value interpretations. This would be particularly relevant for applications where users interact with AI systems through multiple modalities.

Acknowledgments. Enrico Liscio's work was conducted as part of AlgoSoc, a collaborative 10-year research program on public values in the algorithmic society, and the Hybrid Intelligence Centre, both funded by the Dutch Ministry of Education, Culture and Science (OCW) under the Gravitation programme (project numbers 024.005.017 and 024.004.022). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OCW or those of the AlgoSoc consortium as a whole.

References

1. Ajzen, I.: The theory of planned behavior. Organizational Behavior and Human Decision Processes **50**(2), 179–211 (1991)
2. Alshomary, M., Baff, R.E., Gurcke, T., Wachsmuth, H.: The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. pp. 8782–8797. ACL '22, Association for Computational Linguistics, Dublin, Ireland (2022)
3. Araque, O., Gatti, L., Kalimeri, K.: MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. Knowledge-Based Systems **191**, 1–29 (2020)
4. Araque, O., Gatti, L., Kalimeri, K.: The Language of Liberty: A preliminary study. In: Companion Proceedings of the Web Conference 2021. pp. 1–4. WWW '21 Companion, Association for Computing Machinery, Ljubljana, Slovenia (2021)
5. Bahgat, M., Wilson, S.R., Magdy, W.: Towards Using Word Embedding Vector Space for Better Cohort Analysis. In: Proceedings of the International AAAI Conference on Web and Social Media. pp. 919–923. AAAI Press, Atlanta, Georgia (2020)
6. Bench-Capon, T.J.M.: Persuasion in Practical Argument Using Value-based Argumentation Frameworks. Journal of Logic and Computation **13**(3), 429–448 (06 2003)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Cabitza, F., Campagner, A., Basile, V.: Toward a perspectivist turn in ground truthing for predictive computing. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6860–6868 (2023)
9. Creanga, C., Dinu, L.P.: Designing nlp systems that adapt to diverse worldviews. arXiv preprint arXiv:2405.11197 (2024)
10. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A dataset of fine-grained emotions. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4040–4054. Association for Computational Linguistics, Online (Jul 2020)
11. Dong, Y.R., Hu, T., Collier, N.: Can llm be a personalized judge? arXiv preprint arXiv:2406.11657 (2024)
12. Fleisig, E., Abebe, R., Klein, D.: When the majority is wrong: Modeling annotator disagreement for subjective tasks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6715–6726. Association for Computational Linguistics, Singapore (2023)
13. Garten, J., Hoover, J., Johnson, K.M., Boghrati, R., Iskiwitch, C., Dehghani, M.: Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior research methods **50**(1), 344–361 (2018)
14. Goyal, N., Kivlichan, I.D., Rosen, R., Vasserman, L.: Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. Proceedings of the ACM on Human-Computer Interaction **6**, 1–28 (2022)
15. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In: Advances in Experimental Social Psychology, vol. 47, pp. 55–130. Elsevier, Amsterdam, the Netherlands (2013)

16. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review **108**(4), 814 (2001)
17. Hofstede, G.: Culture and organizations. International studies of management & organization **10**(4), 15–41 (1980)
18. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behavior Research Methods **53**, 232–246 (2021)
19. Huang, X., Wormley, A., Cohen, A.: Learning to Adapt Domain Shifts of Moral Values via Instance Weighting. In: Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22). pp. 121–131. ACM, Barcelona, Spain (2022)
20. Itten, A., Mouter, N.: When digital mass participation meets citizen deliberation: combining mini-and maxi-publics in climate policy-making. Sustainability **14**(8), 4656 (2022)
21. Jafari, A.R., Rajapaksha, P., Farahbakhsh, R., Li, G., Crespi, N.: Unveiling human values: Analyzing emotions behind arguments. Entropy **26**(4), 327 (2024)
22. Jiang, A., Vitsakis, N., Dinkar, T., Abercrombie, G., Konstas, I.: Re-examining sexism and misogyny classification with annotator attitudes. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 15103–15125. Association for Computational Linguistics, Miami, Florida, USA (2024)
23. John, O.P., Robins, R.W., Pervin, L.A.: Handbook of personality: Theory and research. Guilford Press (2010)
24. 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)
25. Kiesel, J., Alshomary, M., Mirzakhmedova, N., Heinrich, M., Handke, N., Wachsmuth, H., Stein, B.: 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 (Jul 2023). https://doi.org/10.18653/v1/2023.semeval-1.313, https://aclanthology.org/2023.semeval-1.313
26. Kiesel, J., et al.: Overview of Touché 2024: Argumentation Systems. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Nunzio, G.M.D., Soulier, L., Galuscakova, P., Herrera, A.G.S., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024). Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2024)
27. Kobbe, J., Rehbein, I., Hulpuș, I., Stuckenschmidt, H.: Exploring morality in argumentation. In: Proceedings of the 7th Workshop on Argument Mining. pp. 30–40. Association for Computational Linguistics, Online (Dec 2020)
28. Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C.M., Kalimeri, K., Murukannaiah, P.K.: What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers. pp. 14113–14132. ACL '23, ACL, Toronto, Canada (2023)
29. Liscio, E., Dondera, A.E., Geadau, A., Jonker, C.M., Murukannaiah, P.K.: Cross-domain classification of moral values. In: Findings of the Association for Compu-

tational Linguistics: NAACL 2022. pp. 2727–2745. Association for Computational Linguistics, Seattle, WA, USA (2022)
30. Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R.I., Jonker, C.M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Murukannaiah, P.K.: Value inference in sociotechnical systems. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. pp. 1774–1780 (2023)
31. Liscio, E., Lorandi, M., Murukannaiah, P.K.: News is More than a Collection of Facts: Moral Frame Preserving News Summarization. In: Second Conference on Language Modeling (2025)
32. Liscio, E., Siebert, L.C., Jonker, C.M., Murukannaiah, P.K.: Value preferences estimation and disambiguation in hybrid participatory systems. Journal of Artificial Intelligence Research **82**, 819–850 (2025)
33. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. Mining text data pp. 415–463 (2012)
34. Mackie, J.L.: The subjectivity of values. Essays on moral realism pp. 95–118 (1988)
35. Marcus, R.B.: Moral dilemmas and consistency. The Journal of Philosophy **77**(3), 121–136 (1980)
36. van der Meer, M., Vossen, P., Jonker, C., Murukannaiah, P.: Do Differences in Values Influence Disagreements in Online Discussions? In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 15986–16008. EMNLP '23, ACL, Singapore (2023)
37. Mishra, R., Morren, M.: Eric fromm at touché: Prompts vs finetuning for human value detection. In: 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. pp. 3433–3446. CEUR Proceedings (2024)
38. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational intelligence **29**(3), 436–465 (2013)
39. Orlikowski, M., Pei, J., Röttger, P., Cimiano, P., Jurgens, D., Hovy, D.: Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. arXiv preprint arXiv:2502.20897 (2025)
40. Orlikowski, M., Röttger, P., Cimiano, P., Hovy, D.: The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada (2023)
41. Pavan, M.C., Santos, V.G., Lan, A.G.J., Martins, J., Santos, W.R., Deutsch, C., Costa, P.B., Hsieh, F.C., Paraboni, I.: Morality Classification in Natural Language Text. IEEE Transactions on Affective Computing **3045**(c) (2020)
42. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC. Mahway: Lawrence Erlbaum Associates **71** (2001)
43. Ponizovskiy, V., Ardag, M., Grigoryan, L., Boyd, R., Dobewall, H., Holtz, P.: Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text. European Journal of Personality **34**(5), 885–902 (2020)
44. Qiu, L., Zhao, Y., Li, J., Lu, P., Peng, B., Gao, J., Zhu, S.C.: ValueNet: A New Dataset for Human Value Driven Dialogue System. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. pp. 11183–11191. AAAI '22 (2022)
45. Rezapour, R., Shah, S.H., Diesner, J.: Enhancing the measurement of social effects by capturing morality. In: Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis. pp. 35–45 (2019)
46. Rokeach, M.: The nature of human values. Free press (1973)

47. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
48. Senthilkumar, R.A., Homayounirad, A., Siebert, L.C.: Leveraging large language models to identify the values behind arguments. In: International Workshop on Value Engineering in AI. pp. 87–103. Springer (2024)
49. Sorensen, T., Jiang, L., Hwang, J.D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., et al.: Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 19937–19947 (2024)
50. Wan, R., Kim, J., Kang, D.: Everyone's voice matters: quantifying annotation disagreement using demographic information. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI'23, AAAI Press (2023)
51. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. Transactions on Machine Learning Research (2022)
52. Wilson, S.R., Shen, Y., Mihalcea, R.: Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In: Proceedings of the 10th International Conference on Social Informatics. pp. 455–470. SocInfo '18, Springer, St. Petersburg, Russia (2018)
53. Zhang, J.: Sentiment and language: A socio-semiotic analysis. Philosophy Journal (2013)
54. Zhang, Z., Rossi, R.A., Kveton, B., Shao, Y., Yang, D., Zamani, H., Dernoncourt, F., Barrow, J., Yu, T., Kim, S., et al.: Personalization of large language models: A survey. arXiv preprint arXiv:2411.00027 (2024)

A Annotation Procedure

The annotation procedure involved five computer science students from Europe, each possessing prior knowledge in natural language processing. Among them, Annotator 5 was more familiar with the value annotation task. To ensure consistency, all annotators first reviewed the dataset and labels for each task (see Table 4). They were then instructed to examine each sentence in the dataset and assign zero or more labels from the predefined category set. Since the sentences could be interpreted in multiple ways, annotators were reminded that their judgments were inherently subjective and should reflect their individual perspectives.

Table 4. Overview of the annotated labels.

Category	**Labels (comma-separated)**
Sentiments	Very negative, Somewhat negative, Neutral, Somewhat positive, Very positive
Emotions	admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise
Arguments	If the sentence contains an argument, the annotation consists of one or more extracted text spans corresponding to its premises (e.g., "I care about sustainability"); if no argument is present, the annotation is *None*. Only premises are extracted, conclusions are not considered.
Topics	Municipality and residents engagement in the energy sector, Energy storage and supplying energy in The Netherlands, Wind and solar energy, Market Determination Dynamics, Landscapes and windmills tourism, Hydrogen energy pipeline networks
Values	Be creative, Be curious, Have freedom of thought, Be choosing own goals, Be independent, Have freedom of action, Have privacy, Have an exciting life, Have a varied life, Be daring, Have pleasure, Be ambitious, Have success, Be capable, Be intellectual, Be courageous, Have influence, Have the right to command, Have wealth, Have social recognition, Have a good reputation, Have a sense of belonging, Have good health, Have no debts, Be neat and tidy, Have a comfortable life, Have a safe country, Have a stable society, Be respecting traditions, Be holding religious faith, Be compliant, Be self-disciplined, Be behaving properly, Be polite, Be honoring elders, Be humble, Have life accepted as is, Be helpful, Be honest, Be forgiving, Have the own family secured, Be loving, Be responsible, Have loyalty towards friends, Have equality, Be just, Have a world at peace, Be protecting the environment, Have harmony with nature, Have a world of beauty, Be broadminded, Have the wisdom to accept others, Be logical, Have an objective view

B Prompt Outline

Each prompt begins with a brief instruction that defines the task for the model. We prefix: "You are an expert value annotator. Your task is to extract the most relevant value labels from a given sentence.". This primes the model to act as a value classifier. The prompt explicitly instructs the model to output only a list of values from this predefined set and in a specific format (a Python list of strings) without any extra commentary. If auxiliary task annotations are available, the prompt includes them as additional context. We insert a line after the sentence, saying, for instance: "Emotion Annotations for this sentence: approval, curiosity" (if using emotion context). These annotations are provided in natural language form, one per line, so the model knows the sentence's sentiment category, emotion tags, etc. We explicitly tell the model, "You should also consider your previous annotations on the [Task] detection task," when such context is present, to encourage it to make use of that information. Importantly, we format the prompt to resemble a few-shot learning scenario: we can include a few demonstration examples before asking the model to label the new sentence. Each example consists of a prior sentence, its annotation(s) for the auxiliary task(s), and the correct value labels for that example. For instance, in a per-task prompt, an example might look like: Sentence: "The city reduced speed limits to improve safety." Sentiment Annotation: "Somewhat positive" Values: ["Have a safe country", "Be responsible"].

We then provide the new test sentence with its annotation and ask for the Values output. In an all-tasks prompt, each example and the test input would instead list all four annotations (Argument, Emotion, Sentiment, Topic) followed by the Values for that example. By showing complete input-output examples, the model can learn the pattern that connects task annotations and sentences to the correct value labels.