Hitachi at SemEval-2023 Task 4: Exploring Various Task Formulations Reveals the Importance of Description Texts on Human Values

Masaya Tsunokake, Atsuki Yamaguchi, Yuta Koreeda,
Hiroaki Ozaki and **Yasuhiro Sogawa**
Research and Development Group, Hitachi, Ltd.
Kokubunji, Tokyo, Japan
{masaya.tsunokake.qu, atsuki.yamaguchi.xn, yuta.koreeda.pb,
hiroaki.ozaki.yu, yasuhiro.sogawa.tp}@hitachi.com

Abstract

This paper describes our participation in SemEval-2023 Task 4, ValueEval: Identification of Human Values behind Arguments. The aim of this task is to identify whether or not an input text supports each of the 20 pre-defined human values. Previous work on human value detection has shown the effectiveness of a sequence classification approach using BERT. However, little is known about what type of task formulation is suitable for the task. To this end, this paper explores various task formulations, including sequence classification, question answering, and question answering with chain-of-thought prompting and evaluates their performances on the shared task dataset. Experiments show that a zero-shot approach is not as effective as other methods, and there is no one approach that is optimal in every scenario. Our analysis also reveals that utilizing the descriptions of human values can help to improve performance.

1 Introduction

SemEval-2023 Task 4 involves detecting human values behind argumentative sentences in English. Given a conclusion, stance, and premise, the task is to classify whether they entail each of the 20 pre-defined values (Kiesel et al., 2023). The performance improvement in this task can contribute to more precise analysis of argumentative texts using argument and opinion mining techniques.

In this paper, we explore various task formulations for detecting human values in texts, including sequence classification, question answering, and question answering with chain-of-thought prompting. Previous work on human value detection (Kiesel et al., 2022) thus far has only investigated the effectiveness of sequence classification using BERT (Devlin et al., 2019), and it is still unknown what type of task formulation is effective for human value detection.

To this end, we test a variety of pretrained language models. For sequence classification, we employ encoder-based models such as RoBERTa (Liu et al., 2019) and DeBERTaV3 (He et al., 2021). For question answering, we test both encoder-decoder and decoder models, including T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and GPT-3 (text-davinci-003) (Brown et al., 2020). For question answering with chain-of-thought prompting, we use GPT-3 (text-davinci-003) in a zero-shot manner. We measure the detection performance of all approaches on three shared task datasets (Mirzakhmedova et al.), including the two optional test datasets: Nahj al-Balagha and New York Times. We also verify the effectiveness of three additional training strategies for sequence classification: loss weighting, pretraining on a similar corpus: ValueNet (Qiu et al., 2022), and adding a value description to an input.

In the competition[1], our approach using $\text{RoBERTa}_{\text{LARGE}}$ trained with sequence classification and weighted loss achieved macro F1 scores of 0.51 and 0.34 on the Main and New York Times test datasets and was ranked 7th and 1st, respectively. For the Nahj al-Balagha dataset, the submitted approach using a mixture of the highest performing models with respect to each value on the validation set demonstrated a macro F1 score of 0.34 and was ranked 4th.

Our contributions are three-folds.

- This study is the first to explore various task formulations for detecting human values in argumentative texts (Section 2).
- We conduct five analyses to examine the behavior of our approaches and verify the effec-

[1]This paper reports the results not only with our submitted systems but also other approaches for a comprehensive comparison of various task formulations. Table 4 includes the detailed results of our submitted system on the Main test dataset, and Appendix A details our submitted systems.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1723–1735

tiveness of three additional training strategies for sequence classification (Section 5).

- We demonstrate that utilizing the descriptions of human values can improve detection performance (Section 5.2 and 5.3).

2 Methodology

This section explains our approaches in this work. Table 1 provides an overview of each approach with the input and output formats and corresponding example model names.

2.1 Sequence Classification

Our first approach is sequence classification (**SC**), which has been tested in previous work (Kiesel et al., 2022). Let $s = [w_0, \ldots, w_n]$ be an input sequence, where w denotes a token. In SC, a model takes s as input and outputs a hidden vector $h_s = [e_0, \ldots, e_n]$ corresponding to s. We put a linear layer on top of the model for final classification. Because the task is formulated as multi-label classification, the model computes the score of each value category using a sigmoid function and is trained with the binary cross-entropy loss averaged over all 20 classes (values).

2.2 Question Answering

Because the shared task is fine-grained and has a variety of value categories, it may be difficult for a model to accurately detect human values in texts only using a (conclusion, stance, premise) triple as in SC. If the model can capture subtle differences in value definitions, the correct values in texts can be identified more easily. To this end, we investigate adding descriptions of values to the model's input.

Here, we formulate the task as question answering (**QA**) to accommodate a value description in input. We obtain all descriptions of the 20 values from the shared task website[2] and generate a yes/no question for each value description given a (premise, stance, conclusion) triple (see Table 1 for an example). This means we create 20 samples per triple and need to feed them into a model one by one to obtain predictions for a particular triple. We fine-tune both encoder-decoder and decoder models, including T5, BART, and GPT-3 (text-davinci-003).

2.3 Question Answering with Chain-of-Thought Prompting

Previous studies (Kojima et al., 2022; Wei et al., 2022) have reported that large language models (LLMs), e.g., GPT-3, are effective even in complicated tasks that need reasoning. We assume that the shared task falls in this category; thus, utilizing an LLM can be an effective approach to detecting human values behind texts. One practical problem arises in using an LLM: financial costs. To reduce the cost as much as possible, we use a zero-shot approach with chain-of-thought prompting (Kojima et al., 2022), which shows significant performance improvement over zero-shot LLMs on various reasoning benchmark datasets. This approach is a two-step procedure exemplified in Table 1 (**QA w/ CoT**). Specifically, we employ a similar prompt used for the question answering approach as a first query with the addition of "Let's think step by step." to its end. After obtaining the corresponding output from GPT-3 (text-davinci-003), we concatenate the first prompt, output, and "Therefore, the answer (YES or NO) is" and feed it to GPT-3 again to obtain the final decision.

3 Experimental Setup

Data We used the official training dataset for training and the official validation dataset for monitoring generalization performance. For evaluation, we used the Main, Nahj al-Balagha, and New York Times test datasets. Details of each dataset can be found in the task paper (Kiesel et al., 2023) and on the website[3].

Evaluation Metrics The evaluation metrics are a macro F1 score averaged over all value categories and F1 score for each value.

Models Table 2 presents the list of models tested for each approach. For simplicity, we only compare the results of **bold** models that showed the highest macro F1 score on the validation set in the remainder of this paper.

Implementation Details We implemented each approach with PyTorch (Paszke et al., 2019) and Hugging Face transformers (Wolf et al., 2020) except for GPT-3. For GPT-3, we used the official API provided by OpenAI to use both the base and

[2] https://touche.webis.de/semeval23/touche23-web/index.html#task

[3] https://touche.webis.de/semeval23/touche23-web/index.html

Approach / Input	Output	Example Models
Sequence Classification (SC) (§2.1) I am {stance} the argument: {conclusion}. And, I have the premise: {premise}.	[0, 1] for 20 values	roberta-large microsoft/deberta-v3-large
Question Answering (QA) (§2.2) Question: I am {stance} the conclusion that {conclusion}. That's because I have the premise: "{premise}." Do you think I am explicitly supporting the value that {value}? Answer:	[Yes, No]	t5-large facebook/bart-large
Question Answering with Chain-of-Thought (QA w/ CoT) (§2.3) *# Input 1* Question: I {agree | disagree} with the conclusion that {conclusion}. I have the premise: "{premise}" Do you think that I EXPLICITLY support the value about {value}? Answer: Let's think step by step.	*# Output 1* *# Reasoning texts generated by GPT-3*	GPT-3 (text-davinci-003)
# Input 2 [Input 1] + [Output 1] + "Therefore, the answer (YES or NO) is"	*# Output 2* [Yes, No]	
Sequence Classification with Human Value Description (§5.3) I am {stance} the argument: {conclusion}. And, I have the premise: {premise}. So, I am supporting the value that {value}.	[0, 1]	roberta-large microsoft/deberta-v3-large

Table 1: Input/Output format of each approach. "stance", "conclusion", and "premise" are provided in the official datasets. For QA with Chain-of-Thought, we assigned "agree" when "stance" is "in favor of", while we put "disagree" if "stance" is "against". We obtained value descriptions from the shared task website.[1] Each model identifier corresponds to the model name in the transformers library, except for GPT-3.

Approach	Models
SC	**roberta-large**, microsoft/deberta-v3-large, google/rembert, nghuyong/ernie-2.0-large-en, microsoft/infoxlm-large, google/electra-large-discriminator, google/canine-s
QA	facebook/bart-large, t5-base, t5-large, google/flan-t5-base, **google/flan-t5-large**, GPT-3 (text-davinci-003) (on some values)
QA w/ CoT	**GPT-3** (text-davinci-003) (zero-shot)

Table 2: List of models tested in this paper. In the experiments, we used **bold** models that showed the highest F1 macro score on the validation set for each approach.

fine-tuned text-davinci-003 models. The hyperparameter settings are detailed in Appendix A.

4 Results

Table 3 presents the macro F1 and F1 score of each value category with respect to the three approaches introduced in Section 2. We also include the results of the baseline ("1-Baseline") that always assigns 1 to all instances for reference. Overall, QA exhibited the highest macro F1 scores of 0.54 and 0.42 on the Main and Nahj al-Balagha datasets, respectively. SC showed the highest macro F1 score of 0.28 on the New York Times dataset. QA w/ CoT resulted in the lowest macro F1 score among the three approaches though it outperformed the baseline. This indicates that it is essential to train a model on annotated data for human value detection in order to effectively identify values behind argumentative texts.

As for the value-wise F1 results, the most effective approach depends on the value categories and datasets, indicating that there exists no one-size-fits-all approach for human value detection. However, we see a general trend that QA slightly outperforms SC the majority of the times. The main difference between the two is whether or not an input has a value description. Thus, we hypothesize that utilizing the descriptions of human values can boost the detection performance. We verify our hypothesis in the Sections 5.2 and 5.3.

5 Analysis

5.1 Mitigating Data Imbalances

As shown in Table 4, the official dataset is imbalanced with respect to all value categories, which makes it challenging for a model to acquire the useful representations for the task due to the limited number of positive samples. We found in the preliminary experiments that value-wise F1 scores on the validation data were often low for value categories with low frequency. To address with this problem, we utilized an adjusted loss function with respect to label frequency in SC[4] and verified its effectiveness on the Main test dataset. For im-

[4]We only targeted SC because it achieved the highest macro F1 score on the validation data.

Test set / Approach	All	1. Self-direction: thought	2. Self-direction: action	3. Stimulation	4. Hedonism	5. Achievement	6. Power: dominance	7. Power: resources	8. Face	9. Security: personal	10. Security: societal	11. Tradition	12. Conformity: rules	13. Conformity: interpersonal	14. Humility	15. Benevolence: caring	16. Benevolence: dependability	17. Universalism: concern	18. Universalism: nature	19. Universalism: tolerance	20. Universalism: objectivity
Main																					
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
SC	.51	.49	.69	.16	.24	.61	.43	.49	.29	.72	.63	.59	.58	.45	.16	.54	.42	.73	.83	.42	.55
QA	.54	.57	.70	.15	.47	.66	.40	.54	.31	.78	.67	.66	.57	.31	.10	.50	.33	.76	.86	.47	.53
QA w/ CoT	.31	.26	.47	.20	.06	.29	.07	.05	.22	.43	.42	.29	.31	.08	.06	.36	.21	.59	.58	.31	.40
Nahj al-Balagha																					
1-Baseline	.13	.04	.09	.01	.03	.41	.04	.03	.23	.38	.06	.18	.13	.06	.13	.17	.12	.12	.01	.04	.14
SC	.32	.09	.32	.40	.25	.61	.11	.00	.50	.41	.20	.45	.25	.33	.20	.33	.18	.23	.50	.00	.25
QA	.42	.14	.44	.50	.57	.62	.50	.20	.59	.48	.40	.56	.35	.29	.35	.40	.37	.29	.25	.00	.40
QA w/ CoT	.18	.02	.13	.00	.11	.31	.00	.18	.30	.40	.10	.26	.18	.07	.16	.22	.13	.19	.06	.10	.22
New York Times																					
1-Baseline	.15	.05	.03	-	.03	.28	.03	-	.05	.51	.20	-	.07	.03	.12	.12	.26	.24	.03	.03	.33
SC	.28	.20	.17	-	.00	.40	.00	-	.00	.55	.33	-	.24	.67	.18	.14	.33	.54	.29	.00	.37
QA	.25	.22	.20	-	.00	.34	.00	-	.00	.61	.41	-	.33	.00	.00	.11	.29	.48	.50	.00	.43
QA w/ CoT	.17	.11	.00	-	.00	.00	.00	-	.00	.49	.22	-	.07	.04	.35	.12	.25	.35	.13	.08	.33

Table 3: F1 scores of our approaches on three test datasets. The highest scores for each value category are underlined.

plementation, we adjusted the weight of positive samples in each value category according to the frequency in the training data when computing the binary cross-entropy loss.

As Table 4 shows, SC with adjusted loss slightly outperformed SC on the validation dataset, but not on the Main test dataset in most cases. The Pearson and Spearman correlations between the ratio of positive samples in the training data for each value category and its performance improvement over SC were extremely small, -0.09 and -0.07, respectively. Although the loss is adjusted to improve performance on low-frequency labels, we did not see a correlation between the ratio of positive samples and performance improvement over SC. These results suggest that the adjusted loss function does not have much impact on the performance improvement.

5.2 Pre-training on ValueNet

It is very unlikely that argumentative texts in the official datasets explicitly mention underlying human values. Therefore, it may be difficult for models to acquire the semantic representations for effectively identifying human values. To address this challenge, we pre-trained an SC model on a similar corpus, ValueNet (Qiu et al., 2022), motivated by the assumption that it could provide the model with prior knowledge of human values and thus improve performance. ValueNet contains short texts about social scenarios (e.g., "applying to a far-away university against my dad's wishes") and the relatedness (-1/0/1) between the texts and specific human values. The value categories in ValueNet are similar to those in the shared task as both are based on Schwartz's theory (Schwartz, 2012).[5]

During pre-training, a model predicted the relatedness (-1/0/1) between a given scenario text and value and was trained with the mean squared error loss, following Qiu et al. (2022). We generated input texts with the following template.

```
The premise {scenario text} is supporting
the value towards {value}.
```

As a `{value}`, we used "Definition goal" defined by Schwartz (2012). The model was then fine-tuned on the shared task dataset as described in 2.1. We used RoBERTa $_{\text{LARGE}}$, which achieved the highest macro F1 score on the validation set for the experiments. During fine-tuning, we applied the adjusted loss as this approach exhibited the best F1 score on the validation set (see Section 5.1).

As Table 4 shows, SC with the adjusted loss and ValueNet outperformed or was on par with SC

[5]Some values in this task are integrated in ValueNet, and some do not exist in ValueNet. Appendix E describes the correspondence between the two datasets.

Approach / Ratio in training data	All (Val)	All (Test)	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
Sequence Classification (SC)																						
Original	.527	.51	.49	.69	.16	.24	.61	.43	.49	.29	.72	.63	.59	.58	.45	.16	.54	.42	.73	.83	.42	.55
w/ adjusted loss†	.532	.50	.48	.66	.22	.23	.61	.43	.45	.32	.74	.63	.57	.54	.47	.15	.53	.36	.74	.81	.42	.55
w/ adjusted loss & ValueNet	.525	.52	.51	.65	.26	.37	.59	.49	.52	.31	.73	.63	.57	.55	.48	.22	.53	.37	.72	.84	.42	.55
w/ HVD	.551	.56	.45	.71	.29	.44	.66	.46	.56	.37	.74	.68	.67	.59	.56	.26	.53	.43	.76	.85	.43	.61
Question Answering (QA)																						
Original	.523	.54	.57	.70	.15	.47	.66	.40	.54	.31	.78	.67	.66	.57	.31	.10	.50	.33	.76	.86	.47	.53
Fine-tuned GPT-3	-	-	.58	-	.17	-	-	-	-	.27	-	-	.67	-	-	.39	-	.29	-	-	.39	-
Ratio in training data	-	-	.18	.26	.05	.03	.28	.11	.12	.07	.37	.32	.11	.22	.04	.07	.25	.15	.39	.08	.12	.20

Table 4: Performances of our model variants on Main test dataset. The row with † corresponds to the results of the submitted system.

with the adjusted loss on the Main test dataset in many values. The performances of value categories whose positive ratio in the training data was lower than 0.1 improved compared to those with SC w/ adjusted loss except for "Face." The Pearson and Spearman correlations between the ratio of positive samples in the training data for each value category and its performance improvement over SC w/ adjusted loss were -0.56 and -0.57, respectively. These results suggest that ValueNet improves the performance of low-frequency value categories.

5.3 Sequence Classification with Human Value Descriptions

As observed in Section 5.2, the prior knowledge of human values seems to improve the detection performance of low-frequency human values. Here, we try to incorporate **H**uman **V**alue **D**escriptions into SC (**SC w/ HVD**), which should help a model capture the semantic representations to identify human values. We followed the same setting as that of SC except for the input and output formats shown in Table 1, and employed *deberta-v3-large*, which achieved the highest macro F1 score on the validation dataset. This time, we need to feed an individual query for each value into the model to obtain a prediction.

As seen in Table 4, SC w/ HVD exhibited the highest macro F1-score and outperformed SC on almost all value categories. In addition, the Pearson and Spearman correlations between the ratio of positive samples in the training data for each value category and its performance improvement over SC were -0.55 and -0.61, respectively. These results indicate that incorporating human value descriptions into an input is effective for identifying human values behind argumentative texts, especially for low-frequency labels.

5.4 Fine-tuning with GPT-3

As explained in Table 2, we only fine-tuned GPT-3 (text-davinci-003) on a limited number of value categories due to budget constraints, including "Self-direction: thought," "Stimulation," "Face," "Tradition," "Humility," "Benevolence: dependability," and "Universalism: tolerance," which exhibited low macro F1 score in the validation data regardless of the approaches used.

The results in Table 4 show that some F1 scores improved, but their differences were small except "Humility." This suggests that although fine-tuned GPT-3 does not always show the highest result, there might be human values that fine-tuned GPT-3 is good at identifying. We leave the thorough investigation of such values as future work.

5.5 Prediction Tendency of Question Answering with Chain-of-Thought Prompting

As shown in Table 3, zero-shot question answering with chain-of-thought prompting (QA w/ CoT) by GPT-3 exhibited the lowest macro F1 scores on all datasets. To determine the causes of the low performance of QA w/ CoT, we examined its output trends and error cases. To simplify the discussion, we focused on the validation and Main test datasets.

Table 5 presents the types of answers generated by QA w/ CoT, the number of each type, and the example answers. While the ratio of "Yes" in gold answers is 0.168 on the validation dataset, the ratio of answers starting with "yes" by QA w/ CoT is 0.291. Thus, QA w/ CoT tends to answer "Yes,"

Types of answers	Ratio in the validation dataset	Ratio in the Main test dataset	Example answers and the number of that in the Main test dataset
Answers starting with "yes"	0.291 (11,039/37,920)	0.304 (9,583/31,520)	"YES." (5,889) "YES, you are implicitly supporting" (862)
Answers starting with "no"	0.676 (25,620/37,920)	0.658 (20,742/31,520)	"NO." (16,634) "NO, you do not explicitly" (2,570)
Other	0.033 (1,261/37,920)	0.038 (1,195/31,520)	"not explicitly stated in the premise" (207) "not clear-cut. It" (135), "up to you." (70)

Table 5: Types of answers which correspond to *# Output 2* of QA w/ CoT in Table 1.

Prompt (*# Input 1* in Table 1): Question: I agree with the conclusion that the use of public defenders should be mandatory. I have the premise: "the use of public defenders should be mandatory because some people don't have money for a lawyer and this would help those that don't." Do you think I EXPLICITLY support the value that it is good to be successful in accordance with social norms? Answer: Let's think step by step.

Reasoning Text (*# Output 1* in Table 1): The premise states that the use of public defenders should be mandatory because it would help those who don't have money for a lawyer. This premise does not explicitly support the value that it is good to be successful in accordance with social norms. However, the conclusion that the use of public defenders should be mandatory implies that it is good to be successful in accordance with social norms, as it implies that those who don't have money for a lawyer should be given the same opportunity to be successful as those who do. Therefore, while the premise does not explicitly support the value that it is good to be successful in accordance with social norms, the conclusion does

■: Paraphrasing the given premise, ■: Stating that the premise does not explicitly support the given value,
■: Explaining the interpretation that associates the given conclusion with the value

Table 6: An example reasoning text generated by QA w/ CoT when the model's answer started with "yes."

excessively.

In addition, QA w/ CoT sometimes generates answers that do not follow the instructions. As for an example of an answer starting with "yes," the model answered whether to implicitly support the values, although the first prompt (*# Input 1*) in Table 1 asked whether to explicitly support the values. For all examples of the "Other" type in Table 5, the answers were ambiguous, although the second prompt (*# Input 2*) in Table 1 instructed the model to answer with "Yes" or "No."

To understand how the model wrongly outputs answers starting with "yes" on the validation dataset (i.e., false positive cases), we analyzed the intermediate reasoning texts generated by the model based on the first prompt (*# Input 1*) in Table 1. Some of the reasoning texts stated that a given premise did not explicitly support the given value but the given premise or conclusion implicitly supported the value. Table 6 presents one such example. In these error cases, the model generally paraphrased the given premise first, stated that the premise did not explicitly support the given value, and then explained the interpretation that associates the premise or conclusion with the value. The ratio of reasoning text having the strings of "does not explicit" and "however" among the false positives case on the validation dataset was 0.662 (5,277/7,966). To remedy these cases, we need to further investigate how to make the model not allow implicit support for the values, such as through prompt engineering. Post-hoc modification of the answer based on the content of its reasoning texts is also promising.

6 Discussion and Conclusion

This paper has explored three task formulations for human value detection: SC, QA, and QA w/ CoT, and investigated the effectiveness of each approach on the shared task dataset. Experimental results demonstrated that an one-size-fits-all approach does not exist for this task, but supervised learning with the dedicated task data is necessary to obtain decent detection performance.

As mentioned in Section 4, we hypothesized that utilizing the descriptions of human values can boost detection performance, which was supported by the results in Sections 5.2 and 5.3. Although there is no one task formulation that is always effective for all value categories, different value categories can favor different task formulations. For instance, QA largely outperformed SC w/ HVD in "Self-direction: thought" and "Universalism: tolerance," though the only difference between the two is in their architecture. Investigating this trend in detail can be a future research direction.

Acknowledgments

We would like to thank anonymous reviewers for their valuable feedback. We would also like to thank Gaku Morio, Yuichi Sasazawa, and Ken-ichi Yokote for their initial feedback on the experimental design and Dr. Masaaki Shimizu for the maintenance and management of large computational resources.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 Task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. The Touché23-ValueEval dataset for identifying human values behind arguments. *arxiv preprint arXiv:2301.13771*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hyperparameter	Value
Batch size	32
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	2e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Attention Dropout	0.1
Seed	61

Table 7: Hyperparameters of best model on Main and New York Times dataset.

Appendices

A Submitted Systems

Here, we detail our submitted systems that achieved the highest macro F1 score among our runs in each dataset.

A.1 Main

Our most accurate system for the Main test dataset utilized RoBERTa$_{\text{LARGE}}$ trained with sequence classification and loss weighting[6]. We used one NVIDIA Tesla V100 (32GB) and optimized the model with AdamW (Loshchilov and Hutter, 2019). Table 7 shows the hyperparameter settings.

A.2 Nahj al-Balagha

Our most accurate system for the Nahj al-Balagha test dataset used a mixture of models and task formulations. We selected a highest-performing model for each value category on the basis of a F1 score on the validation set. Table 8 shows the detailed configuration for each value category.

A.3 New York Times

Our most accurate system for the New York Times test dataset is the same as that one for the Main test dataset.

B Cost of GPT-3

The total cost of QA w/ CoT was $2324.63 to obtain predictions for the validation set and the three test datasets in a zero-shot manner, while that for QA was $964.04 to fine-tune GPT-3 and make predictions on the test datasets over the seven value categories listed in Table 4.

[6]See details in Section 5.1

C Fine-tuning with Instructions

Some pre-trained models (e.g., Flan-T5) are fine-tuned with texts beginning with instructions, and they have shown promising results on various NLP benchmarks (Chung et al., 2022). Here, we also verified the effectiveness of the QA approach with instructions (**QA w/ Instruct**) in the task of human value detection. The only difference between QA w/ Instruct and QA is whether or not an input has an instruction sentence as demonstrated in Table 18.

As shown in Table 19, QA w/ Instruct outperformed QA on the validation dataset, but it did not always improve F1 scores on the test dataset. While the F1 score of "Conformity: interpersonal" largely improved compared to QA, that of "Hedonism" largely degraded. Therefore, no positive impact on detection performance was observed with respect to the usage of QA w/ Instruct.

D Prediction Tendency

Here, we analyze the tendency of prediction for the sequence classification (SC), question answering (QA), and question answering with chain-of-thought prompting (QA w/ CoT) based on precision and recall to better understand the characteristics of each approach in human value detection. Table 20 presents the precision and recall of the three approaches on the Main test dataset. QA consistently achieved the highest precision for all value categories, while its recall values were consistently lower than those of SC except "Hedonism" and "Power: resources." QA w/ CoT achieved the highest recall values in several value categories but exhibited significantly low precision, indicating that GPT-3 with zero-shot chain-of-thought prompting tends to say "YES" to a given question regardless of correctness in this task. This is consistent with the tendency described in Section 5.5.

One possible reason for the low recall values in QA is that the ratio of "yes" in the training data was quite low (0.17), forcing a model to generate "no" most of the time. We assume that negative sampling can be effective for improving recall values for QA.

E ValueNet

Table 21 presents the correspondence between human values in Task 4 (Mirzakhmedova et al.) and ValueNet (Qiu et al., 2022). Table 21 also presents

Value	Task Formulation	Model	Hyperparameters
Self-direction: thought	SC w/ adjusted loss	google/rembert	Table 9
Self-direction: action	QA	t5-large	Table 10
Stimulation	SC w/ adjusted loss	roberta-large	Table 7
Hedonism	SC w/ adjusted loss	google/rembert	Table 9
Achievement	QA	facebook/bart-large	Table 11
Power: dominance	SC w/ adjusted loss	roberta-large	Table 7
Power: resources	SC	nghuyong/ernie-2.0-large-en	Table 12
Face	SC w/ adjusted loss	roberta-large	Table 7
Security: personal	QA	t5-large	Table 10
Security: societal	QA	t5-large	Table 10
Tradition	SC	roberta-large	Table 13
Conformity: rules	SC w/ adjusted loss and ValueNet	roberta-large	Table 14
Conformity: interpersonal	SC w/ adjusted loss	microsoft/deberta-v3-large	Table 15
Humility	QA	Fine-tuned GPT-3 (text-davinci-003)	Table 16
Benevolence: caring	SC	google/rembert	Table 17
Benevolence: dependability	SC w/ adjusted loss	google/rembert	Table 9
Universalism: concern	QA	facebook/bart-large	Table 11
Universalism: nature	SC	roberta-large	Table 13
Universalism: tolerance	SC	roberta-large	Table 13
Universalism: objectivity	SC w/ adjusted loss and ValueNet	roberta-large	Table 14

Table 8: Model configurations for Nahj al-Balagha test dataset.

Hyperparameter	Value
Batch size	16
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	2.5e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Classification dropout	0.1
Seed	23

Table 9: Hyperparameters for sequence classification with adjusted loss using google/rembert.

the frequency of each human value category on both training dataset.

Hyperparameter	Value
Batch size	16
Maximum number of epochs	5
Learning rate scheduler	linear
Peak learning rate	2e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length for encoder	512
Sequence length for decoder	128
Attention dropout	0.1
Seed	42

Table 10: Hyperparameters for question answering using t5-large.

Hyperparameter	Value
Batch size	64
Maximum number of epochs	5
Learning rate scheduler	linear
Peak learning rate	2e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length for encoder	512
Sequence length for decoder	128
Attention dropout	0.1
Seed	61

Table 11: Hyperparameters for question answering using facebook/bart-large.

Hyperparameter	Value
Batch size	32
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	2e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Attention dropout	0.1
Seed	23

Table 12: Hyperparameters for sequence classification using nghuyong/ernie-2.0-large-en.

Hyperparameter	Value
Batch size	32
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	3e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Attention dropout	0.1
Seed	42

Table 13: Hyperparameters for sequence classification using roberta-large.

Hyperparameter	Value
Batch size	32
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	2e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Attention dropout	0.1
Seed	23

Table 14: Hyperparameters for sequence classification with adjusted loss and ValueNet using roberta-large.

Hyperparameter	Value
Batch size	16
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	1.5e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Attention dropout	0.1
Seed	61

Table 15: Hyperparameters for sequence classification with adjusted loss using microsoft/deberta-v3-large.

Hyperparameter	Value
`max_tokens`	1
`temperature`	0
`n_epochs`	4
`batch_size`	8
`learning_rate_multiplier`	0.1
`prompt_loss_weight`	0.01

Table 16: Hyperparameters for fine-tuned GPT-3 (text-davinci-003).

Hyperparameter	Value
Batch size	16
Maximum number of epochs	10
Learning rate scheduler	linear
Peak learning rate	2.0e-05
Warmup steps	0
Weight decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	256
Classification dropout	0.1
Seed	61

Table 17: Hyperparameters for sequence classification using google/rembert.

Approach / Input	Output	Example Models
Question Answering (QA) (§2.2) Question: I am {stance} the conclusion that {conclusion}. That's because I have the premise: "{premise}." Do you think I am explicitly supporting the value that {value}? Answer:	[Yes, No]	t5-large facebook/bart-large
Question Answering with Instruction (QA w/ Instruct) (Appendix C) Answer the following yes/no question. I am {stance} the conclusion that {conclusion}. That's because I have the premise: "{premise}." Can you say that I am explicitly supporting the value that {value}?	[Yes, No]	google/flan-t5-large

Table 18: Input/Output format of QA approach and QA approach with instructions. "stance", "conclusion", and "premise" are provided in the official datasets. We obtained value descriptions from the shared task website.[1]

Approach	All (Val)	All (Test)	1. Self-direction: thought	2. Self-direction: action	3. Stimulation	4. Hedonism	5. Achievement	6. Power: dominance	7. Power: resources	8. Face	9. Security: personal	10. Security: societal	11. Tradition	12. Conformity: rules	13. Conformity: interpersonal	14. Humility	15. Benevolence: caring	16. Benevolence: dependability	17. Universalism: concern	18. Universalism: nature	19. Universalism: tolerance	20. Universalism: objectivity
Question Answering (QA)																						
Original	.523	.54	.57	.70	.15	.47	.66	.40	.54	.31	.78	.67	.66	.57	.31	.10	.50	.33	.76	.86	.47	.53
w/ Instruction	.525	.53	.54	.70	.17	.36	.68	.40	.53	.34	.78	.67	.66	.57	.38	.10	.51	.31	.75	.87	.47	.54
Fine-tuned GPT-3	-	-	.58	-	.17	-	-	-	-	.27	-	-	.67	-	-	.39	-	.29	-	-	.39	-

Table 19: Performances of QA approach and its variants on Main test dataset.

Metric / Approach	All	1. Self-direction: thought	2. Self-direction: action	3. Stimulation	4. Hedonism	5. Achievement	6. Power: dominance	7. Power: resources	8. Face	9. Security: personal	10. Security: societal	11. Tradition	12. Conformity: rules	13. Conformity: interpersonal	14. Humility	15. Benevolence: caring	16. Benevolence: dependability	17. Universalism: concern	18. Universalism: nature	19. Universalism: tolerance	20. Universalism: objectivity
Precision																					
SC	.46	.40	.67	.27	.24	.54	.40	.45	.28	.63	.53	.50	.49	.47	.15	.45	.36	.63	.83	.35	.57
QA	.59	.57	.75	.41	.59	.63	.63	.47	.48	.73	.67	.66	.54	.65	.20	.51	.52	.69	.90	.53	.72
QA w/ CoT	.23	.16	.33	.16	.05	.43	.11	.09	.15	.37	.30	.23	.24	.04	.06	.27	.13	.50	.53	.20	.32
Recall																					
SC	.56	.63	.71	.12	.23	.70	.46	.54	.29	.84	.78	.72	.70	.43	.19	.66	.51	.89	.83	.52	.54
QA	.49	.57	.65	.09	.38	.68	.30	.63	.23	.83	.66	.66	.59	.21	.07	.49	.24	.85	.82	.42	.41
QA w/ CoT	.46	.64	.81	.27	.08	.22	.06	.04	.42	.52	.71	.40	.44	.87	.07	.57	.45	.73	.65	.65	.53

Table 20: Precision and Recall scores of our approaches on Main test dataset. The highest precision and recall for each value category is underlined.

Human Value	**Task4 Freq.**	**ValueNet Freq.**
Self-direction: thought	988	325
Self-direction: action	1,395	
Stimulation	247	1,281
Hedonism	172	2,160
Achievement	1,512	854
Power: dominance	610	878
Power: resources	625	
Face	382	-
Security: personal	2,000	3691
Security: societal	1,728	
Tradition	568	1,301
Conformity: rules	1,177	1,884
Conformity: interpersonal	207	
Humility	395	-
Benevolence: caring	1,332	7,667
Benevolence: dependability	806	
Universalism: concern	2,081	1,333
Universalism: nature	427	
Universalism: tolerance	664	
Universalism: objectivity	1,054	
ALL	5,393	21,374

Table 21: Correspondence between human values in Task4 and ValueNet and their frequency in the training dataset.