Eric Fromm at Touché: Prompts vs FineTuning for Human Value Detection

Notebook for the Touché Lab at CLEF 2024

Ranjan Mishra[2], Meike Morren[1]

[1]Department of Marketing, School of Business and Economics, Vrije Universiteit Amsterdam
[2]Tinbergen Institute, Netherlands

Abstract

Human values are notoriously difficult to predict as they are often of nuanced nature, culturally embedded and varying across geographies. Generative Large Language Models (LLMs) have become very powerful to mimic how people use language, including value-laden content. We explore the opportunities for supervised fine-tuning and prompt engineering the LLMs in order to better perform a downstream task such as finding value-laden content in text. We compare fine-tuning, which heavily relies on labeled data, to the more flexible approach of prompt engineering that requires less or no labeled data at all. Our goal in this paper is three-fold: 1) assess the capabilities of closed source (GPT-3.5 and GPT-4o) versus open source (Gemini and Llama3) LLMs, 2) analyse the influence of domain-specific information by comparing fine-tuning with prompts, and 3) compare multi-label with single-label approaches.

Keywords

Generative AI, Prompt Engineering, Supervised Fine-Tuning

1. Introduction

In recent years, Generative AI (GenAI) has established itself as the state of the art in the field of Natural Language Processing (NLP) by enabling the creation of highly sophisticated models. These models, often referred to as large language models (LLMs) are extensively trained on a vast amount of data allowing them to deliver state of the art performances on a wide range of NLP tasks across different domains. Given their flexibility in training, these LLMs can be used in different ways. One can fine-tune a pre-trained LLM by using Supervised Fine-Tuning (SFT) with a curated dataset for a specific language modeling task. This allows to obtain an improved version of the model that can perform the particular downstream task better. A more flexible and efficient alternative to fine-tuning is prompt engineering, where prompts are queried in a specific format to the LLMs to generate a desired response. An effective prompt design can mitigate the need for extensive fine-tuning [1]. In our paper, we compare two aforementioned approaches and their influence on the prediction of human values, and we aim to compare this effect across different open and closed source models. We present our results in the CLEF Touché's workshop [2].

2. Background

Higher-order constructs such as human values are likely picked up by transformer models [1]. One way to effectively use these pre-trained transformer models to capture nuances in texts containing human values is through fine-tuning. It involves taking a pre-trained model and adopting to a specific language modeling task by further training on a smaller task-specific dataset. The idea is that, through fine-tuning, the new model captures both the general linguistic features from the pre-training phase as well as improve performance on the specific task by adjusting the model's parameters during fine-tuning.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
674757rm@eur.nl (R. Mishra); meike.morren@vu.nl (M. Morren)
0000-0000-0000-0000 (R. Mishra); 0000-0001-6350-356X (M. Morren)

[1]https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Fine-tuning can be done in various ways, including supervised, unsupervised and semi-supervised approaches. For our case, we use supervised fine-tuning (SFT) which involves further training the model on a labelled dataset to adapt it to a new task. SFT involves selecting a relevant pre-training model for the task, preparing a labeled dataset tailored to the task and then extracting a new model with adjusted parameters that captures the specific nuances of the ask. The main advantage of SFT is the improved performance on the specific task while also being highly resource efficient, requiring less data and computational capacity compared to training a new model from scratch.

In our paper, we use four models, two from OpenAI (GPT 3.5, GPT 4o), Gemini-1.0-pro and Llama3-70B-Instruct, all of which are based on the transformer architecture introduced by Vaswani et al. (2017) [3]. The main advantage of this architecture is its attention mechanism, which allows the models to focus on different parts of the input text selectively, enabling them to capture long-range dependencies and contextual relationships more effectively than previous architectures RNNs and LSTMs. This results in better handling of complex language structures and understanding nuanced meanings. An extensive overview of the performance of these models across different tasks can be studies in their technical reports. Our choice of these particular models is influenced by our familiarity, domain knowledge as well as the prospect of comparison between closed source (GPT 3.5, GPT-4o, Gemini 1.0 pro) and open source (Llama3-70b-instruct) models.

To maximise the benefit from these capabilities of the LLMs, we integrate description of human values directly into the prompts. By including the information about all possible values in the prompt but instructing the model to only report one value per sentence, we ensure that the model assigns a single value to each sentence. Prompting is quite sensitive to the information fed, meaning even small changes in prompts can lead to significantly different results which emphasizes the importance of an effective prompt design [1]. Therefore, we focus on how the information provided in the prompts influences the prediction of human values. The informed zero-shot multilabel (ML) prompt (see Appendix A.1) includes both the task of identifying human values as well as the descriptions of the values given in the coding manual [4]. We also use single label (SL) prompting where we only give a description of one value (see Appendix A.3) which allows us to obtain multiple values per sentence.

Prompting also allows us to give examples with this description which is a bridge between fine-tuning (giving many examples) and zero-shot prompting (giving only description). We apply few-shot SL prompting by carefully selecting examples from the training set so that the model is able to learn to distinguish between the positive and negative examples for each value. After some experimentation, we suffice with 3 positive examples, and 3 negative examples. The example sentences are selected from a dataset of sentences based on the words they have in common with the value-labeled sentences. Before matching sentences based on words, we remove the words that are common across all sentences (see Appendix B.2). For the negative examples, we select sentences that are a) randomly drawn from those sentences not annotated by the vocal value, b) annotated with a related value adjacent in the circle, and c) annotated with an opposed value. This way, we hope that we show the algorithm specialized information on what constitutes a value and what does not.

3. Approach

From the training set, we selected sentences to be used as examples in prompts as well as labeled data used for fine-tuning. We remove sentences with fewer than 15 characters are excluded from this selection as they are less likely to be informative about a human value, reducing the training set by 44123 sentences. when we also remove those labeled by 0.5, which might be less clear, our final training dataset is 42210. Second, we remove the stopwords (we augmented the nltk list with 124 words, see B.1), connector words (gensim), numbers (both alphabetically and numerically written), and tokens smaller than 2. We keep hyphened words and nouns, adjectives, and adverbs. On this subset we run a phrase model to identify frequently co-occurring words. From this final vocabulary of 24172 tokens, we identify the most frequent words occurring across all sentences (see B.2). Excluding these overall common words, we search per value for the most frequent words and match the negative and positive

examples based on these words.

To explore various approaches to zero-shot and few-shot prompting and compare with fine-tuning, we select a subset from the validation sample. For prompting, we selected max 600 sentences per value of which 300 were positive examples, and the other 300 were divided among 4 sets of negative examples (of which 2 were random negative examples, 1 was related negative example, and 1 was opposed negative example). If there were fewer than 300 positive examples, we selected all positive examples, and matched with an equal set of negative examples (divided across the random, opposed and related values). Since the negative examples could be labeled for values other than the vocal value, the total subset contained more than 300 positive examples for some values. In total we have .. sentences for the validation subset used for testing (see appendix A.3 and A.1. All of our models are tested on these subsamples from the original validation set.

To fine-tune the models, we used the training set to select sentences. We used the same approach as above but only for maximally 240 positive examples per value (for SL), or 20 positive examples (for ML) to reduce the computational resources needed. This way, we cap the dataset used for fine-tuning at 480 sentences. Again, we tested the models on the sub samples from the validation set. For fine-tuning Gemini, we convert the training data into a jsonl format and use the VertexAI API to initiate and run a fine-tuning job. When completed, the job returns evaluation metrics for the training data which includes the training loss, token accuracy at training step and number of predicted tokens at a training step [2]. These metrics can be visualised both using an API call as well as the Vertex AI Dashboard. For fine-tuning in OpenAI using the Davinci model, we used the 480 sentences and the labels with hyphens between (so self-direction-thought). This resulted in fewer random responses. But still there were responses such as: ' self-direction-direction-thought','-thoughtominityetal' or 'freedom-dominance-th'. After performing the fine-tuning job for both single and multi-label, we evaluate their performance on the validation set.

4. Results

We present our validation set results and discuss the influence of different choices on model performances: 1) open vs closed source 2) fine-tuning vs prompting using single and multi-label approaches. As can be seen in 2, our best performing model is the open source Llama3-70b-instruct with an overall f1-score of 0.70, 6 points higher than the best performing closed source models (gemini-1.0-pro SFT (SL) and GPT-4o few-shot (SL). This signifies that even open-access models can deliver state of the art performance in a task like human values detection that requires nuanced and contextual understanding of language. In comparing fine-tuning to prompting, we first analyse it on the single label. Here, the fine-tuning for single label for Gemini seems to relatively match the performance of the best performing prompting approach, which Llama3 being an exception. This highlights that creating a fine-tuned model for single labels and aggregating them for predicting all labels might give good results. However, prompting still seems to be the best performing approach for predicting single labels. In contrast, the multi-label approaches seem to be the worst performing for both the fine-tuning as well as prompting. For fine-tuning, this can be caused by the lack of sufficient training data for each value for the model to properly understand the nuances in them. For prompting, we think that this can be partly explained by the fact that these language models do not robustly make use of information presented in long input contexts[5].

We also see that compared to zero-shot, few-shot approaches lead to a slight gain in performance, indicating the usefulness of including positive and negative examples for each value in prompt design. For GPT, it seems that the more recent models are more effective with a higher F1 score for GPT4o compared to GPT3.5. We only tested this for zero-shot single label). Adding positive and negative examples to the prompt increases the performance. When adding on top of the examples, the context of the sentence (i.e. the three sentences in the text preceding the sentence that was labeled), the

[2]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning

Table 1
Achieved F_1-score of each trained model on the validation dataset for subtask 1. A ✓ indicates that the submission used the automatic translation to English.

		F_1-score																			
Validation Subset	**EN**	**All**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
GPT-3.5 zero-shot (ML)	✓	38	32	33	42	59	69	32	32	38	32	31	63	30	33	33	32	32	33	32	32
GPT-4o zero-shot (ML)	✓	48	38	38	44	54	64	52	46	36	59	49	55	36	35	38	49	35	56	79	37
GPT-3.5 Supervised Fine Tuning (SFT) (ML)	✓	42	41	38	39	48	49	47	41	38	48	46	46	49	35	28	38	39	46	53	40
GPT-3.5 zero-shot (SL)	✓	57	47	58	59	48	61	61	50	40	55	59	70	57	62	56	39	53	47	69	75
GPT-3.5 few-shot (SL)	✓	63	41	53	71	72	62	64	59	59	58	57	76	67	59	60	55	61	66	78	75
GPT-4o few-shot (SL)	✓	64	45	62	67	67	60	71	59	57	60	56	78	73	67	61	58	61	61	81	74
GPT-3.5 context zero-shot (SL)	✓	58	48	57	64	46	62	66	35	29	55	60	71	70	64	56	39	57	71	73	72
GPT-3.5 context few-shot (SL)	✓	62	45	52	72	76	62	43	54	54	60	58	74	68	61	57	53	61	78	73	73
gemini-1.0-pro Supervised Fine Tuning (SFT) (SL)	✓	64	57	51	12	77	69	61	68	73	68	68	84	67	52	66	67	54	65	84	70
gemini-1.0-pro Supervised Fine Tuning (SFT) (ML)	✓	21	15	13	05	35	32	23	24	05	35	14	38	33	08	22	22	10	17	24	39
llama3-70b-instruct zero-shot (SL)	✓	70	49	67	67	61	75	76	72	75	65	69	85	73	70	58	75	75	76	91	78
llama3-70b-instruct zero-shot (ML)	✓	26	12	24	17	24	37	23	13	14	25	19	50	38	00	36	25	17	24	52	48

Table 2
Achieved F_1-score of each submission on the test dataset for subtask 1. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

		F_1-score																			
Submission (test set)	**EN**	**All**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
GPT3.5 few shot (SL)	✓	23	08	12	13	20	27	18	27	12	15	32	31	33	07	03	19	19	35	50	11
GPT-4o informed zero-shot (ML)	✓	25	15	10	10	18	25	18	09	24	21	30	46	33	09	15	26	15	41	55	20
valueeval24-bert-baseline-en	✓	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02

performance deteriorates slightly. Note that Llama3 has only been used as a zero-shot due to time constraints.

If we zoom in on the values, we see that Llama3 performs well across all values, while other generative LLMs perform worse. For instance, GPT3.5 has a much lower F1 score for values across the board, except for tradition and universalism-nature. However, Llama3 also outperforms GPT3.5 here with a very impressive F1 score of 85, respectively 91. Some values are notoriously difficult to predict, such as self-direction thought. Even Llama3 was unable to achieve a higher F1 score than .49. Surprisingly, fine-tuning with GEMINI proved to be very successful and obtained an F1 of .57 for this value. We can hypothesize that for some values such as self-direction thought, fine-tuning leads to a better result as the model better learns the nuances in the value through sufficient training examples whereas for

some an effective prompt design seems to give the best results. This also highlights the importance of combining these two approaches to achieve an overall better result.

Table 2 shows that our ML model does slightly better above the baseline model on the test set C.4. As the SL predictions took a lot of time (for GPT3.5, it took about 3-4 hours per value for the single-label model, and Llama3 it took about 7-8 hours per value), we only include the results of our best performing GPT3.5 model: the few-shot single label prompt. Contrary to our expectations the prompt did not do much better than our previous multilabel prompt using GPT4o. The value self-direction-thought was not completely finished which could explain the low F1 score here, but even without this one value, we don't see an improvement of SL-GPT3.5 over ML-GPT4o. Despite these results, the single label few shot outperformed this model in our validation subsets. The most likely reason could be that our validation subsets have very different distributions of values and words than the test set.

5. Discussion

In our paper, we looked at the capabilites of open and closed source models as well as the influence of fine-tuning and prompting with different single and multi-label approaches. Based on our validation set, prompting gives the best results when trying to predict human values using text, hence signifying the importance of an effective prompt design, with few-shot approaches showing slight gain in performances compared to zero shot approaches. Given the time-limitedness, further research can be focused on looking at the text similarities and differences between test and our validation subset. We can also estimate SL prompting approaches with GPT-4o as well as run LLama3 SL on the entire test set given enough computational capacity and finally compare SL SFT for openai with ML SFT and note the gain in performance or lack thereof.

References

[1] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[2] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, L. Scheunemann, Values and Identities - a Policymaker's Guide, Technical Report KJ-NA-30800-EN-N, European Commission's Joint Research Centre, Luxembourg, 2021. doi:10.2760/349527.

[5] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173.

A. Appendix: Prompts

A.1. Multi-Label

Assess which value relates to text. Follow description below in format VALUE: description. SELF-DIRECTION–THOUGHT: Freedom to cultivate one's own ideas and abilities
SELF-DIRECTION–ACTION: Freedom to determine one's own actions
STIMULATION: Excitement, novelty, and change
HEDONISM: Pleasure and sensuous gratification
ACHIEVEMENT: Success according to social standards
POWER–DOMINANCE: Power through exercising control over people
POWER–RESOURCES: Power through control of material and social resources
FACE: Security and power through maintaining one's public image and avoiding humiliation
SECURITY–PERSONAL: Safety in one's immediate environment
SECURITY–SOCIETAL: Safety and stability in the wider society
TRADITION: Maintaining and preserving cultural, family, or religious traditions
CONFORMITY–RULES: Compliance with rules, laws, and formal obligations
CONFORMITY–INTERPERSONAL: Avoidance of upsetting or harming other people
HUMILITY: Recognizing one's insignificance in the larger scheme of things
BENEVOLENCE–DEPENDABILITY: Being a reliable and trustworthy member of the in-group
BENEVOLENCE–CARING:Devotion to the welfare of in-group members
UNIVERSALISM–CONCERN: Commitment to equality, justice, and protection for all people
UNIVERSALISM–NATURE: Preservation of the natural environment
UNIVERSALISM–TOLERANCE: Acceptance and understanding of those who are different from oneself
Return VALUE. If text reflects no value, return NEUTRAL.

A.2. Single Label

Assess if the text relates to UNIVERSALISM–TOLERANCE: Acceptance and understanding of those who are different from oneself. Return 1 if it does, 0 if not.

A.3. Few shot

Assess if the text relates to SELF–DIRECTION–THOUGHT: Freedom to cultivate one's own ideas and abilities. Return 1 if it does, 0 if not. Here are some examples:

Haimov explains that it is important for the child to be involved in the process, so that he understands that even if he is headed for a certain institution, sometimes it is not the right step for him. : 1

President Donald Trump says the US Supreme Court has not properly addressed mass election fraud. : 1

Stabilize eco-bonuses and support efficient district heating for upgrading and decarbonization of public and private heritage buildings.: 0

People who wanted to obtain information on the issue accelerated their research.: 0

This series of experiments is the first step in a multi-year experiment program of the Ministry of Defense (the directorate for research and development of the military and technological infrastructure - AB) and the defense industries to develop a land and air laser system to deal with threats at different ranges at high powers.: 0

B. Appendix: Words

B.1. Stopwords

B.2. Common words

Self-direction: action	Self-direction: thought	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	All texts
people	children	right	development	good	last	Israel	market	political	water
new	school	different	order	fun	number	power	economic	state	safe
time	President	Trump	public	really	company	Russian	gas	apology	treatment
country	education	important	change	moment	companies	Russia	energy	party	way
years	already	political	technology	children	market	US	economy	media	security y
year	party	issue	education	love	already	EU	Russia	Russian	body
government	right	several	energy	food	system	President	euros	public	Mineral
first	freedom	idea	innovation	speech	way	military	production	campaign	important
European	free	things	young	still	well	Ukraine	money	Prime	beneficial l
Minister	state	researchers	business	Many	Israel	police	EU	image	good
many	information	decision	opportunities	home	good	sanctions	investment	part	risk
countries	action	President	work	true	work	control	way	fact	place
even	group	name	research	little	able	political	well	never	school
world	well	way	possible	day	best	pressure	sector	role	home
also	EU	research	future	happy	percent	Turkey	companies	Ministry	health

Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance
security	Israel	law	EU	everyone	children	together	social	energy	different
measures	cultural	rules	relations	much	support	support	children	climate	differences
order	family	Court	Greece	day	family	cooperation	education	green	racism
social	children	EU	meeting	important	education	Israel	women	emissions	diversity
health	state	decision	talks	humble	companies	President	rights	renewable	society
Israel	national	court	cooperation	American	families	well	refugees	use	today
police	history	case	states	night	better	NATO	system	change	meeting
system	God	legal	Turkish	situation	child	relations	citizens	areas	together
energy	part	order	Turkey	team	important	Turkey	support	environmental	course
protection	Jewish	right	agreement	thankful	citizens	solidarity	right	global	political
crisis	Allah	public	never	grateful	students	good	work	sustainable	differently
economic	education	work	interest	season	health	way	school	gas	tolerance
public	heritage	state	chance	anything	workers	members	opportunities	production	issue
state	faith	Ministry	everyone	whole	free	EU	young	development	peace
necessary	language	already	together	lot	social	Prime	public	carbon	discrimination

C. Appendix: OpenAI

C.1. Zero-shot ML

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.329032	0.500000	0.245192	0.500000	208
Self-direction: action	0.333663	0.501946	0.746032	0.501946	505
Stimulation	0.421372	0.541637	0.700067	0.541637	509
Hedonism	0.592675	0.644231	0.792135	0.644231	104
Achievement	0.692830	0.704829	0.729844	0.704829	583
Power: dominance	0.325175	0.500000	0.240933	0.500000	579
Power: resources	0.325581	0.500000	0.241379	0.500000	580
Face	0.389513	0.523870	0.668259	0.523870	163
Security: personal	0.326923	0.500000	0.242857	0.500000	105
Security: societal	0.313953	0.500000	0.228814	0.500000	590
Tradition	0.635952	0.677326	0.798913	0.677326	337
Conformity: rules	0.306147	0.500000	0.220613	0.500000	587
Conformity: interpersonal	0.331210	0.500000	0.247619	0.500000	105
Humility	0.333333	0.500000	0.250000	0.500000	40
Benevolence: caring	0.327212	0.500000	0.243176	0.500000	403
Benevolence: dependability	0.327623	0.500000	0.243631	0.500000	314
Universalism: concern	0.330317	0.500000	0.246622	0.500000	592
Universalism: nature	0.329567	0.500000	0.245787	0.500000	356
Universalism: tolerance	0.327869	0.500000	0.243902	0.500000	82
Mean	0.384208	0.531255	0.398725	0.531255	354.842105

C.2. Zero-shot ML (GPT4o)

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.389098	0.528302	0.752475	0.528302	208
Self-direction: action	0.382206	0.523276	0.715813	0.523276	505
Stimulation	0.444883	0.552972	0.701315	0.552972	509
Hedonism	0.548611	0.615385	0.782609	0.615385	104
Achievement	0.641937	0.665530	0.710124	0.665530	583
Power: dominance	0.523553	0.572276	0.612637	0.572276	579
Power: resources	0.458125	0.523571	0.542262	0.523571	580
Face	0.365027	0.511822	0.623428	0.511822	163
Security: personal	0.590541	0.637800	0.732252	0.637800	105
Security: societal	0.499145	0.575752	0.645444	0.575752	590
Tradition	0.550340	0.616032	0.755430	0.616032	337
Conformity: rules	0.359992	0.513519	0.575959	0.513519	587
Conformity: interpersonal	0.351852	0.509434	0.750000	0.509434	105
Humility	0.386602	0.525000	0.756410	0.525000	40
Benevolence: caring	0.493737	0.576481	0.696442	0.576481	403
Benevolence: dependability	0.354696	0.512422	0.746774	0.512422	314
Universalism: concern	0.561806	0.619680	0.740923	0.619680	592
Universalism: nature	0.791186	0.793496	0.800323	0.793496	356
Universalism: tolerance	0.378788	0.523810	0.750000	0.523810	82
Mean	0.477480	0.573503	0.704769	0.573503	354.842105

C.3. Supervised Finetuning ML

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.412444	0.472346	0.454635	0.472346	208
Self-direction: action	0.375717	0.511179	0.573454	0.511179	505
Stimulation	0.391322	0.484725	0.463307	0.484725	509
Hedonism	0.487179	0.519231	0.525641	0.519231	104
Achievement	0.493900	0.528693	0.538008	0.528693	583
Power: dominance	0.474970	0.503548	0.504364	0.503548	579
Power: resources	0.416076	0.508095	0.519784	0.508095	580
Face	0.383482	0.485919	0.460247	0.485919	163
Security: personal	0.337805	0.489651	0.406863	0.489651	105
Security: societal	0.484343	0.513368	0.515700	0.513368	590
Tradition	0.462425	0.467741	0.466663	0.467741	337
Conformity: rules	0.489845	0.507557	0.507981	0.507557	587
Conformity: interpersonal	0.351852	0.509434	0.750000	0.509434	105
Humility	0.285714	0.400000	0.222222	0.400000	40
Benevolence: caring	0.377595	0.471914	0.433128	0.471914	403
Benevolence: dependability	0.396299	0.500832	0.502480	0.500832	314
Universalism: concern	0.469198	0.496210	0.495305	0.496210	592
Universalism: nature	0.533835	0.537948	0.539046	0.537948	356
Universalism: tolerance	0.403372	0.510119	0.532381	0.510119	82
Mean	0.422493	0.495711	0.495327	0.495711	354.842105

C.4. Zero-shot SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.475630	0.545320	0.591760	0.545320	208
Self-direction: action	0.581450	0.612597	0.654453	0.612597	505
Stimulation	0.586714	0.623173	0.677015	0.623173	509
Hedonism	0.484685	0.576923	0.770833	0.576923	104
Achievement	0.616815	0.649464	0.709813	0.649464	583
Power: dominance	0.614849	0.615789	0.615852	0.615789	579
Power: resources	0.506604	0.527143	0.531476	0.527143	580
Face	0.401412	0.529895	0.681777	0.529895	163
Security: personal	0.550000	0.610022	0.712781	0.610022	105
Security: societal	0.594653	0.601389	0.602546	0.601389	590
Tradition	0.709248	0.729035	0.794491	0.729035	337
Conformity: rules	0.573897	0.584789	0.585867	0.584789	587
Conformity: interpersonal	0.628830	0.659833	0.733563	0.659833	105
Humility	0.563636	0.625000	0.785714	0.625000	40
Benevolence: caring	0.399468	0.627181	0.446013	0.418121	403
Benevolence: dependability	0.534250	0.587951	0.652088	0.587951	314
Universalism: concern	0.473877	0.714247	0.482888	0.476164	592
Universalism: nature	0.690581	0.708903	0.761707	0.708903	356
Universalism: tolerance	0.753754	0.758929	0.771875	0.758929	82
Mean	0.565282	0.625662	0.661185	0.602128	354.842105

C.5. Few-shot SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.417733	0.492323	0.484873	0.492323	208
Self-direction: action	0.534062	0.535459	0.535731	0.535459	505
Stimulation	0.715059	0.715224	0.719304	0.715224	509
Hedonism	0.720430	0.730769	0.770833	0.730769	104
Achievement	0.628770	0.647462	0.675343	0.647462	583
Power: dominance	0.645333	0.655932	0.687710	0.655932	579
Power: resources	0.559652	0.562381	0.563124	0.562381	580
Face	0.590816	0.604367	0.617737	0.604367	163
Security: personal	0.584018	0.617102	0.661250	0.617102	105
Security: societal	0.573713	0.596991	0.609961	0.596991	590
Tradition	0.764715	0.771529	0.793786	0.771529	337
Conformity: rules	0.675029	0.678283	0.675832	0.678283	587
Conformity: interpersonal	0.595561	0.599057	0.603175	0.599057	105
Humility	0.605003	0.625000	0.656740	0.625000	40
Benevolence: caring	0.556410	0.581805	0.601835	0.581805	403
Benevolence: dependability	0.612154	0.621321	0.630588	0.621321	314
Universalism: concern	0.658162	0.678858	0.728490	0.678858	592
Universalism: nature	0.788282	0.790734	0.798026	0.790734	356
Universalism: tolerance	0.754785	0.758333	0.765931	0.758333	82
Mean	0.630510	0.645417	0.662119	0.645417	354.842105

C.6. Few-shot SL (GPT4o)

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.448520	0.546430	0.646784	0.546430	208
Self-direction: action	0.620388	0.650339	0.709828	0.650339	505
Stimulation	0.674399	0.689271	0.718614	0.689271	509
Hedonism	0.672856	0.701923	0.813253	0.701923	104
Achievement	0.600360	0.639865	0.712416	0.639865	583
Power: dominance	0.716752	0.717760	0.717817	0.717760	579
Power: resources	0.591579	0.601071	0.607982	0.601071	580
Face	0.578354	0.625602	0.714617	0.625602	163
Security: personal	0.600137	0.636710	0.699629	0.636710	105
Security: societal	0.561422	0.571412	0.573649	0.571412	590
Tradition	0.788847	0.795278	0.819482	0.795278	337
Conformity: rules	0.730796	0.739653	0.738435	0.739653	587
Conformity: interpersonal	0.670071	0.697932	0.789236	0.697932	105
Humility	0.615385	0.650000	0.734375	0.650000	40
Benevolence: caring	0.583410	0.628944	0.710506	0.628944	403
Benevolence: dependability	0.610371	0.644014	0.705616	0.644014	314
Universalism: concern	0.619124	0.656119	0.744511	0.656119	592
Universalism: nature	0.819764	0.821310	0.825753	0.821310	356
Universalism: tolerance	0.746444	0.761310	0.815374	0.761310	82
Mean	0.644683	0.672365	0.726204	0.672365	354.842105

C.7. Context zero-shot SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.479109	0.550222	0.604604	0.550222	208
Self-direction: action	0.572290	0.600430	0.631968	0.600430	505
Stimulation	0.646528	0.678114	0.752841	0.678114	509
Hedonism	0.467637	0.567308	0.768041	0.567308	104
Achievement	0.625524	0.656131	0.714782	0.656131	583
Power: dominance	0.662955	0.665090	0.666266	0.665090	579
Power: resources	0.354369	0.542738	0.364602	0.361825	580
Face	0.296793	0.547741	0.505519	0.365161	163
Security: personal	0.550000	0.610022	0.712781	0.610022	105
Security: societal	0.601666	0.605324	0.605051	0.605324	590
Tradition	0.706533	0.726004	0.788529	0.726004	337
Conformity: rules	0.705278	0.714757	0.714418	0.714757	587
Conformity: interpersonal	0.644893	0.669086	0.726874	0.669086	105
Humility	0.563636	0.625000	0.785714	0.625000	40
Benevolence: caring	0.393815	0.625444	0.453447	0.416962	403
Benevolence: dependability	0.574656	0.612309	0.663650	0.612309	314
Universalism: concern	0.718488	0.722763	0.733290	0.722763	592
Universalism: nature	0.732942	0.744815	0.785144	0.744815	356
Universalism: tolerance	0.727657	0.735119	0.753205	0.735119	82
Mean	0.580251	0.642022	0.670038	0.611918	354.842105

C.8. Context few-shot SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.454714	0.515908	0.527778	0.515908	208
Self-direction: action	0.517328	0.519965	0.520295	0.519965	505
Stimulation	0.722974	0.723089	0.727319	0.723089	509
Hedonism	0.760369	0.769231	0.815972	0.769231	104
Achievement	0.624210	0.644229	0.673806	0.644229	583
Power: dominance	0.433450	0.659516	0.461945	0.439677	579
Power: resources	0.548872	0.552381	0.553282	0.552381	580
Face	0.540455	0.555271	0.562351	0.555271	163
Security: personal	0.596154	0.626362	0.669426	0.626362	105
Security: societal	0.585175	0.606366	0.618702	0.606366	590
Tradition	0.746617	0.753717	0.774514	0.753717	337
Conformity: rules	0.683777	0.687529	0.684939	0.687529	587
Conformity: interpersonal	0.608245	0.609035	0.610235	0.609035	105
Humility	0.573333	0.600000	0.633333	0.600000	40
Benevolence: caring	0.535228	0.570406	0.596237	0.570406	403
Benevolence: dependability	0.618975	0.623290	0.626965	0.623290	314
Universalism: concern	0.649815	0.675845	0.741064	0.675845	592
Universalism: nature	0.785372	0.787972	0.795740	0.787972	356
Universalism: tolerance	0.730263	0.733929	0.740809	0.733929	82
Mean	0.616596	0.642844	0.649195	0.631274	354.842105

D. Appendix: GEMINI

D.1. Supervised Fine Tuning Gemini SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.567442	0.557143	0.559633	0.575472	210
Self-direction: action	0.512097	0.528265	0.531381	0.494163	513
Stimulation	0.125874	0.521073	0.750000	0.068702	522
Hedonism	0.769231	0.798077	0.897436	0.673077	104
Achievement	0.690722	0.700000	0.712766	0.670000	600
Power: dominance	0.617594	0.641414	0.669261	0.573333	594
Power: resources	0.677054	0.620000	0.588670	0.796667	600
Face	0.608187	0.588957	0.590909	0.626506	163
Security: personal	0.731707	0.688679	0.652174	0.833333	106
Security: societal	0.680982	0.653333	0.630682	0.740000	600
Tradition	0.837349	0.843023	0.868750	0.808140	344
Conformity: rules	0.666667	0.673333	0.680556	0.653333	600
Conformity: interpersonal	0.524272	0.533333	0.540000	0.509434	105
Humility	0.666667	0.725000	0.846154	0.550000	40
Benevolence: caring	0.671264	0.652068	0.640351	0.705314	411
Benevolence: dependability	0.544170	0.598131	0.631148	0.478261	321
Universalism: concern	0.656881	0.688333	0.730612	0.596667	600
Universalism: nature	0.846939	0.833795	0.786730	0.917127	361
Universalism: tolerance	0.705882	0.695122	0.697674	0.714286	82
Mean	0.640672	0.654745	0.678099	0.649180	370.421053

D.2. Supervised Fine Tuning Gemini ML

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.146341	0.938596	0.214286	0.111111	27
Self-direction: action	0.125000	0.901754	0.142857	0.111111	36
Stimulation	0.051282	0.935088	0.200000	0.029412	34
Hedonism	0.347826	0.947368	0.571429	0.250000	32
Achievement	0.325000	0.905263	0.333333	0.317073	41
Power: dominance	0.231579	0.871930	0.215686	0.250000	44
Power: resources	0.240964	0.889474	0.208333	0.285714	35
Face	0.057143	0.942105	0.111111	0.038462	26
Security: personal	0.354839	0.929825	0.282051	0.478261	23
Security: societal	0.141176	0.871930	0.125000	0.162162	37
Tradition	0.384615	0.943860	0.322581	0.476190	21
Conformity: rules	0.337662	0.910526	0.282609	0.419355	31
Conformity: interpersonal	0.088889	0.928070	0.068966	0.125000	16
Humility	0.222222	0.963158	0.333333	0.166667	18
Benevolence: caring	0.226415	0.928070	0.187500	0.285714	21
Benevolence: dependability	0.105263	0.940351	0.111111	0.100000	20
Universalism: concern	0.170213	0.931579	0.166667	0.173913	23
Universalism: nature	0.242424	0.956140	0.250000	0.235294	17
Universalism: tolerance	0.392157	0.945614	0.384615	0.400000	25
Mean	0.211047	0.924172	0.229270	0.223080	27.889

E. Appendix: LLAMA3

E.1. Zero Shot SL

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.49	0.62	0.76	0.36	210
Self-direction: action	0.67	0.70	0.75	0.61	513
Stimulation	0.67	0.73	0.84	0.56	522
Hedonism	0.61	0.72	1.00	0.44	104
Achievement	0.75	0.72	0.69	0.83	600
Power: dominance	0.76	0.73	0.71	0.82	579
Power: resources	0.72	0.61	0.57	0.96	580
Face	0.75	0.76	0.80	0.71	163
Security: personal	0.65	0.69	0.77	0.56	106
Security: societal	0.69	0.57	0.54	0.96	590
Tradition	0.85	0.85	0.88	0.82	337
Conformity: rules	0.73	0.64	0.59	0.94	587
Conformity: interpersonal	0.70	0.70	0.71	0.70	105
Humility	0.58	0.68	0.82	0.45	40
Benevolence: caring	0.75	0.72	0.71	0.79	403
Benevolence: dependability	0.75	0.73	0.71	0.80	314
Universalism: concern	0.76	0.73	0.69	0.86	592
Universalism: nature	0.91	0.90	0.85	0.97	356
Universalism: tolerance	0.78	0.78	0.82	0.74	82
Mean	0.705	0.716	0.748	0.709	384.68

E.2. Zero Shot ML

Value	F1	Accuracy	Precision	Recall	N
Self-direction: thought	0.120000	0.920145	0.125000	0.115385	26
Self-direction: action	0.240000	0.931034	0.428571	0.166667	36
Stimulation	0.173913	0.931034	0.307692	0.121212	33
Hedonism	0.242424	0.954628	0.800000	0.142857	28
Achievement	0.368421	0.912886	0.388889	0.350000	40
Power: dominance	0.238806	0.907441	0.285714	0.205128	39
Power: resources	0.133333	0.929220	0.300000	0.085714	35
Face	0.137931	0.954628	0.400000	0.083333	24
Security: personal	0.247619	0.856624	0.158537	0.565217	23
Security: societal	0.193548	0.773140	0.127119	0.405405	37
Tradition	0.500000	0.952813	0.419355	0.619048	21
Conformity: rules	0.380952	0.952813	0.666667	0.266667	30
Conformity: interpersonal	0.000000	0.967332	0.000000	0.000000	15
Humility	0.363636	0.974592	0.666667	0.250000	16
Benevolence: caring	0.254545	0.925590	0.200000	0.350000	20
Benevolence: dependability	0.166667	0.963702	0.500000	0.100000	20
Universalism: concern	0.244444	0.876588	0.164179	0.478261	23
Universalism: nature	0.520000	0.956443	0.393939	0.764706	17
Universalism: tolerance	0.488889	0.958258	0.550000	0.440000	25
Mean	0.263954	0.926258	0.362228	0.289979	26.736842