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

Notebook for the Touché Lab at CLEF 2024

Ranjan Mishra[1] & Meike Morren[2]

[1]Tinbergen Institute, Netherlands
[2]Vrije Universiteit Amsterdam, Netherlands

Generative AI for Human Value Detection

- Higher-order constructs like human values are likely to be picked up by transformer models [1].
- Generative AI (GenAI) and Large Language Models (LLMs) established as state of the art in NLP
- Two primary adaptation approaches:
  - **Supervised Fine-Tuning (SFT)**
  - **Prompt Engineering**
    - Zero-Shot Multi-Label
    - Zero-Shot Single Label
    - Few-Shot
- Models:
  - **Closed Source:** GPT3.5, GPT-4o, gemini-1.0-pro
  - **Open Source:** llama3-70b-instruct
- Comparison of these approaches and their influence on predicting human values (Subtask 1)

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

Prompt Engineering: Zero-shot Single label (SL)

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.

Prompt Engineering: Zero-shot Multi-label (ML)

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.

F_1-Score Results for Subtask 1 - Test Dataset

Table 1: Achieved F_1-score on the test dataset for subtask 1.

		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
llama3-70b-instruct zero-shot (SL)	✓	18	09	17	16	17	24	21	19	12	16	21	24	23	06	03	16	13	29	37	13
valueeval24-bert-baseline-en	✓	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02

Using the validation set

Data Preparation

- Shorter (<15 characters) and ambiguous (labeled 0.5) excluded for being less informative about human values.
- Final training set of 42,210 sentences
- Additional preprocessing steps:
  - Removed stopwords, connector words, numbers (both written and numeric), and tokens smaller than 2.
  - Kept hyphenated words, nouns, adjectives, and adverbs.
  - Ran Phrase model to identify frequently co-occurring words
- Identified the most frequent words occurring across all sentences.
- Per value, frequent words were used to match positive and negative examples.

Most Frequent Occuring Words

Table 2: Most common words across different values in validation subsets

Self-direction: action	Stimulation	Hedonism	Achievement	All texts
people	right	development	good	water
new	different	order	fun	safe
time	Trump	public	really	treatment
country	political	technology	moment	way
years	issue	education	children	security
year	freedom	energy	speech	body
government	idea	innovation	still	important
first	things	young	home	beneficial
European	researchers	business	Many	good
Minister	decision	opportunities	home	risk
many	President	work	true	place
countries	name	research	little	school
even	way	possible	day	home
world	research	future	happy	health
also	EU	opportunities	speech	mineral

Selection for Prompting (Zero-shot and Few-shot)

- Subset selected from the validation sample for testing prompting approaches.
- For each value, we selected a maximum of 600 sentences:
  - 300 positive examples.
  - 300 divided among 4 sets of negative examples (random, related, and opposed).
- If fewer than 300 positive examples available, we used all positive examples with matching negative examples.

Prompt Engineering: Few-shot Single label (SL)

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

Supervised Fine-Tuning (SFT)

- Fine-tuning dataset creation mirrored the sentence selection process for prompting.
- Max of 240 positive examples per value for single-label (SL) fine-tuning.
- Max of 20 positive examples per value for multi-label (ML) fine-tuning.
- Total dataset for fine-tuning capped at 480 sentences to optimize computational resources.

Fine-Tuning Process

- Gemini Fine-Tuning
  - Training data converted to JSONL format for Gemini.
  - Used the Vertex AI API to run a fine-tuning job.
  - Job provided evaluation metrics:
    - Training loss, token accuracy at training step, and predicted tokens.
  - Metrics visualized through both API and Vertex AI Dashboard [2].
- OpenAI DaVinci Fine-Tuning
  - Training set of 480 sentences used for fine-tuning Davinci.
  - Labels structured with hyphens (e.g., self-direction-thought).

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

Results Validation Dataset

F_1-Score Results for Subtask 1 - Validation Dataset

Table 3: Achieved F_1-score on the validation dataset for subtask 1.

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
		F_1-score																			
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

Validation Set Results: Model Performance

- Few-shot vs Zero-shot Prompting

Table 4: Comparison of F_1-scores for GPT-3.5 models using zero-shot and few-shot SL prompting

Model	**Zero-shot F_1-score**	**Few-shot F_1-score**
GPT-3.5 (SL)	57	63
GPT-3.5 with Context (SL)	58	62

- Multi-label approaches worst performing across the board

Table 5: Comparison of F_1-scores for multi-label (ML) and single-label (SL) approaches.

Model	**ML F_1-score**	**SL F_1-score**
GPT-3.5 zero-shot	38	57
gemini-1.0-pro Supervised Fine Tuning (SFT)	21	64
llama3-70b-instruct zero-shot	26	70

Validation Set Results: Model Performance

- Some values difficult to predict than others

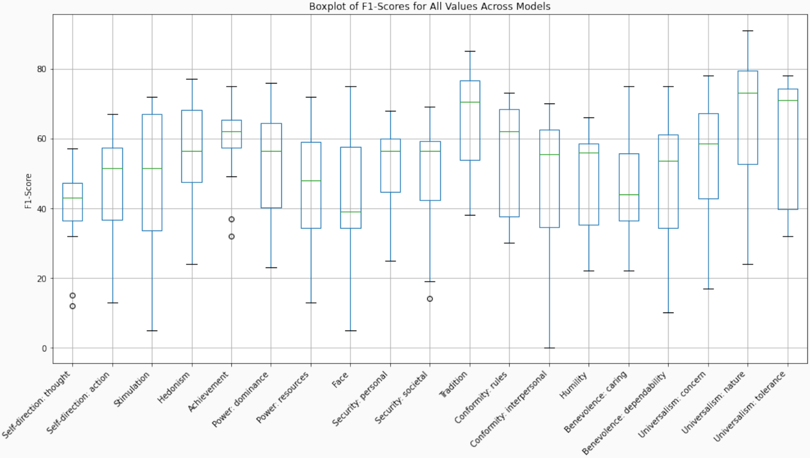

Figure 1: Boxplot of F1-Scores for All Values Across Models

Discussion

Points for discussion

We've learned that:

- Fine Tuning does worse than prompting
- Single-label models outperforms Multi-label models
- In Validation Subset: LLama3 (70B) single label performs best
- In Testset: GPT4o multi-label performs best (GPT4o single label was too expensive)

We wonder why:

- Results validation subset differ in testset
- Even though validation and test set are very similar in terms of number of sentences, sentence length (characters), vocabulary, entropy.
- Validation subset is a bit different but not dramatically
- Multi-label GPT4o performs better in test set (compared to single label GPT4o) while in the validation set this is reversed.

Thank You for Your Attention!

Comparing datasets: Statistics at sentence level

	Test (N=14569)	Valid (14904)	Valid subset (N=4183)
mean chars	127.37	126.91	142.52
std	89.04	87.68	85.53
min	1	1	15
25%	67	66	82
50%	110	110	126
75%	166	167	183
max chars	2148	2188	856

Comparing datasets: Frequent Words

Test (N=19278)	Valid (N=16392)	Valid subset (N=9844)
also	also	also
people	people	new
new	new	people
time	government	EU
country	time	government
year	years	time
years	country	country
government	European	way
European	first	years
first	year	European
Minister	Minister	today
public	even	world
countries	countries	Minister
state	many	order
many	public	public
even	President	important
EU	EU	Bulgaria
last	way	work
President	well	social
percent	state	part
well	today	first
way	last	possible
system	already	day
order	system	energy
already	world	countries
day	percent	system
world	Turkey	measures
number	work	Israel
Europe	companies	development
children	important	crisis

Comparing datasets: Entropy at text level

	Test set (N=522)	Valid (N=522)	Valid subset (N=511)
mean	4.429639	4.429408	4.347069
std	0.094378	0.097833	0.110457
min	4.173949	4.116592	3.876894
25%	4.369347	4.370443	4.284216
50%	4.420451	4.423430	4.345551
75%	4.477135	4.473919	4.402887
max	4.904444	5.142850	4.986116