Evaluation of few-shot learning capabilities in polish language models

Tsimur Hadeliya   Dariusz Kajtoch

ML Research at Allegro, Poznań, Poland

Introduction

Recent research reports that pre-trained language models can effectively solve natural language problems using only few examples. This approach called few-shot learning(FSL) and gained much popularity in recent years. However, vast majority of research in few-shot learning conducted exclusively for English, while other languages remains unexplored. To address this gap for polish language, we conducted experiments using two main approaches in FSL: In-context learning (ICL) and Parameter Efficient Fine-Tuning (PEFT). To get relevant and reliable results we stick to classification tasks during the experiments and construct few-shot classification benchmark based on publicly available datasets.

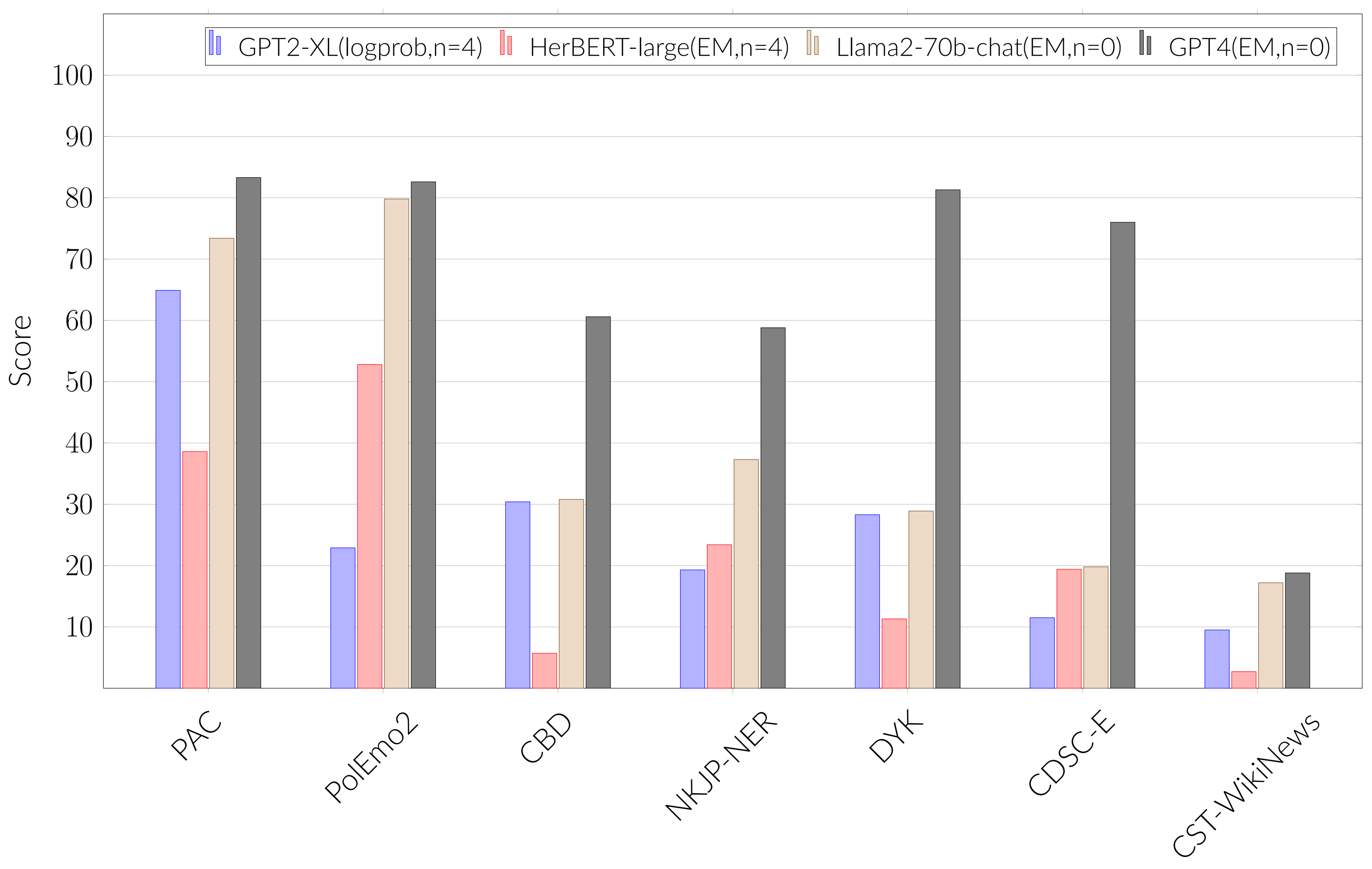
Benchmark

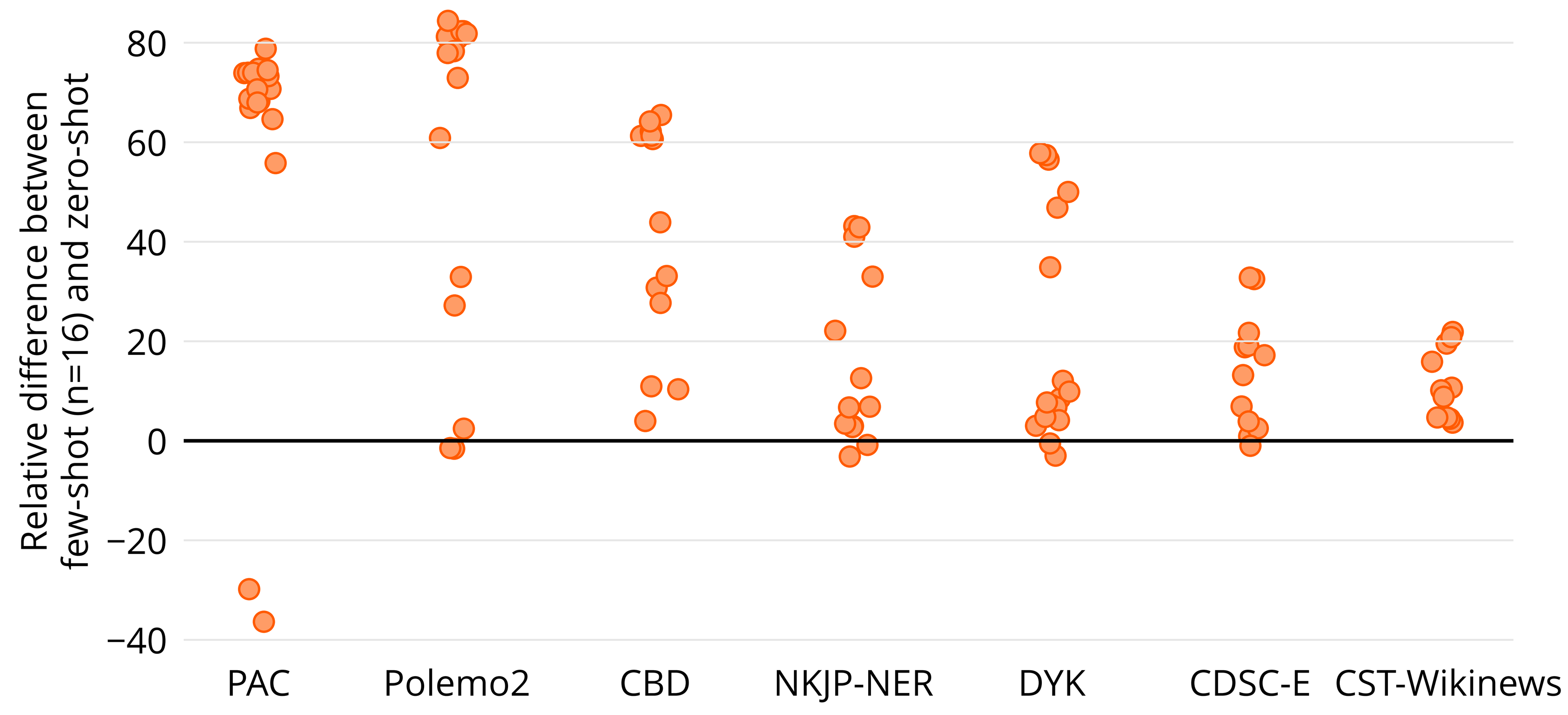
Name	Input	#Classes	Metrics	Avg. len	Domain
Lepiszcze					
PAC	text	2	F1-binary	185	legal texts
DYK	text pair	2	F1-binary	288	Wikipedia
CDSC-E	text pair	3	Accuracy	144	image captions
Polemo2	text	4	Accuracy	758	online reviews
KLEJ					
CBD	text	2	F1-binary	93	social media
NKJP-NER	text	6	Accuracy	85	national corpus
Other					
CST-Wikinews	text pair	12	Accuracy	232	Wikinews

Main findings

- **Few-shot approaches shows weaker performance compared to baseline models**
  Except GPT-4, all models and methods shows much weaker performance compared to full fine-tuning and even the simplest baselines.
- **PEFT outperforms ICL almost for all models**
  ICL methods is useful with LLM, but can't guarantee stable performance/
- **Polish LM from the box is not suitable to in-context learning methods**
  The main problems are: short context window and high sensitivity for input format.
- **Instruction tuning of LLM transfers to Polish language**
  Significant difference is observed between base and chat version of Llama-2-70b models.

In-context learning performance of selected models





Difference in performance between the few- shot setting (n = 16) and the zero-shot setting for the GPT-3.5. Results were collected for all our templates.

Model	AVG	PAC	Polemo2	CBD	NKJP-NER	DYK	CDSC-E	CST-Wikinews
Metric used		F1	Acc	F1	Acc	F1	Acc	Acc
Baseline								
Random Guessing	$26.6_{\pm 0.4}$	$57.5_{\pm 0.7}$	$24.8_{\pm 1.0}$	$20.9_{\pm 1.6}$	$16.8_{\pm 0.8}$	$25.5_{\pm 1.7}$	$33.6_{\pm 0.7}$	$6.9_{\pm 0.9}$
Most frequent	42.6	80.6	41.3	23.6	34.3	28.9	74.4	15.4
/Full FT/ HerBERT-large	$79.9_{\pm 0.6}$	$91.1_{\pm 0.0}$	$90.9_{\pm 0.0}$	$53.2_{\pm 3.2}$	$94.0_{\pm 0.0}$	$68.8_{\pm 2.1}$	$93.4_{\pm 0.0}$	$67.9_{\pm 1.0}$
PEFT								
/SF/ SBERT-large	47.1	$68.8_{\pm 6.5}$	$69.9_{\pm 10.7}$	$44.4_{\pm 5.4}$	$30.7_{\pm 6.5}$	$27.7_{\pm 2.9}$	$\underline{72.3}_{\pm 5.7}$	$16.2_{\pm 3.2}$
/SF/ HerBERT-large	44.1	$70.7_{\pm 11.9}$	$46.0_{\pm 11.2}$	$42.5_{\pm 10.9}$	$25.8_{\pm 11.6}$	$\underline{40.6}$	$67.2_{\pm 14.5}$	$15.7_{\pm 1.9}$
/SF/ RoBERTa-large	$\underline{47.6}$	$66.8_{\pm 16.1}$	$\underline{\mathbf{83.6}}_{\pm 2.7}$	$\underline{44.5}_{\pm 10.6}$	$\underline{35.5}_{\pm 4.4}$	26.1	$62.2_{\pm 10.5}$	$14.3_{\pm 2.5}$
/LP/ SBERT-large	43.4	$67.5_{\pm 7.0}$	$60.3_{\pm 4.6}$	$40.1_{\pm 4.3}$	$30.4_{\pm 5.9}$	$27.2_{\pm 1.8}$	$62.8_{\pm 5.8}$	$15.3_{\pm 3.0}$
/FT/ SBERT-large	31.2	$33.6_{\pm 31.3}$	$47.0_{\pm 5.8}$	$32.0_{\pm 30.7}$	$28.1_{\pm 7.7}$	$6.4_{\pm 10.3}$	$61.0_{\pm 17.6}$	$10.1_{\pm 4.3}$
/LP/ Ada	40.9	$\underline{72.9}_{\pm 6.3}$	$55.2_{\pm 5.0}$	$30.7_{\pm 4.2}$	$29.1_{\pm 2.5}$	$25.4_{\pm 4.1}$	$58.2_{\pm 11.7}$	$14.8_{\pm 3.9}$
/LP/ DaVinci	42.7	$67.6_{\pm 7.0}$	$58.9_{\pm 1.3}$	$36.9_{\pm 9.7}$	$30.1_{\pm 6.6}$	$29.9_{\pm 2.7}$	$60.1_{\pm 11.2}$	$15.5_{\pm 3.5}$
/LP/ Gecko	37.7	$61.8_{\pm 4.5}$	$42.9_{\pm 6.4}$	$23.7_{\pm 1.3}$	$24.8_{\pm 4.3}$	$25.0_{\pm 3.8}$	$68.8_{\pm 6.3}$	$\underline{17.0}_{\pm 2.8}$
In-context learning								
GPT-3.5 (EM) (n=0)	55.4	$82.2_{.014}$	$81.6_{.005}$	$50.0_{.046}$	$44.9_{.001}$	$53.1_{.0}$	$62.9_{.0}$	$13.3_{.003}$
GPT-3.5 (EM) (n=16)	59.5	$73.9_{\pm 3.6}$	$81.9_{\pm 2.1}$	$\underline{\mathbf{64.1}}_{\pm 1.9}$	$46.1_{\pm 2.9}$	$64.1_{\pm 1.8}$	$66.7_{\pm 7.6}$	$\underline{\mathbf{19.8}}_{\pm 2.7}$
GPT-4 (EM) (n=0)	$\underline{65.9}$	$83.3_{\pm .0}$	$\underline{82.6}_{\pm .009}$	$60.6_{\pm .028}$	$\underline{\mathbf{58.8}}_{\pm .004}$	$\underline{\mathbf{81.3}}_{\pm .002}$	$\underline{\mathbf{76.0}}_{\pm .0}$	$18.8_{\pm .003}$
Llama-2-70b-chat (EM) (n=0)	41.0	$73.4_{\pm .090}$	$79.8_{\pm .002}$	$30.8_{\pm .122}$	$37.3_{\pm .087}$	$28.9_{\pm .0}$	$19.8_{\pm .0}$	$17.2_{\pm .016}$
Llama-2-70b (EM) (n=0)	14.6	$41.8_{\pm .576}$	$11.3_{\pm .678}$	$0.3_{\pm .662}$	$21.4_{\pm .324}$	$19.2_{\pm .055}$	$6.6_{\pm .0}$	$1.6_{\pm .859}$
Bison-text (EM) (n=0)	52.2	$80.2_{\pm .006}$	$80.7_{\pm .009}$	$42.6_{\pm .077}$	$47.5_{\pm .027}$	$61.6_{\pm .016}$	$35.0_{\pm .001}$	$17.7_{\pm .003}$
Bison-text (EM) (n=16)	-	$\underline{\mathbf{83.7}}_{\pm 1.4}$	$81.8_{\pm 2.8}$	-	$45.7_{\pm 3.6}$	$76.6_{\pm 0.7}$	$66.4_{\pm 7.3}$	$19.5_{\pm 2.2}$
Krakowiak-7b (EM) (n=0)	20.5	$38.0_{\pm .624}$	$28.4_{\pm .056}$	$0.5_{\pm .687}$	$23.9_{\pm .002}$	$24.4_{\pm .03}$	$19.0_{\pm .002}$	$9.1_{\pm .109}$
GPT-2-xl (EM) (n=0)	15.0	$23.8_{\pm 33.4}$	$20.0_{\pm 27.4}$	$16.9_{\pm 29.3}$	$21.3_{\pm 9.6}$	$14.9_{\pm 33.2}$	$8.1_{\pm 14.4}$	$0.0_{\pm 0.0}$
GPT-2-xl (EM) (n=4)	28.9	$65.1_{\pm 10.7}$	$32.5_{\pm 11.0}$	$29.3_{\pm 10.9}$	$16.1_{\pm 3.3}$	$35.6_{\pm 28.6}$	$12.1_{\pm 7.6}$	$11.6_{\pm 7.6}$
GPT-2-xl (EM) (n=16)	-	$68.4_{\pm 4.9}$	$-_{-}$	$19.9_{\pm 8.7}$	$22.1_{\pm 6.6}$	$23.1_{\pm 22.8}$	$10.2_{\pm 5.0}$	$11.5_{\pm 6.6}$
GPT-2-xl (logprob) (n=0)	20.2	$45.6_{\pm 27.0}$	$20.3_{\pm 5.8}$	$3.6_{\pm 4.7}$	$24.9_{\pm 1.6}$	$29.0_{\pm 18.2}$	$11.9_{\pm 5.1}$	$6.2_{\pm 5.3}$
GPT-2-xl (logprob) (n=4)	26.7	$64.9_{\pm 4.9}$	$22.9_{\pm 4.3}$	$30.4_{\pm 13.3}$	$19.3_{\pm 10.0}$	$28.3_{\pm 3.0}$	$11.5_{\pm 6.8}$	$9.5_{\pm 3.9}$
HerBERT-large (iter) (n=0)	11.2	$20.0_{44.7}$	$40.0_{54.8}$	$0.0_{0.0}$	$5.4_{12.1}$	$0.0_{0.0}$	$13.2_{12.2}$	$0.0_{0.0}$
HerBERT-large (iter) (n=4)	22.0	$38.6_{36.1}$	$52.8_{7.5}$	$5.7_{11.2}$	$23.4_{2.6}$	$11.3_{8.9}$	$19.4_{1.0}$	$2.7_{1.1}$

Performance on test data

Underline numbers - best across method(PEFT or ICL); **Bold numbers** - best across two methods (PEFT and ICL)

Methodology

We *F1-binary* and *Accuracy* metrics to follow approach used in KLEJ benchmark and have comparable results with models tested on this datasets.
To obtain reliable results, we conduct 5 experiments with different seeds and calculate mean with standard deviation (reported under-line).

Baseline

- **Random guessing** - Sample label from training data distribution.
- **Most frequent** - Use the most frequent label from train dataset as constant prediction. This method shows inbalance in datasets.

PEFT

- **/SF/** - fine-tuning with SetFit method.
- **/LP/** - Linear probing. Logistic regression on top of LM representations.
- **/FT/** - Head-based Fine-tuning.
- **/Full FT/** - Fine-tuning on all training data.

In-context learning

- *eval method*:
  - **EM** - Exact match. Check label in generated substring. If no label is matched, output special label and calculate as wrong prediction.
  - **logprob** - Calculate log probability of sequence with given label. Choose sequence with highest probability. Always choose one of the proposed labels.
  - **iter** - Iteratively add "[MASK]" token to generate label sequence. When generated EM approach is used.
- **(n=k)** - number of demonstrations (k) used in prompt.

References

[1] Augustyniak et al., 2022, this is the way: designing and compiling lepiszcze, a comprehensive nlp benchmark for polish.

[2] Bach et al., 2022, PromptSource: An integrated development environment and repository for natural language prompts.

[3] Radford et al., 2019, language models are unsupervised multitask learners.

[4] Rybak et al., 2020, KLEJ: Comprehensive benchmark for Polish language understanding.

[5] Touvron et al., 2023, llama 2: Open foundation and fine-tuned chat models.

allegro ML Research