

Small Language Models are Good Too An Empirical Study of Zero-Shot Classification

Pierre Lepagnol^{1,2}, Thomas Gerald¹, Sahar Ghannay¹, Christophe Servan^{1,3}, Sophie Rosset¹
¹Université Paris-Saclay, CNRS, LISN, ²SCIAM, ³QWANT

Our Study on Zero-shot Text classification

What:

- An Evaluation of the performance of **9 major models** with different sizes leading to 63 combinations on **15 classifications datasets**.

How:

- Using Zero-Shot Prompting with **Pattern-Verbalizer** method and Statistical Tools.

Why:

- Understand the performance of models on different tasks and datasets.

Contributions:

1. Showing the effectiveness of small models in zero-shot classification,
2. a fully open-source repository: The code is available online in this repository.



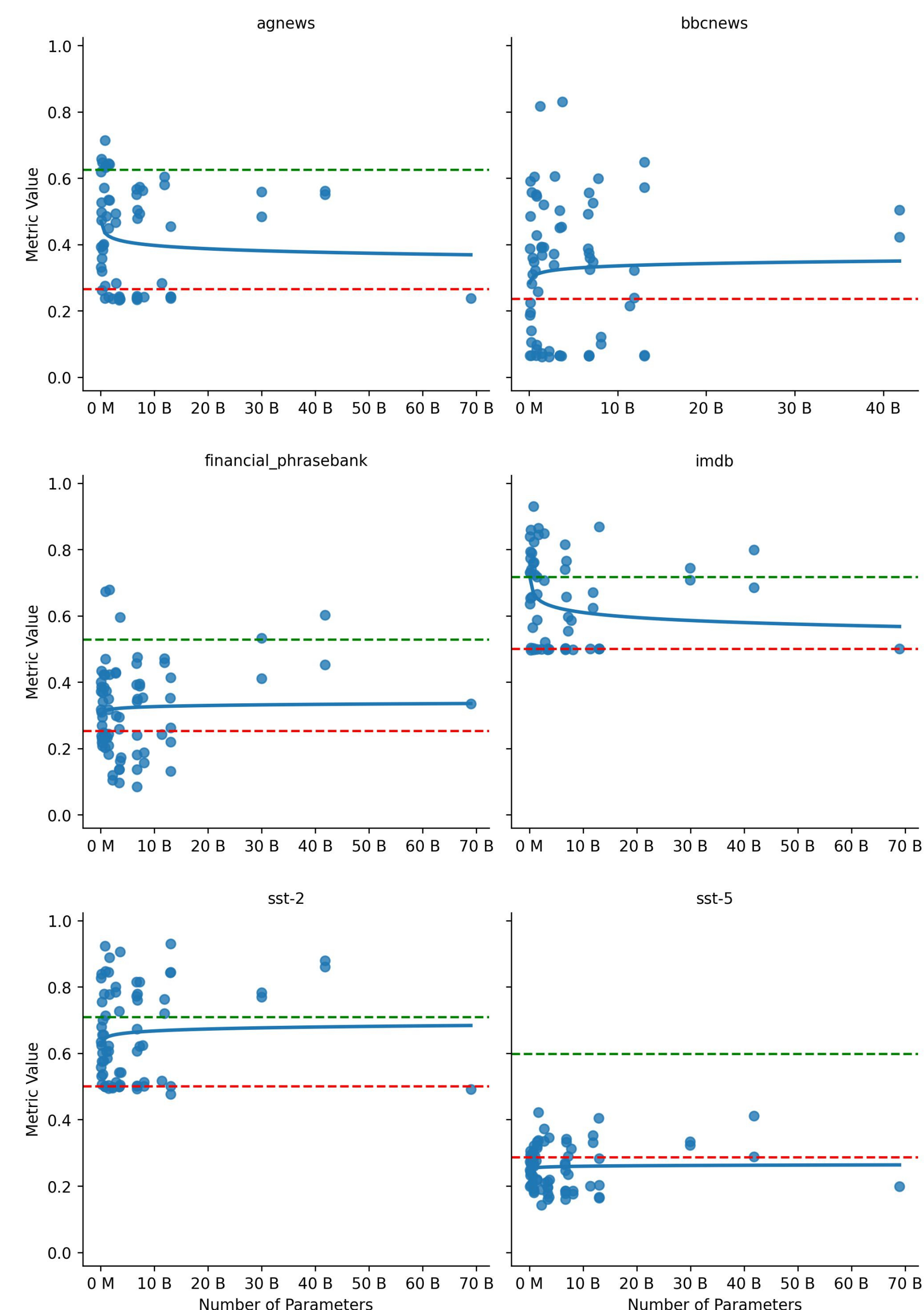
Benchmarked Models

- Bloom/z & mT0 (bigscience)
- Falcon (Tii)
- LaMini-Family (MBZUAI)
- Pythia (EleutherAI)
- MPT (mosaicml)
- Dolly-v2 (databricks)
- Open_llama v2 (openlm-research)
- Orca_Family (pankajmathur)
- T5 (Google) & Bart (Facebook)

Datasets

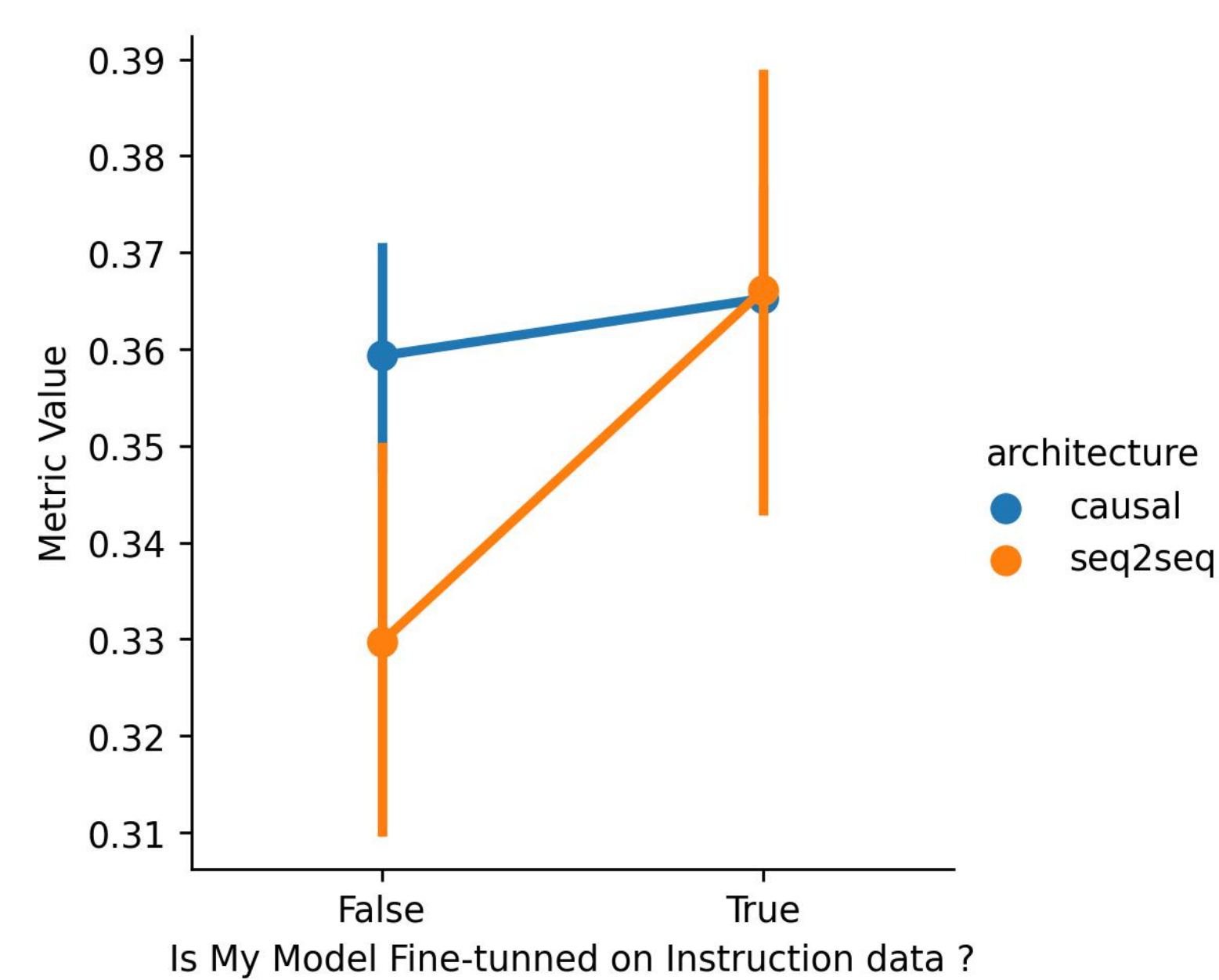
- Topic Classification (AGNews, BBCNews, financial phrasebank)
- Sentiment Classification (ETHOS, IMDB, SST2, SST5, Yelp)
- Relation Classification (CDR, Chemprot, SemEval, Spouse)
- Spam Classification (SMS, Youtube)
- Question Classification (TREC)

Model Size doesn't really matter



- The model size doesn't have a statistically significant impact on performance

Impact of Instruction-Tuning on Performance



Seq2Seq Models Benefit more of instruction-tuning than Causal Models

Beating the SOTA with Small Models

dataset	SOTA Scores	Best Score	Model Used	Number of parameters
agnews	0.625	0.734	MBZUAI/LaMini-GPT-124M	163.0 Millions
bbcnews	NaN	0.869	bigscience/mt0-large	1.2 Billions
cdr	NaN	0.717	bigscience/bloomz-3b	3.6 Billions
chemprot	0.172	0.192	bigscience/bloomz-3b	3.6 Billions
ethos	0.667	0.597	bigscience/bloomz-1b1	1.5 Billions
financial_phrasebank	0.528	0.744	MBZUAI/LaMini-GPT-774M	838.4 Millions
imdb	0.718	0.933	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
semeval	0.435	0.270	bigscience/mt0-xxl	12.9 Billions
sms	0.340	0.699	mosaicml/mpt-7b	6.6 Billions
spouse	0.630	0.521	gpt2	163.0 Millions
sst-2	0.710	0.956	bigscience/bloomz-3b	3.6 Billions
sst-5	0.598	0.485	tiuae/falcon-40b-instruct	41.8 Billions
trec	NaN	0.324	mosaicml/mpt-7b-instruct	6.6 Billions
yelp	0.888	0.977	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
youtube	0.468	0.716	tiuae/falcon-40b	41.8 Billions

Acknowledgements

This work is supported by the ANRT (Association nationale de la recherche et de la technologie) with a CIFRE fellowship granted to SCIAM. (CIFRE N°2022/1608)
This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014242).