LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

SCIAM

Small Language Models are Good Too

An Empirical Study of Zero-Shot Classification

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, Sophie Rosset

Laboratoire Interdisciplinaire de Science Numérique & SCIAM

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Although LLMs are extremely useful, they come with significant drawbacks.

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Although LLMs are extremely useful, they come with significant drawbacks.

They require:

1. Massive amounts of sophisticated datasets to perform well and not be too sensitive to the prompt design.

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Although LLMs are extremely useful, they come with significant drawbacks.

They require:

1. Massive amounts of sophisticated datasets to perform well and not be too sensitive to the prompt design.
2. A lot of resources to train and run, such as massive GPU infrastructure.

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Although LLMs are extremely useful, they come with significant drawbacks.

They require:

1. Massive amounts of sophisticated datasets to perform well and not be too sensitive to the prompt design.
2. A lot of resources to train and run, such as massive GPU infrastructure.

Motivation

Large Language Models (LLMs) have become a **go-to solution** for tackling lots of tasks **without any training examples** to learn from.

Although LLMs are extremely useful, they come with significant drawbacks.

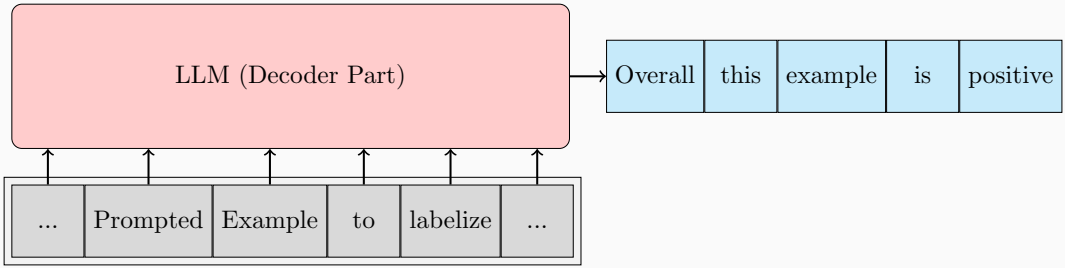
They require:

1. Massive amounts of sophisticated datasets to perform well and not be too sensitive to the prompt design.
2. A lot of resources to train and run, such as massive GPU infrastructure.

But for simpler tasks like classification, do we really need large(r) models?

Lots of ways to use LLMs as classifiers:

Generate free response with multiple tokens + parse the response to detect the class.



⟶ Parsing the output string makes small models hard to use in zero-shot

Generate a single next token and map it on the class.

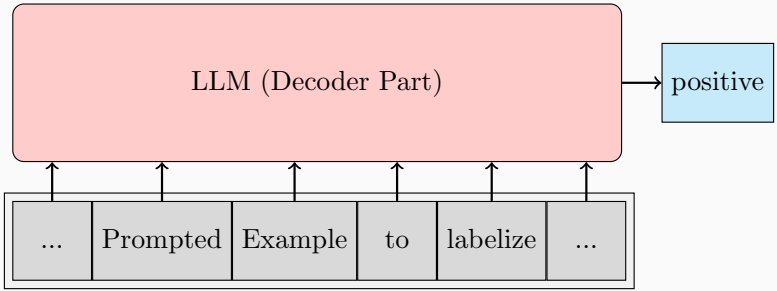


Figure 1: (Pattern-Verbalizer for Zero-shot Setting) from Schick, 20211.

⟶ Much more suitable in zero-shot

Our Study, our Experiments

We have benchmarked:

- 63 models:

Our Study, our Experiments

We have benchmarked:

- 63 models:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models

Our Study, our Experiments

We have benchmarked:

- 63 models:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 datasets**:

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 datasets**:
  - Topic Classification (AGNews, BBCNews, financial_phrasebank)

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 datasets**:
  - Topic Classification (AGNews, BBCNews, financial_phrasebank)
  - Sentiment Classification (ETHOS, IMDB, SST2, SST5, Yelp)

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 datasets**:
  - Topic Classification (AGNews, BBCNews, financial_phrasebank)
  - Sentiment Classification (ETHOS, IMDB, SST2, SST5, Yelp)
  - Relation Classification (CDR, Chemprot, SemEval, Spouse)

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 datasets**:
  - Topic Classification (AGNews, BBCNews, financial_phrasebank)
  - Sentiment Classification (ETHOS, IMDB, SST2, SST5, Yelp)
  - Relation Classification (CDR, Chemprot, SemEval, Spouse)
  - Spam Classification (SMS, Youtube)

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 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)

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 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)
- Using statistical analysis tools:

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 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)
- Using statistical analysis tools:
  - **Biweight Midcorrelation Coefficient**: for correlation analysis

Our Study, our Experiments

We have benchmarked:

- **63 models**:
  - Causal Decoder (Like BLOOM or GPT) & Encoder-Decoder (Like T5) models
  - Instruction-Tuned or not
  - From **70M** to **40B** parameters
- On **15 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)
- Using statistical analysis tools:
  - **Biweight Midcorrelation Coefficient**: for correlation analysis
  - **ANCOVA** (Analysis of Covariance): for comparing categorial aspect for models

Results

Small Models can beat the SOTA in zero-shot i

dataset	SOTA Scores	Majority Class - Scores	Best Score	Model Used	Number of parameters
agnews	0.625	0.266	0.734	MBZUAI/LaMini-GPT-124M	163.0 Millions
bbcnews	NaN	0.236	0.869	bigscience/mt0-large	1.2 Billions
cdr	NaN	0.676	0.717	bigscience/bloomz-3b	3.6 Billions
chemprot	0.172	0.049	0.192	bigscience/bloomz-3b	3.6 Billions
ethos	0.667	0.566	0.597	bigscience/bloomz-1b1	1.5 Billions
financial phrasebank	0.528	0.254	0.744	MBZUAI/LaMini-GPT-774M	838.4 Millions
imdb	0.718	0.500	0.933	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
semeval	0.435	0.054	0.270	bigscience/mt0-xxl	12.9 Billions
sms	0.340	0.464	0.699	mosaicml/mpt-7b	6.6 Billions
spouse	0.630	0.479	0.521	gpt2	163.0 Millions
sst-2	0.710	0.501	0.956	bigscience/bloomz-3b	3.6 Billions
sst-5	0.598	0.286	0.485	tiiuae/falcon-40b-instruct	41.8 Billions
trec	NaN	0.072	0.324	mosaicml/mpt-7b-instruct	6.6 Billions
yelp	0.888	0.522	0.977	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
youtube	0.468	0.528	0.716	tiiuae/falcon-40b	41.8 Billions

Model Size does’nt really matter

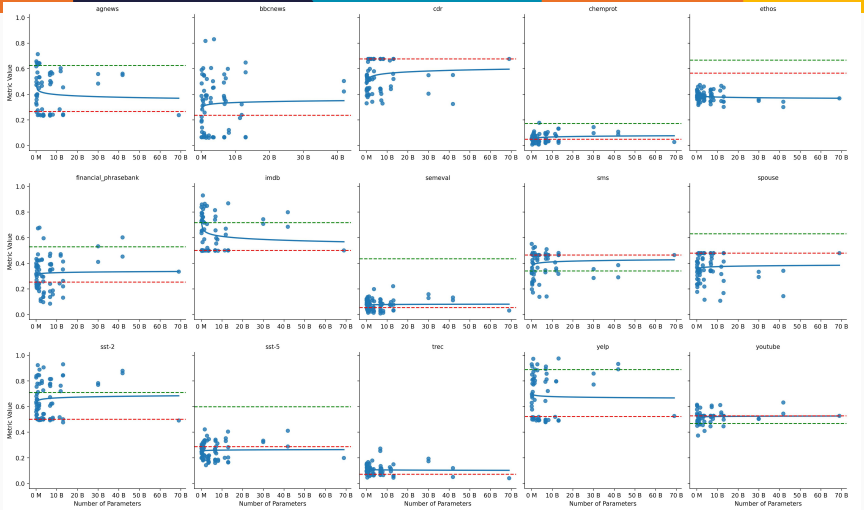
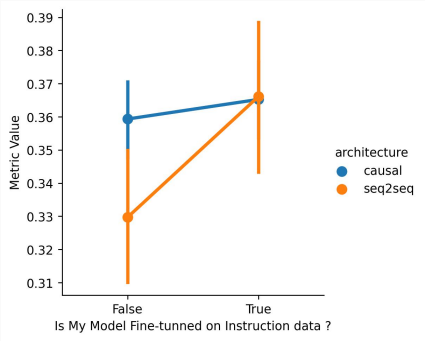


Figure 2: Performance Score Analysis Across Different Model Sizes Using Linear Regression

Seq2Seq Models Benefit more of instruction-tuning than Causal Models



dataset	statistic	pvalue	Equal Variances
causal	0.1825	0.6693	True
seq2seq	6.9406	0.0086	False

The increase in scores is more significant for Seq2Seq models than for Causal models.

Conclusion

- Don’t always need large models for zero-shot classification.
- Small models can be competitive with SOTA.
- Instruction-tuning is more beneficial for Seq2Seq models than for Causal models.
- Small models are good too!

Conclusion

- Don't always need large models for zero-shot classification.
- Small models can be competitive with SOTA.
- Instruction-tuning is more beneficial for Seq2Seq models than for Causal models.
- Small models are good too!

Future work:

- Investigate the variability of the prompt design on the performance of small models.
- Lot's of new models camed out since our study.
- Investigate the impact of Alignment Methods on the performance of small models.

Schick, Timo, and Hinrich Schütze. 2021. "Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference." arXiv. http://arxiv.org/abs/2001.07676.