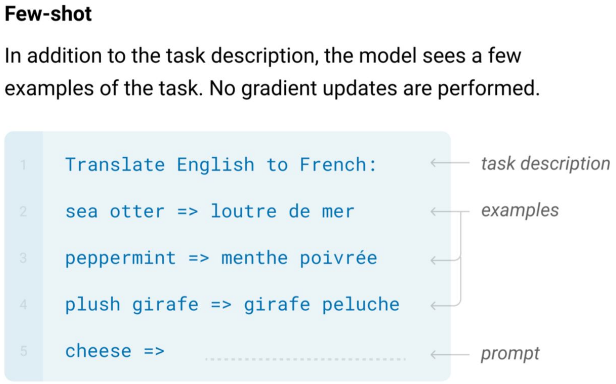
Prompting Large Language Model for Machine Translation: A Case Study

Biao Zhang, Barry Haddow, Alexandra Birch

Edinburgh
University of Edinburgh
Natural Language Processing
NLP

The 40th International Conference on Machine Learning (ICML 2023)

Prompting LLM gains popularity and also works for MT



GPT-3's in-context learning or few-shot prompting for machine translation

		0-shot		1-shot		Few-shot		Supervised
Src	Tgt	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Finetuned SOTA
en	fr	32.9[a]	**38.5**	28.3[b]	**37.5**	33.9[a] (9)	**44.0**	45.6[c]
en	de	25.4[a]	**31.8**	26.2[b]	**31.8**	26.8[a] (11)	**37.4**	41.2[d]
en	ro	16.7[a]	**24.2**	20.6[b]	**28.2**	20.5[a] (9)	**28.7**	33.4[e]
fr	en	35.5[a]	**41.1**	33.7[b]	**37.4**	38.0[a] (9)	**42.8**	45.4[f]
de	en	38.9[a]	**43.8**	30.4[b]	**43.9**	40.6[a] (11)	**47.5**	41.2[g]
ro	en	36.8[a]	**39.9**	38.6[b]	**42.1**	37.3[a] (9)	**43.8**	39.1[h]

MT results (BLEU) on WMT datasets for PaLM

Brown et al., 2020; Chowdhery et al. 2022

…… but **it is non-trivial**

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:       ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese => ...........               ← prompt
```

GPT-3's in-context learning or few-shot prompting for machine translation

Src	Tgt	0-shot Prior SOTA	0-shot PaLM 540B	1-shot Prior SOTA	1-shot PaLM 540B	Few-shot Prior SOTA	Few-shot PaLM 540B	Supervised Finetuned SOTA
en	fr	32.9[a]	**38.5**	28.3[b]	**37.5**	33.9[a] (9)	**44.0**	45.6[c]
en	de	25.4[a]	**31.8**	26.2[b]	**31.8**	26.8[a] (11)	**37.4**	41.2[d]
en	ro	16.7[a]	**24.2**	20.6[b]	**28.2**	20.5[a] (9)	**28.7**	33.4[e]
fr	en	35.5[a]	**41.1**	33.7[b]	**37.4**	38.0[a] (9)	**42.8**	45.4[f]
de	en	38.9[a]	**43.8**	30.4[b]	**43.9**	40.6[a] (11)	**47.5**	41.2[g]
ro	en	36.8[a]	**39.9**	38.6[b]	**42.1**	37.3[a] (9)	**43.8**	39.1[h]

MT results on WMT datasets for PaLM

A systematic study of how prompting works for MT is still missing!

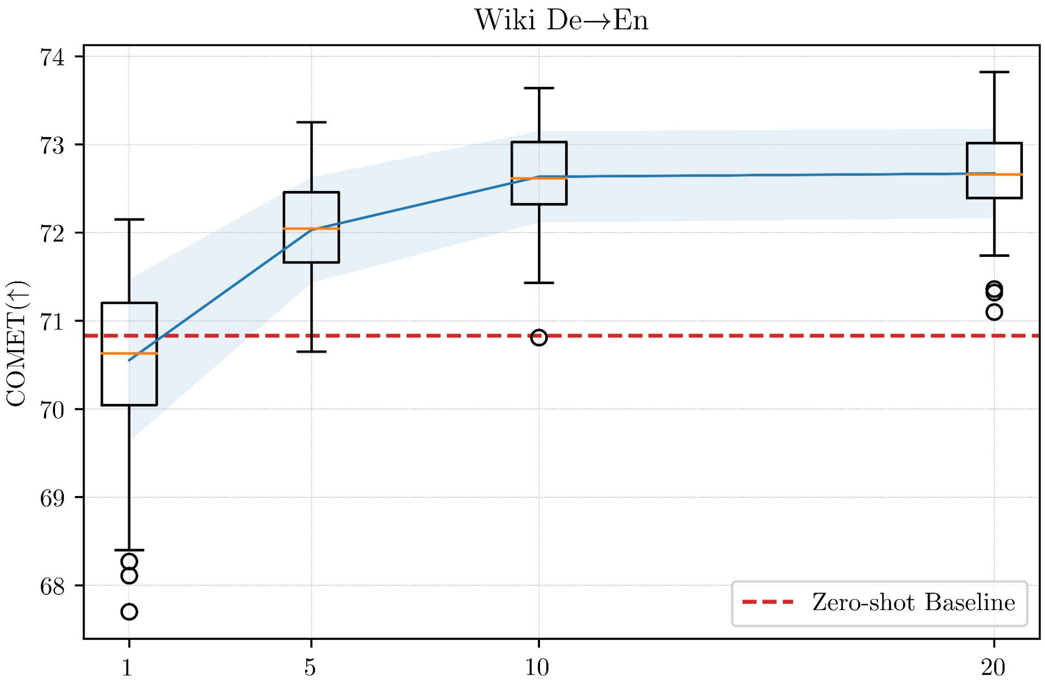
Research Questions

- What’s the best ***prompting strategy*** for MT?
- Can we use ***monolingual data*** for few-shot prompting?
- Can we ***transfer prompt*** across different settings?

Experimental setup

- Model
  - GLM-130B [Chinese and English] (quantized version)
- Languages
  - English (En), German (De), Chinese (Zh)
- Dataset
  - Wiki (Flores, En-De-Zh), WMT21 (En-De, En-Zh), Multi-Domain (IT, Law and Medical, De-En)
  - PDC for document-level translation (En-Zh)
  - Ablation set: 100 samples from dev as ablation test, the rest as ablation dev
- SacreBLEU, COMET and Spearman correlation

Demonstration greatly affects prompting quality



* The curve: **more examples gives better quality**

* Each box: **quality varies greatly over different examples**

* Results on Wiki De→En Ablation sets for few-shot prompting.
* We randomly sample 100 times from the pool.

Example is important; How to select?

Feature	Model	Case dep	Description
SLength	None	No	Source length
TLength	None	No	target length
LMScore	**GLM-130B**	No	log likelihood of GLM
MTScore	**COMET QE**	No	MT quality of prompt example
SemScore	**LASER2**	No	cosine semantic similarity of prompt example
CaseSemScore-Src	**LASER2**	**Yes**	SemScore(source example, test input)
CaseSemScore-Tgt	**LASER2**	**Yes**	SemScore(target example, test input)

input and model agnostic

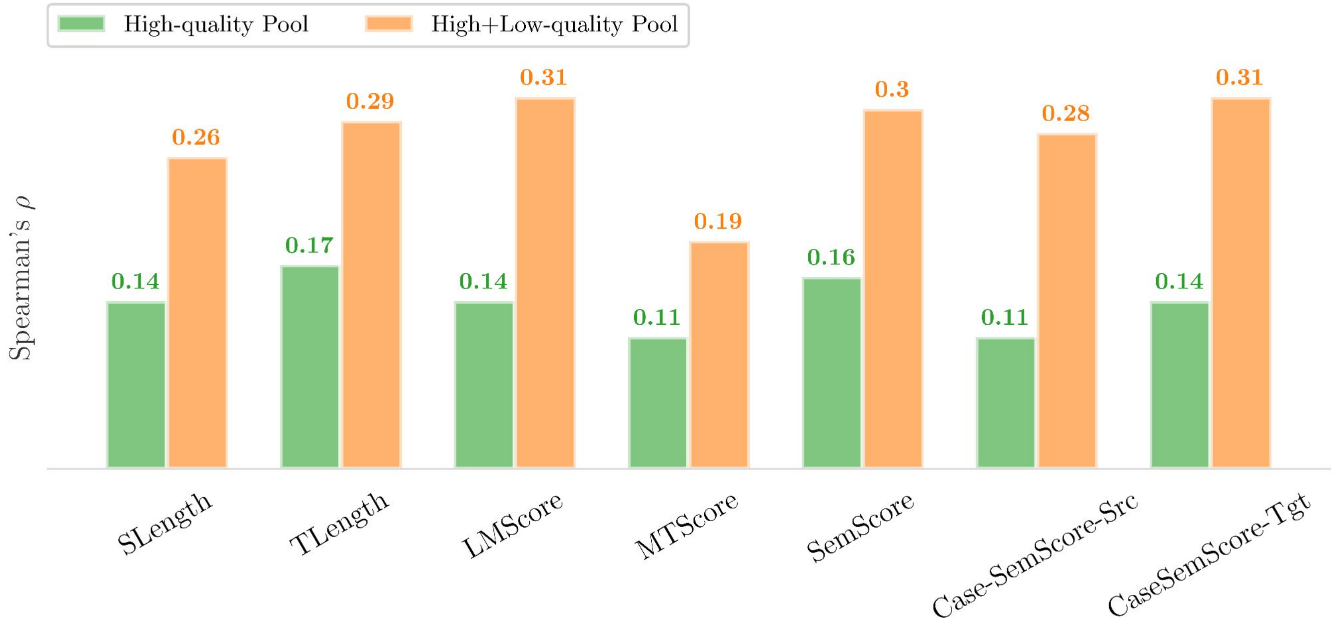
model dependent but input agnostic

input and model dependent

*** We extract a set of features and check their correlation with prompting results**

* We focus on **1-shot prompting** to simplify the setup

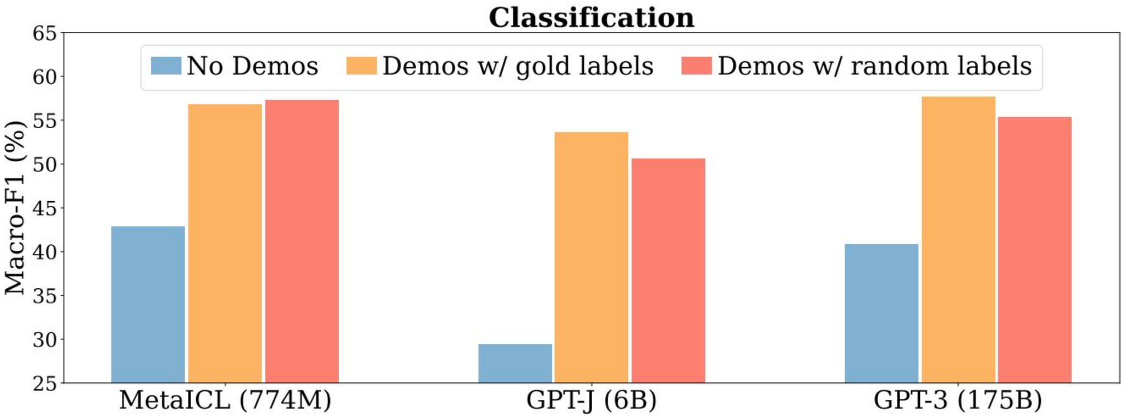
Features show significant yet weak correlation



* Average Spearman scores (over 6 directions) for 1-shot prompting on Wiki Ablation sets.
* **High-quality pool**: FLORES Ablation dev set; **Low-quality pool**: WikiMatrix v1.
* We randomly sample **600 demonstrations** to compute the correlation.

Do we need genuine MT pairs for demonstration?

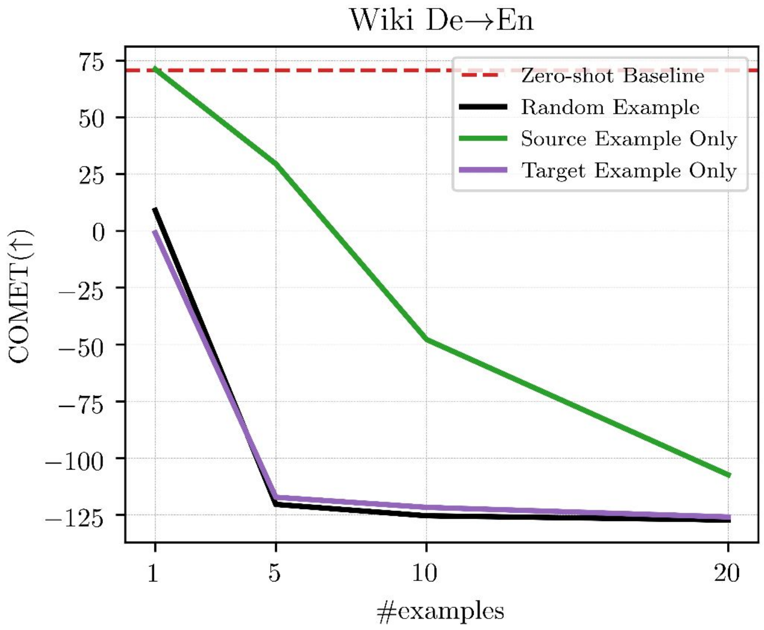
- Ground truth demonstrations may be unimportant (Min et al., 2022).



Does this also apply to MT? Further, can we use monolingual data for prompting?

Example taken from *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? (Min et al., 2022)*

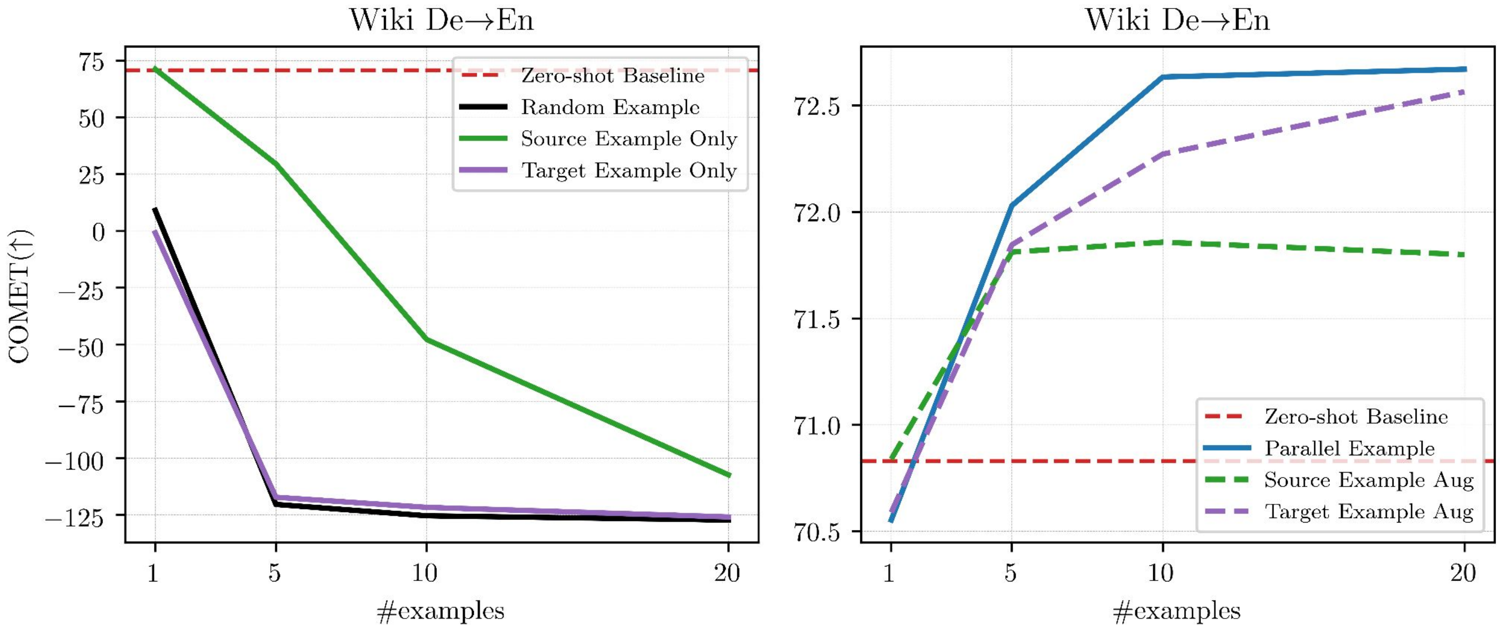
MT prompting requires genuine input-output mapping



* Using random examples hurts prompting
* Source-only or target-only monolingual prompting doesn’t work

Results on Wiki De→En Ablation sets; we randomly sample 50 demonstrations and report average performance.

Can we use mono data? Yes! For/backward translation



* Pseudo parallel data (based on zero-shot prompting) benefits prompting

* Back-translation performs better than forward-translation

Choosing demonstrations is hard; can we transfer it?

Setting *A*

We have rich resources and found good demonstrations, D_1 and D_2 , and

$$\text{perf}(D_1) > \text{perf}(D_2)$$

Setting *B*

1. $\text{perf}(D_1)$ or $\text{perf}(D_2) >$ zero-shot prompting ?
2. $\text{perf}(D_1) > \text{perf}(D_2)$?

Cross-lingual transfer

e.g. En→De ⇒ De→Zh

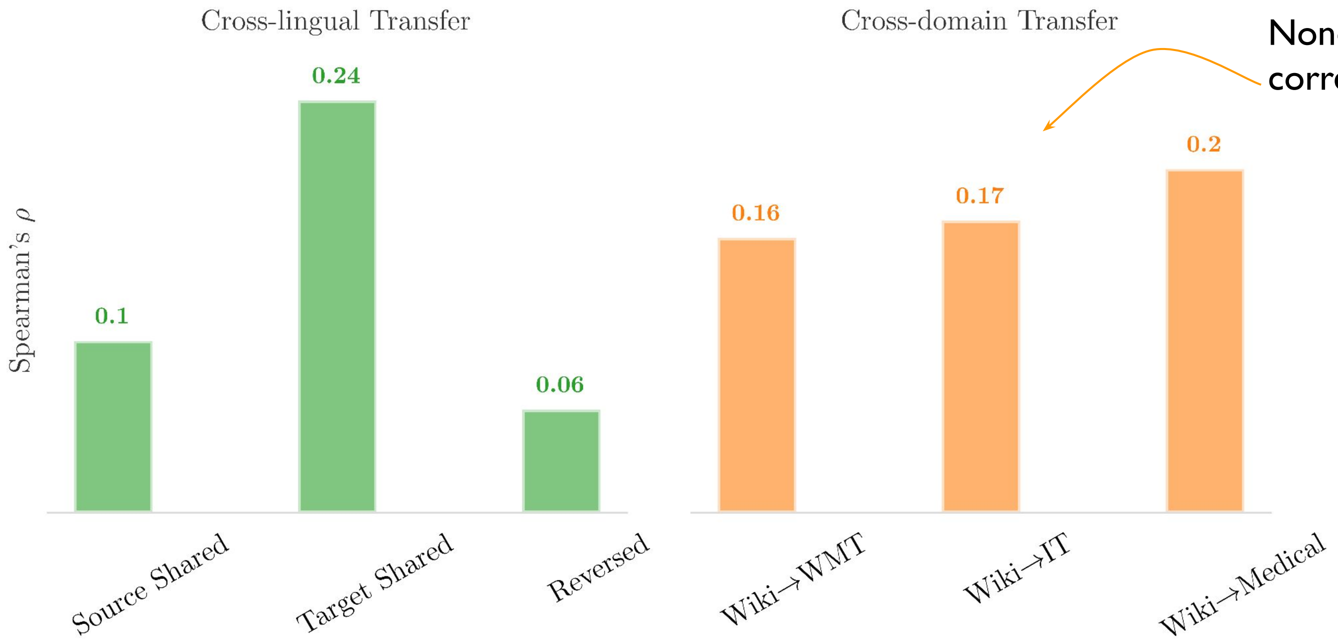
Cross-domain transfer

e.g. Wiki to WMT, IT, Medical

Sentence-to-document transfer

e.g. Sent-MT to Doc-MT

The superiority of a demonstrate **doesn't generalize**



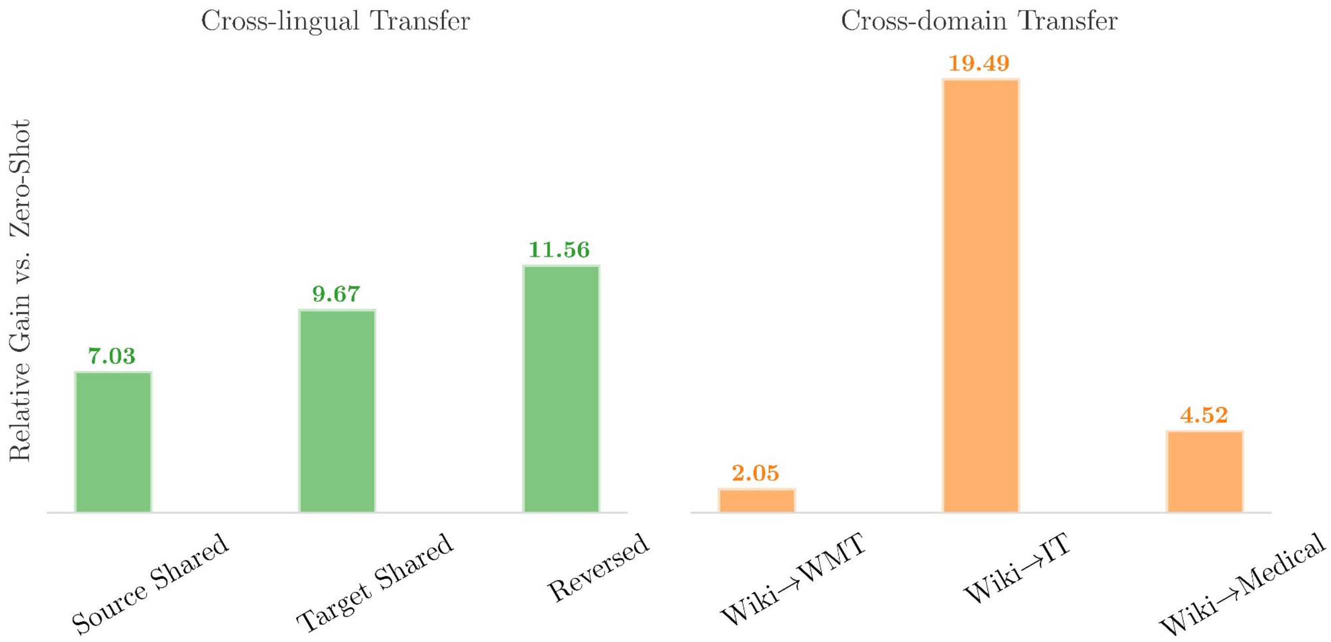
None of the correlations are strong.

* Source/target shared: transfer when source/target language is the same.

* Reversed: transfer between reversed language pairs.

* We randomly sample 200-300 demonstrations to obtain the correlation on Ablation sets.

Out-of-setting demonstration **beats** zero-shot MT



Few-shot prompting with out-of-setting demonstrations is preferred than zero-shot prompting

Takeaways:

- Prompting performance varies greatly across templates
- Selecting examples via simple features is not very promising
- MT prompting doesn't work with monolingual data alone; Use pseudo parallel examples instead.
- Transfer learning of prompting is feasible
- Prompting LLMs for MT still faces problems

Check out our paper for more details
https://arxiv.org/abs/2301.07069