Prompting Large Language Model for Machine Translation: A Case Study

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the 40th International Conference on Machine Learning

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Prompting Large Language Model for Machine Translation: A Case Study

Biao Zhang  Barry Haddow  Alexandra Birch

Abstract
Research on prompting has shown it to have excellent performance with little or even no supervised training across many tasks. However, prompting for machine translation is still under-explored in the literature. We fill this gap by offering a systematic study on prompting strategies for translation, examining various factors for prompt template and demonstration example selection. We further explore the use of monolingual data and the feasibility of cross-lingual, cross-domain, and sentence-to-document transfer learning in prompting. Extensive experiments with GLM-130B (Zeng et al., 2022) as the testbed show that 1) the number and the quality of prompt examples matter, where using suboptimal examples degenerates translation; 2) several features of prompt examples, such as semantic similarity, show significant Spearman correlation with their prompting performance; yet, none of the correlations are strong enough; 3) using pseudo parallel prompt examples constructed from monolingual data via zero-shot prompting could improve translation; and 4) improved performance is achievable by transferring knowledge from prompt examples selected in other settings. We finally provide an analysis on the model outputs and discuss several problems that prompting still suffers from.

1. Introduction
Large language models (LLMs) pretrained on massive unlabeled corpora have shown impressive emergent abilities under model scaling which enable prompting for downstream applications (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022b; Zhang et al., 2022a; Chowdhery et al., 2022). Different from task-specific finetuning, constructing task-specific prompts by rephrasing test examples with descriptive task instructions and executing the task by feeding prompts to LLMs directly. It can be further enhanced through in-context learning by providing a few labeled examples (or prompt examples) as a demonstration (Brown et al., 2020). As a new paradigm, prompting LLMs has achieved state-of-the-art performance over a range of natural language processing (NLP) tasks (Chung et al., 2022; Goyal et al., 2022; Wei et al., 2022c;a; Chowdhery et al., 2022).

In this paper, we focus on prompting LLMs for machine translation (MT). MT represents a complex task requiring transforming a source input into its semantically equivalent target output in a different language, which combines sequence understanding and generation. It offers a unique platform to assess the cross-lingual generation capability of LLMs, and the assessment may shed light on pretraining/finetuning algorithm design for achieving universal LLMs (Chowdhery et al., 2022). While a few studies have reported translation results (Brown et al., 2020; Reynolds & McDonell, 2021; Chowdhery et al., 2022), a systematic study on how prompting works for MT is still missing in the literature.

We aim at filling this gap by thoroughly examining different prompting setups using the recently released GLM-130B (Zeng et al., 2022), particularly concerning three aspects: the prompting strategy, the use of unlabeled/monolingual data, and the feasibility of transfer learning. Prompting has shown varying sensitivity to the choice of prompt templates and examples (Zhao et al., 2021). For MT, prior studies adopted different templates (Brown et al., 2020; Wei et al., 2022a; Chowdhery et al., 2022), and we reevaluate them to figure out the optimal one. We further design a set of features for prompt examples and explore which one(s) could explain the prompting performance, according to which we develop the example selection strategy.

Since leveraging monolingual data to improve MT has long been of interest, we would like to determine whether and how such data can be used in prompt example construction. We make a step in this direction by studying the effect of data augmentation using back-/forward-translation (Sennrich et al., 2016b; Zhang & Zong, 2016) via zero-shot prompting. In addition, neural MT and pretrained LLMs

---

*Now at Google DeepMind; work done prior to joining Google.
1School of Informatics, University of Edinburgh. Correspondence to: Biao Zhang <biaojiaxing@google.com>, Barry Haddow <bhaddow@inf.ed.ac.uk>, Alexandra Birch <a.birch@ed.ac.uk>.

have shown encouraging transfer abilities (Devlin et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020; Xue et al., 2021) but transfer learning for prompting has received little attention. Whether prompt examples are transferable across different settings, such as from one domain/language pair to another and from sentence-level examples to document-level translation, is yet to be addressed.

We address the above concerns with GLM-130B as the testbed and conduct extensive experiments on FLORES and WMT evaluation sets. We mainly study translation for three languages: English, German and Chinese. We also provide a quantitative and qualitative analysis to disclose problems when prompting for MT, which might offer insights for future study. Our main findings are listed as below:

- Prompting performance varies greatly across templates, and language-specific templates mainly work when translating into languages LLMs are pretrained on. An English template in a simple form works best for MT.

- Several features of prompt examples, such as sequence length, language model score, and semantic similarity, correlate significantly with its prompting performance while the correlation strength is weak in general. Selecting examples based on these features can outperform the random strategy, but not consistently.

- Using monolingual examples for prompting hurts translation. By contrast, constructing pseudo parallel examples via back-/forward-translation is a good option. Back-translation performs better and is more robust.

- Prompting shows some degree of transferability. Using demonstrations from other settings can improve translation over the zero-shot counterpart, while the superiority of a demonstration in one setting can barely generalize to another.

- Prompting for MT still suffers from copying, mistranslation of entities, hallucination, inferior direct non-English translation, and prompt trap where translating the prompt itself via prompting becomes non-trivial.

2. Setup

Prompting for MT  Given a pretrained and fixed LLM $L$, MT prompting first converts each test input $X$ to a prompt according to a template $T$ and then generates the translation $Y$ by feeding the prompt to $L$. In this study, we consider zero-shot and few-shot prompting for translation.

Zero-shot prompting only has access to the test input $X$, while few-shot prompting assumes that a few extra labeled examples (or prompt/demonstration examples) $D^P = \{X'_i, Y'_i\}_{i=1}^K$ are available and can be used as a demonstration. Particularly, we adopt the following template for zero-shot prompting based on the results in Section 3:

$$[\text{src}]: X \rightarrow [\text{tgt}]: Y$$

where $[\text{src}]$ and $[\text{tgt}]$ denote test language(s), i.e., the source and target language name of the test language pair, respectively. For few-show prompting, we concatenate the given prompt examples:

$$[\text{src}]: X'_1 \rightarrow [\text{tgt}]: Y'_1 \ldots [\text{src}]: X'_K \rightarrow [\text{tgt}]: Y'_K$$

where $[\text{src}]$ and $[\text{tgt}]$ denote prompt language(s), i.e., the source and target language name of the prompt example, respectively. By default, prompt examples and test data are in the same language pair. However, when considering cross-lingual transfer for prompting, prompt examples might be in a different language pair.

We also explore template language, which denotes the language in which the template is expressed. For example, the Chinese template “中：X 英文：” represents the Chinese counterpart of the following English template “Chinese: X English: “.

Setting  We experiment with GLM-130B, a LLM with 130B parameters pretrained on Chinese and English “monolingual” corpora, which was reported to outperform GPT-3 and OPT-175B on several NLP tasks (Zeng et al., 2022). Note GLM-130B is a raw LLM without any further fine-tuning. We use its INT4-quantized version, which is more affordable and suffers little performance degradation. We adopt beam search for MT with a beam size of 2, and perform experiments with 4 RTX 3090 or A100-40G GPUs.

We work on three languages: English (En), German (De), and Chinese (Zh). We perform major analysis on FLORES (Wiki domain, En-De-Zh, NLLB Team et al., 2022) and WMT21 (News domain, En-De, En-Zh, Akhbardeh et al., 2021), and also report results on Multi-Domain (IT, Law and Medical domain, De-En, Aharoni & Goldberg, 2020) to examine domain robustness and transfer ability, and PDC (News domain, Zh→En, Sun et al., 2022) for document-level translation. To understand the relation between prompt examples and their prompting performance, we construct an Ablation set for Wiki, WMT and Multi-Domain (IT and Medical) based on the dev set of FLORES, WMT21 and Multi-Domain, separately, where we randomly sample 100 instances as the ablation test set and use the rest as the default example selection pool. To distinguish, we will refer to the official dev and test set as Full set. Detailed statistics are listed in Table 1.

We evaluate translation performance using both a surface-based metric, detokenized case-sensitive BLEU↑ from
Prompting Large Language Model for Machine Translation: A Case Study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language(s)</th>
<th>Test Set</th>
<th>Selection Pool (Default)</th>
<th>Source (#sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>English</td>
<td>100</td>
<td>897</td>
<td>FLORES eng_Latn.dev (997)</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>100</td>
<td>897</td>
<td>FLORES deu_Latn.dev (997)</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>100</td>
<td>897</td>
<td>FLORES zho_Hans.dev (997)</td>
</tr>
<tr>
<td>WMT</td>
<td>English-German</td>
<td>100</td>
<td>2900</td>
<td>newstest2013 (3000)</td>
</tr>
<tr>
<td></td>
<td>IT</td>
<td>100</td>
<td>1900</td>
<td>Multi-Domain Dev Set (2000)</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>100</td>
<td>1900</td>
<td>Multi-Domain Dev Set (2000)</td>
</tr>
</tbody>
</table>

(a) Ablation Sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Source</th>
<th>Test Set</th>
<th>High-quality Pool (Default)</th>
<th>Low-quality Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>English</td>
<td>FLORES</td>
<td>eng_Latn.devtest (1012)</td>
<td>eng_Latn.dev (997)</td>
<td>En-Zh* (0.79M)</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>FLORES</td>
<td>deu_Latn.devtest (1012)</td>
<td>deu_Latn.dev (997)</td>
<td>De-En* (1.57M)</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>FLORES</td>
<td>zho_Hans.devtest (1012)</td>
<td>zho_Hans.dev (997)</td>
<td>De-Zh* (0.13M)</td>
</tr>
<tr>
<td>WMT</td>
<td>English-German</td>
<td>WMT</td>
<td>newstest2021 (1002/1000)</td>
<td>newstest2020 (1418)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IT</td>
<td>Multi-Domain</td>
<td>Test Set (2000)</td>
<td>-</td>
<td>Train Set (0.22M)</td>
</tr>
<tr>
<td></td>
<td>Law</td>
<td>Multi-Domain</td>
<td>Test Set (2000)</td>
<td>-</td>
<td>Train Set (0.47M)</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>Multi-Domain</td>
<td>Test Set (2000)</td>
<td>-</td>
<td>Train Set (0.25M)</td>
</tr>
<tr>
<td>PDC</td>
<td>Chinese-English</td>
<td>News</td>
<td>Test Set (4858/148 Docs)</td>
<td>Dev Set (2881)</td>
<td></td>
</tr>
</tbody>
</table>

(b) Full Sets

Table 1: Statistics of Ablation sets and Full sets. Numbers in brackets denote the number of instances. *: data from WikiMatrix.v1 (Schwenk et al., 2021).

SacreBLEU (Post, 2018) (with the option -tok zh for Chinese), and a model-based metric, COMET from unbabel-comet with the model wmt20-comet-da (Rei et al., 2020).

3. Prompting Strategy for MT

To perform MT, prompting needs to cast the translation problem into a language modeling problem via the prompt. Thus, the format of the prompt, including its wording, directly affects how the LLM understands the task and its behavior. For MT, we are interested in the following research questions:

- Which template should we use for MT prompting? And what language for the template?
- Does demonstration matter for MT prompting? How to select optimal prompt examples?

We address them through extensive experiments on the Wiki Ablation sets.

Zero-shot prompting performance varies greatly across templates. We start with zero-shot prompting and explore the effect of different templates. Depending on how to describe MT and partially inspired by prior studies (Brown et al., 2020; Chowdhery et al., 2022; Wei et al., 2022a), we compare 6 templates and evaluate them on the Wiki Ablation sets covering 6 language pairs (En±De, En±Zh, De±Zh). Table 2 shows the results (we list detailed results in Table 10, Appendix). The template affects zero-shot quality substantially, and the simple template A in English specifying just the source and target language name achieves the best overall results. In follow-up experiments, we thus focus on template A.

Language-specific template delivers mixed results. Table 2 also shows the prompting results of German and Chinese templates, which often largely underperform their English counterparts. Since German is not a major pretraining language in GLM-130B, a German template degenerates the translation substantially. By contrast, a Chinese template yields improved quality when translating into Chinese (see Table 10). Still, an English template works best on average.

The preference of GLM-130B to English template also shows that the level of language understanding and cross-lingual ability in GLM-130B varies across languages, even though it’s pretrained on the same amount of monolingual Chinese and English tokens. This might be caused by the fact that more cross-lingual code-switched data is mixed into the English pretraining data (note English is used more globally than Chinese), but might also suggest that improving the language understanding of LLM requires more advanced training algorithms beyond scaling training data.
Prompting Large Language Model for Machine Translation: A Case Study

Table 2: COMET scores averaged over 6 language pairs for zero-shot prompting with different templates and different template languages on Wiki Ablation sets. w/ and w/o denote whether adding line breaks into the template or not; ∘ indicates the position of the line break. [src] and [tgt] denote source and target test language name, respectively, and [input] denotes the test input; all of them are placeholders. English, German and Chinese indicate template languages. Best results are shown in bold.

<table>
<thead>
<tr>
<th>ID</th>
<th>Template (in English)</th>
<th>English</th>
<th>German</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[src]: [input] ∘ [tgt]:</td>
<td>38.78</td>
<td>-26.15</td>
<td>14.82</td>
</tr>
<tr>
<td>B</td>
<td>[input] ∘ [tgt]:</td>
<td>-88.62</td>
<td>-135.97</td>
<td>-66.55</td>
</tr>
<tr>
<td>C</td>
<td>[input] ∘ Translate to [tgt]:</td>
<td>-87.63</td>
<td>-106.30</td>
<td>-63.38</td>
</tr>
<tr>
<td>D</td>
<td>[input] ∘ Translate from [src] to [tgt]:</td>
<td>-113.80</td>
<td>-153.80</td>
<td>-76.79</td>
</tr>
<tr>
<td>E</td>
<td>[src]: [input] ∘ Translate to [tgt]:</td>
<td>20.81</td>
<td>16.69</td>
<td>-8.61</td>
</tr>
</tbody>
</table>

Figures 1 and 2: COMET scores for few-shot prompting as a function of the number of prompt examples (K = 1, 5, 10, 20) on Wiki Ablation sets. For each setup, we randomly sample 100 times from the example pool and show the performance distribution via box plots. Dashed red line denotes the zero-shot baseline; blue curve and shadow area denote the mean and standard deviation.

The performance of demonstration is not stable. However, we also see high performance variance under the same K. It’s possible that a demonstration with 5 examples outperforms its 10 or 20 counterpart. Figure 1 also shows that 1-shot prompting underperforms zero-shot prompting in many cases, even on average. This echoes with previous findings on other NLP tasks (Zhao et al., 2021; Liu et al., 2022) and also highlights the significance of developing effective example selection strategies.

Note that few-shot prompting greatly improves translation into Chinese. The reason based on our manual analysis is that the zero-shot baseline tends to translate into traditional Chinese with messy codes, whereas prompt examples help (the reference text is always simplified Chinese).

Using more prompt examples for demonstration improves translation significantly on average. We next study few-shot prompting following the template A but in format (2) with K varying from 1 to 20. We evaluate multiple demonstrations for each K via random sampling to reduce data biases. Figure 1 shows that the more examples used, the better average performance (more results are shown in Figure 5, Appendix), albeit at the cost of using more GPU memory and increasing the inference time per token as in Figure 2.

Several features correlate with prompting performance significantly yet weakly. We thus turn to explore example selection for prompting. Our idea is to extract a couple of diverse features from demonstrations and examine whether any of them are informative enough to be used as an indicator for the selection. In this study, we simplify our analysis by focusing on 1-shot prompting, which ignores the ordering of prompt examples (we return to few-shot prompting later). Particularly, we extract and analyze 7 features of a demonstration:

1. **ID Template (in English)**
2. **English**
3. **German**
4. **Chinese**
5. **Zero-shot Baseline**
6. **Wiki De→En**
7. **Wiki En→De**
8. **Wiki En→Zh**
9. **Wiki Zh→En**

Using more prompt examples for demonstration improves translation significantly on average. We next study few-shot prompting following the template A but in format (2) with K varying from 1 to 20. We evaluate multiple demonstrations for each K via random sampling to reduce data biases. Figure 1 shows that the more examples used, the better average performance (more results are shown in Figure 5, Appendix), albeit at the cost of using more GPU memory and increasing the inference time per token as in Figure 2.

Several features correlate with prompting performance significantly yet weakly. We thus turn to explore example selection for prompting. Our idea is to extract a couple of diverse features from demonstrations and examine whether any of them are informative enough to be used as an indicator for the selection. In this study, we simplify our analysis by focusing on 1-shot prompting, which ignores the ordering of prompt examples (we return to few-shot prompting later). Particularly, we extract and analyze 7 features of a demonstration:
Prompting Large Language Model for Machine Translation: A Case Study

<table>
<thead>
<tr>
<th>Feature</th>
<th>BLEU + LQ</th>
<th>COMET + LQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLength</td>
<td>0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>TLength</td>
<td>0.23</td>
<td>0.32</td>
</tr>
<tr>
<td>LMScore</td>
<td>0.20</td>
<td>0.33</td>
</tr>
<tr>
<td>MTScore</td>
<td>0.04</td>
<td>0.14</td>
</tr>
<tr>
<td>SemScore</td>
<td>0.19</td>
<td>0.30</td>
</tr>
<tr>
<td>CaseSemScore-Src</td>
<td>0.14</td>
<td>0.29</td>
</tr>
<tr>
<td>CaseSemScore-Tgt</td>
<td>0.14</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 3: Spearman’s ρ between demonstration features and their prompting performance for 1-shot prompting on Wiki Ablation sets. We randomly sample 600 demonstrations from each pool to calculate the correlation. HQ: examples are from the default high-quality pool; LQ: examples are from the low-quality pool based on WikiMatrix.v1.

S(T)Length  the number of source (target) tokens;

LMScore GLM-130B-based, length-normalized log likelihood of the demonstration;

MTScore translation quality of the prompt example from COMET QE wmt20-comet-qa-da (Rei et al., 2020);

SemScore semantic score based on the cosine similarity of the demonstration’s source and target sentence embeddings from LASER2 (Heffernan et al., 2022);

CaseSemScore-Src similarity to the input that averages over SemScores between the test input and the demonstration’s source;

CaseSemScore-Tgt similar to CaseSemScore-Src but compares to demonstration’s target;

We sample multiple demonstrations randomly and inspect the Spearman’s correlation between feature values and prompting performance. We consider high-quality and low-quality pool for sampling.

Table 3 summarizes the results and Figure 3 illustrates the relation between COMET and LMScore (more results are given in Table 11 and Figures 6, 7, Appendix). With the high-quality pool, different demonstrations yield similar translation results (see blue points) despite their feature values varying greatly. Several features show insignificant and inconsistent correlation, particularly for De→En and Zh→En. This suggests developing selection policy for high-quality example pool is non-trivial.

After mixing with demonstrations from the low-quality pool, the significance gets strengthened. LMScore and CaseSemScore-Tgt shows the highest correlation on average followed by TLength and SemScore. MTScore behaves much worse which might be caused by its instability on sentence-level evaluation (Moghe et al., 2022). However, we didn’t see significant difference in terms of Spearman’s ρ between input-relevant and input-agnostic features (Agrawal et al., 2022), neither among surface-based, LLM-based or semantic-based features. Surprisingly, the simple feature, S/TLength, yields reasonably high correlation. We argue that long examples could offer LLM with more signals about the task’s input and output space. This finding suggests that researchers should select long unlabeled sentences for annotation to improve prompting. Yet, most Spearman’s ρs are much smaller than 0.5, indicating a weak/fragile relation.
We also consider 5-shot prompting, where we concatenate we excluded examples that are too long (more than 100 tokens; to ensure the informativeness) during the selection. Using prompt examples selected based on the proposed features yields improved performance. Unfortunately, none of them can guarantee optimal translation performance.

In general, selecting prompt examples of high translation quality, high semantic similarity, high LLM likelihood, long sequence length and high similarity to test inputs are all preferable strategies. Unfortunately, none of them can guarantee optimal translation performance. Roughly following their relative 1-shot translation performance (SemScore > LMScore > TLength on average). In Table 4, this combined strategy outperforms the random one by varying degrees.

4. Monolingual Data for Prompting

A longstanding concern in MT is how to utilize unlabeled data to improve translation. While prompting enables few-shot learning reducing the data requirement, exploring whether demonstration could benefit from monolingual examples is still valuable, both for MT study and for understanding the role of demonstration in prompting.

Min et al. (2022) argue that the key role of demonstration lies in its support of the input space, the label space and the prompt format, rather than the genuineness of the examples. They found that randomly replacing labels in demonstration barely hurts performance on classification tasks. We reexamine this argument in the context of MT by studying the following three prompting settings: 1) random examples constructing sentence pairs from monolingual sources and targets randomly; 2) source/target example only using monolingual source/target alone for prompting.

Directly using monolingual data for demonstration doesn’t work. Figure 4 (top) shows a totally different story (see Figures 8 and 9 in Appendix for more results): monolingual example-based demonstration almost always hurts translation, and the more examples used, the more degeneration yielded. Using random examples misleads the prompting and performs the worst in general; compared to target-only examples, using source examples yields slightly
better results except translating into Chinese. This indicates that the genuine source-target mapping should be retained in the demonstration, and also indicates that MT features unique challenges which deserves more attention when studying prompting.

**Pseudo parallel examples by forward-/back-translation benefits prompting.** Inspired by data augmentation in MT (Sennrich et al., 2016b; Zhang & Zong, 2016), we next resort to constructing pseudo parallel data. We first adopt GLM-130B to translate the source or target examples via zero-shot prompting, and then use the generated parallel examples as demonstration. Despite low quality, Figure 4 (bottom) shows that this is an effective way to improve prompting, and using more examples often produces better results, partially echoing with the findings on prompting-based unsupervised MT (Han et al., 2021; Patel et al., 2022). We also observe that back-translation (i.e. translating target monolingual examples) performs better and behaves more robustly than forward-translation (i.e. translating source examples instead), which even approaches prompting with genuine parallel examples.

### 5. Transfer Learning for Prompting

After obtaining a performant demonstration, we are interested in to what extent its capability could be transferred across different settings, especially from one domain/language pair to another and from sentence-level to document-level translation. While previous studies demonstrate the feasibility with continuous prompts on classification tasks (Wang et al., 2021a), transfer for hard prompting on MT has never been investigated.

Assume that demonstrations $D_1$ and $D_2$ are selected in setting $S_1$ and that $D_1$ performs better (i.e. $D_1 > D_2$). We have the following research questions:

- Could we also expect $D_1 > D_2$ in setting $S_2$?
- Whether using demonstrations from $S_1$ could outperform zero-shot prompting in $S_2$?

We next study them via experiments with 1-shot prompting.

**The superiority of a demonstration doesn’t generalize across settings.** If the ranking $D_1 > D_2$ holds across settings, the results of the same set of demonstrations in different settings should show high and significant Spearman’s correlation. However, the correlations in Table 5 and 6 are weak and often insignificant (more results are given in Table 15, 16, and 17), even for the same language pairs in different directions (Reversed) and for similar domains (Wiki ⇒ WMT). This suggests that we will need setting-specific demonstration to get the optimal translation quality.

**Using out-of-setting demonstrations can benefit translation.** However, we can still gain from using out-of-setting demonstrations as shown by the positive gains in Table 5 and 6, where we find that transfer in target-shared and reversed settings is relatively easier, and that transfer across distant domains can be successful particularly when in-setting example pool is of low quality. This is also supported by the transfer to document-level translation, where both BLEU and document-specific evaluation get improved as shown in Table 7. Results in Table 19 show that the transfer is
We also observe a phenomenon specific to prompting: prompt trap where prompting behaves unpredictable when its input is mixed with prompt template phrases. In the second case in Table 8, the model copies the template phrases, rather than translating them into Chinese. This means that translating prompt itself (not just the input) becomes non-trivial, and that users may attack prompting-based translation systems by manipulating the input format.

We find that the translation quality between German and Chinese is very poor (see Table 13). We argue that the cross-lingual ability of GLM-130B mainly centers around English (although GLM-130B was pretrained on Chinese as well), and thus explore pivoting translation instead. Table 9 shows that pivoting through English greatly improves non-English translation. It’s still unclear whether the current LLM pretraining recipe could achieve promising non-English-centric cross-lingual ability. We might need to consider adding parallel data into the LLM pretraining or finetuning.

### 7. Related Work

The capability of prompting heavily depends on its surface representation, where small modifications to the prompt could cause high variance in its performance. This inspires researchers to develop advanced prompting strategies to get the most from LLMs. Gao et al. (2021) proposed to generate prompt templates automatically using T5 (Xue et al., 2021) rather than adopting manual templates. Liu et al. (2022) reported selecting prompt examples close to the test input via a kNN-based retriever, Sorensen et al. (2022) resorted to an information-theoretic approach based on mutual information, while Zhang et al. (2022b) formulated example selection as a sequential decision problem and solved it by reinforcement learning. For reasoning tasks, Wei et al. (2022c) developed chain-of-thought (CoT) prompting letting the model output the intermediate reasoning steps, which inspires researchers to further explore CoT selection (Fu et al., 2022) and decomposition (Zhou et al., 2022). In contrast to the studies just mentioned, which focus on NLP tasks other than MT, we explore prompting strategies exclusively for translation.

Prompting uses instructions to guide LLMs, which is closely...

### Table 8: Case study of translation errors by prompting. Top: copying (in red), mistranslation of date (in blue), misunderstanding of source (wave lines); Bottom: prompt trap where the model fails to translate the prompt phrase (in bold).

<table>
<thead>
<tr>
<th>Setting</th>
<th>0-shot</th>
<th>1-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De→Zh</td>
<td>Zh→De</td>
</tr>
<tr>
<td>Direct</td>
<td>2.80</td>
<td>10.05</td>
</tr>
<tr>
<td>Pivoting</td>
<td><strong>19.23</strong></td>
<td><strong>19.53</strong></td>
</tr>
</tbody>
</table>

Table 9: COMET scores for direct vs. pivoting translation for De↔Zh on Wiki Full sets. In 1-shot prompting, we randomly sample 3 demonstrations and report average performance. **Pivoting:** source → English → target.
related to neural MT with special prefixes. In multilingual NMT, a target language tag is often appended to the source input to indicate the translation direction (Johnson et al., 2017; Arivazhagan et al., 2019; Zhang et al., 2020). Special attribute tags can also be used to control properties of the model output, such as politeness (Sennrich et al., 2016a), diversity (Shu et al., 2019), and quality (Caswell et al., 2019).

Besides, retrieved phrases and sentences can be augmented to the input to improve translation quality (Zhang et al., 2018; Gu et al., 2018). With the popularity of prompting LLMs, researchers see value in incorporating prompts into neural MT (Li et al., 2022; Tan et al., 2021; Garcia & Firat, 2022). Still, these methods rely on pretraining or finetuning the model rather than prompting frozen LLMs.

Very recently, concurrent to our work, Vilar et al. (2022) examined the capability of prompting PaLM for translation and discovered that prompting with high-quality examples even chosen randomly performs on par with or better than the one using input-relevant examples. By contrast, Agrawal et al. (2022) explored strategies to select input-specific examples, and observed that input-relevant examples based on n-gram overlap significantly improves the capability of prompts. Our study resonates with both their findings and also explains their conflict: while the quality and input-based semantic similarity correlate with prompting performance significantly, the correlation strength is unfortunately not strong enough so using them as indicators to select examples may produce mixed results. Note that apart from example selection, we also studied using monolingual data and transfer learning for MT prompting.

8. Conclusion and Future Work

In this paper, we presented a systematic study on prompting for MT, exploring topics ranging from prompting strategy, the use of unlabelled monolingual data, to transfer learning. We found that prompt template and demonstration example selection both have substantial impact on translation. Some prompt example features correlate significantly with prompting performance; treating them as criteria for example selection benefits translation to some extent but not consistently as the correlations are not strong enough.

Prompting for MT requires retaining the source-target mapping signals in the demonstration. Directly applying monolingual data for prompting sounds interesting but doesn’t work. Constructing pseudo parallel prompt examples by back-/forward-translation via zero-shot prompting is a simple yet effective solution. Regarding transfer learning, we saw positive results when applying a (sentence-level) demonstration to other domains, other language pairs or document-level translation. Unfortunately, the optimality of the demonstration doesn’t generalize across settings and the transfer performance is also unstable. We argue that MT provides a set of unique challenges and call for more efforts on evaluating prompting LLMs for MT.

Prompting also faces a number of other issues, like off-target generation and prompt traps, which we plan to address in the future. We acknowledge that our study heavily depends on the INT-4 quantized GLM-130B, which, unlike GPT and PaLM, was pretrained with both bidirectional and unidirectional training objectives. The quantization might weaken the model’s capability and deteriorate some unknown aspects. We thus are interested in examining whether our findings can generalize to other LLMs, like GPT-3, OPT and PaLM. We would also like to explore further how to improve the cross-lingual ability in LLMs. Finally, while our study focuses on prompting, how to finetune LLMs for MT and when/whether finetuning is preferred over prompting are yet to be investigated.

Acknowledgments

We thank the reviewers for their insightful comments. This work was funded by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee [grant number 10039436 – UTTER]. The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References


Promoting Large Language Model for Machine Translation: A Case Study


Li, Y., Yin, Y., Li, J., and Zhang, Y. Prompt-driven neural machine translation. In Findings of the Association for


Prompting Large Language Model for Machine Translation: A Case Study


A. Appendix
Prompting Large Language Model for Machine Translation: A Case Study

Figure 5: COMET (top) and BLEU (bottom) scores for few-shot prompting as a function of the number of prompt examples ($K = 1, 5, 10, 20$) on Wiki Ablation sets. For each setup, we randomly sample 100 times from the example pool and show the performance distribution via box plots. Dashed red line denotes the zero-shot baseline; blue curve and shadow area denote the mean and standard deviation.

Figure 6: Scatter plotting between BLEU and LMScore for 1-shot prompting on Wiki De$\leftrightarrow$En, En$\leftrightarrow$Zh Ablation sets.
Table 10: Detailed zero-shot results for prompting with different templates and different template languages on Wiki Ablation sets. Template (A) in English achieves the overall best performance measured by BLEU and COMET. Avg: average result over different language pairs. Best results in each section are underlined; best results in each column are in bold.
Table 11: Detailed Spearman’s ρ between demonstration features and their prompting performance (COMET and BLEU) for 1-shot prompting on Wiki Ablation sets. We randomly sample 600 demonstrations from each pool to calculate the correlation. High-quality examples are from the default selection pool while Low-quality examples are from WikiMatrix.v1. †/‡: statistically significant at p < 0.05/0.01. Gray cells indicate insignificance; Red cells indicate ρ > 0.5.

Table 12: Top-ranked parallel examples according to SemScore on WikiMatrix.v1 En-De and En-Zh. Despite showing high semantic similarity, these examples are not very informative. We thus dropped them at selection.
### Table 13: Detailed test results for zero-shot and few-shot prompting on Wiki Full sets with different selection strategies.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>COMET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De ↔ En</td>
<td>En ↔ Zh</td>
</tr>
<tr>
<td>NLLB-200 (54.5B)*</td>
<td>45.80 39.60 25.90 20.60 31.20 31.90 32.50</td>
<td>75.43 66.37 26.22 54.01 32.97 69.82 54.14</td>
</tr>
<tr>
<td>Zero-Shot</td>
<td>37.80 20.50 21.70 9.60 28.60 26.30 24.08</td>
<td>69.04 36.06 48.79 14.63 60.54 66.98 49.34</td>
</tr>
</tbody>
</table>

#### 1-Shot Translation (high-quality pool)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>COMET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De ↔ En</td>
<td>En ↔ Zh</td>
</tr>
<tr>
<td>Random</td>
<td>37.67 21.23 28.70 9.07 34.87 26.30 26.31</td>
<td>67.77 35.56 47.23 11.75 60.69 65.75 48.29</td>
</tr>
<tr>
<td>SemScore</td>
<td>38.40 21.37 29.17 9.47 35.50 26.50 26.73</td>
<td>68.55 35.49 43.54 13.14 59.84 66.98 47.92</td>
</tr>
<tr>
<td>LMScore</td>
<td>37.80 21.43 28.13 9.40 35.40 26.73 26.48</td>
<td>67.79 37.00 45.66 13.63 61.87 66.45 48.73</td>
</tr>
</tbody>
</table>

#### 5-Shot Translation (high-quality pool)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>COMET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De ↔ En</td>
<td>En ↔ Zh</td>
</tr>
<tr>
<td>Random</td>
<td>39.03 22.00 29.37 10.07 37.07 26.30 26.31</td>
<td>70.30 36.46 51.77 16.74 63.77 67.62 51.11</td>
</tr>
<tr>
<td>SemScore</td>
<td>38.13 21.93 30.50 10.20 36.87 26.30 26.31</td>
<td>70.12 38.40 52.29 16.88 64.40 67.85 51.66</td>
</tr>
<tr>
<td>LMScore</td>
<td>38.87 22.03 30.20 9.97 35.83 26.13 26.77</td>
<td>69.74 37.01 51.01 16.63 61.74 67.74 50.65</td>
</tr>
<tr>
<td>TLength</td>
<td>38.57 22.00 29.50 10.00 35.90 26.53 27.08</td>
<td>68.94 37.16 50.80 15.80 63.01 67.29 50.50</td>
</tr>
</tbody>
</table>

#### 1-shot Translation (Low-quality Pool)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>COMET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De ↔ En</td>
<td>En ↔ Zh</td>
</tr>
<tr>
<td>Random</td>
<td>36.73 20.53 22.23 8.23 34.63 26.13 24.75</td>
<td>66.82 34.15 10.11 -1.94 57.97 66.08 38.86</td>
</tr>
<tr>
<td>Ours</td>
<td>37.90 21.27 20.50 9.37 34.47 26.17 24.94</td>
<td>68.46 33.78 0.19 12.07 58.05 66.75 39.88</td>
</tr>
</tbody>
</table>

Table 14: Detailed test results on WMT Full sets. a, b, c: results from Zeng et al. (2021), Qian et al. (2021), and Wang et al. (2021b), respectively.
Figure 8: Results for few-shot prompting with monolingual data on Wiki Ablation sets for De→Zh.

Figure 9: BLEU scores for few-shot prompting with monolingual data on Wiki Ablation sets.

Table 15: Detailed Spearman’s ρ for cross-lingual transfer under 1-shot prompting on Wiki Ablation sets. Gray cells indicate insignificance.
Table 16: Detailed translation results (relative against the zero-shot baseline) for cross-lingual transfer under 1-shot prompting on Wiki Ablation sets. Blue cells indicate positive gains.

Table 17: Spearman’s $\rho$ and relative performance (in BLEU) for cross-domain transfer under 1-shot prompting.

Table 18: BLEU scores for direct vs. pivoting translation for De$\leftrightarrow$Zh on Wiki Full sets.

Table 19: Cross-lingual and cross-domain transfer results on Multi-Domain Full sets under 1-shot prompting. For cross-domain transfer, we adopt the SemScore-based strategy for example selection using the default Wiki/WMT Full candidate pool; for cross-lingual transfer, we extend the selected examples in Multi-Domain 1-shot translation (low-quality pool) by translating the English sentences to French and Chinese using Google Translate. Results are averaged over 3 different demonstrations.