Prompting Large Language Model for Machine Translation: A Case Study

Biao Zhang **Barry Haddow** **Alexandra Birch**

School of Informatics, University of Edinburgh
b.zhang@ed.ac.uk, bhaddow@inf.ed.ac.uk, a.birch@ed.ac.uk

Abstract

Research on prompting has shown 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, prompting constructs task-specific prompts by rephrasing test examples with descriptive task instructions and executes 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; 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; 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 and Zong, 2016) via zero-shot prompting. In addition, neural MT and pretrained LLMs 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 hardly 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 $\mathcal{L}$, MT prompting first converts each test input X to a prompt according to a template $\mathcal{T}$ and then generate the translation Y by feeding the prompt to $\mathcal{L}$. 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*) $\mathcal{D}^P = \{X'_i, Y'_i\}_{i=1}^K$ 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:

$$\texttt{[src]:}\ X\ \texttt{[tgt]:} \tag{1}$$

where [src] and [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:

$$\texttt{[psrc]:}\ X'_1\ \texttt{[ptgt]:}\ Y'_1 \ldots\ \texttt{[psrc]:}\ X'_K\ \texttt{[ptgt]:}\ Y'_K\ \texttt{[src]:}\ X\ \texttt{[tgt]:} \tag{2}$$

where [psrc] and [ptgt] 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 "中文：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 finetuning. 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 and A100-40G GPUs.

We work on three languages: English (En), German (De), and Chinese (Zh). We perform major

ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	`[src]: [input]` ◇ `[tgt]:`	**38.78**	**31.17**	-26.15	-16.48	**14.82**	**-1.08**
B	`[input]` ◇ `[tgt]:`	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	`[input]` ◇ Translate to `[tgt]:`	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	`[input]` ◇ Translate from `[src]` to `[tgt]:`	-113.80	-89.16	-153.80	-130.65	-76.79	-67.71
E	`[src]: [input]` ◇ Translate to `[tgt]:`	20.81	16.69	**-24.33**	**-5.68**	-8.61	-30.38
F	`[src]: [input]` ◇ Translate from `[src]` to `[tgt]:`	-27.14	-6.88	-34.36	-9.22	-32.22	-44.95

Table 1: 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**.

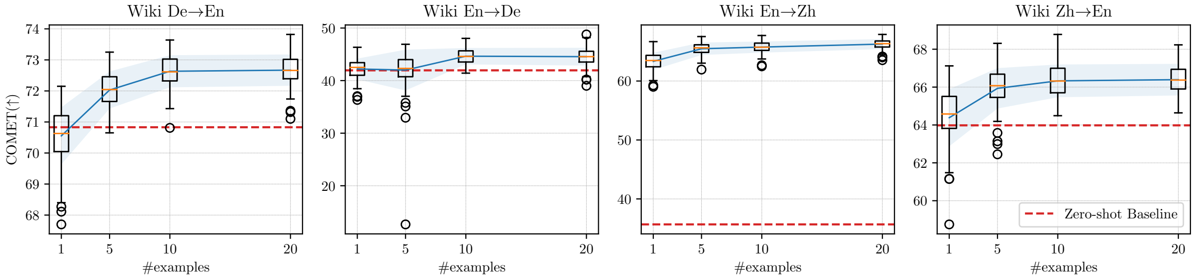


Figure 1: COMET scores for *few-shot* prompting as a function of the number of prompt examples ($K = 1, 5, 10, 20$) on Wiki Ablation sets. For each setup, we randomly sample 100 times from the example pool and show the performance distribution via box plots. Dashed red line denotes the zero-shot baseline;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

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 and 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 9, Appendix.

We evaluate translation performance using both a surface-based metric, detokenized BLEU↑ from SacreBLEU (Post, 2018), and a model-based metric, COMET↑ from `unbabel-comet` with *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 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 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 1 shows the results (we list detailed results in Table 10, Appendix). The template affects zero-shot quality substantially, and the simple template Ⓐ in English specifying just the source and target language name achieves the best overall results. In follow-up experiments, we thus focus on template Ⓐ.

Feature	BLEU		COMET	
	HQ	+ LQ	HQ	+ LQ
SLength	0.21	0.31	0.14	0.26
TLength	**0.23**	0.32	**0.17**	0.29
LMScore	0.20	**0.33**	0.14	**0.31**
MTScore	0.04	0.14	0.11	0.19
SemScore	0.19	0.30	0.16	0.30
CaseSemScore-Src	0.14	0.29	0.11	0.28
CaseSemScore-Tgt	0.14	0.30	0.14	**0.31**

Table 2: 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.

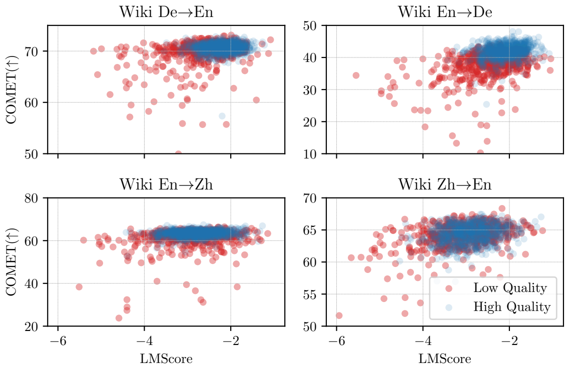


Figure 2: Visualization between COMET and LMScore for *1-shot* prompting on Wiki Ablation sets. While correlations are significant, data points are scattered like clouds.

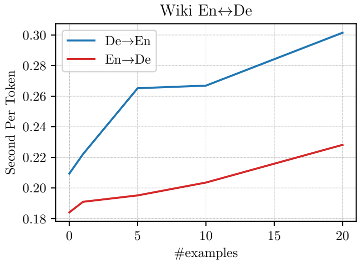


Figure 3: Inference time per token in seconds for *zero-/few-shot* prompting on Wiki En-De Ablation sets. Numbers are averaged over 3 runs with 3 distinct demonstrations on 4 A100-40G GPUs.

Language-specific template delivers mixed results. Table 1 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 performance 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 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.

Using more prompt examples for demonstration improves translation significantly on average. We next study few-shot prompting following the template Ⓐ 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 3.

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, where prompt examples help (the reference text is always simplified Chinese).

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 demonstration 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 leave few-shot analysis to future). Particularly, we extract and analyze 7 features of a demonstration:

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

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
1-Shot Translation (high-quality pool)				
Random	26.31	48.29	21.27	30.70
SemScore	26.73	49.34	21.82	31.28
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
5-Shot Translation (high-quality pool)				
Random	**27.46**	51.11	21.82	33.87
SemScore	27.36	**51.66**	**22.37**	34.30
LMScore	27.17	50.65	22.04	**35.19**
TLength	27.08	50.50	21.75	34.29
1-shot Translation (Low-quality Pool)				
Random	24.75	38.86	22.06	30.70
Ours	24.94	39.88	22.23	30.87

Table 3: BLEU and COMET scores for *zero-shot and few-shot* prompting on Wiki and WMT Full sets with different selection strategies. *Ours*: the proposed combined strategy; *Random*: random sampling; *SemScore, LMScore* and *TLength* denote selecting top-ranked examples based on the corresponding feature values. We select 3 demonstrations for each translation direction and report average performance; the final score is further averaged over different language pairs. Underlined results denote the best in each section, while **Bold** results are the overall best.

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

MTScore translation quality of the prompt example from COMET QE model *wmt20-comet-qe-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 2 summarizes the results and Figure 2 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.

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.

Using prompt examples selected based on the proposed features yields improved performance. We next verify the above findings on the Full sets. We explore selection strategies based on SemScore, LMScore and TLength (i.e. use top-ranked examples) as they show high average correlation. We didn't analyze CaseSemScore-Tgt as it's more complicated and doesn't make significant difference. Note we excluded too long (more than 100 tokens) or too short (less than 10 tokens) examples during selection. We also consider 5-shot prompting, where we concatenate top-ranked 5 examples in an ascending order (Liu et al., 2022).

Table 3 shows that, with high-quality pool, adopting the feature-based strategy is likely to outperform the random baseline, and the SemScore-based strategy performs well across different settings (detailed results are available in Table 13 and 14, Appendix). These strategies also generalize to 5-shot prompting to some extent. For selection from low-quality pool, we propose a combined strategy: we

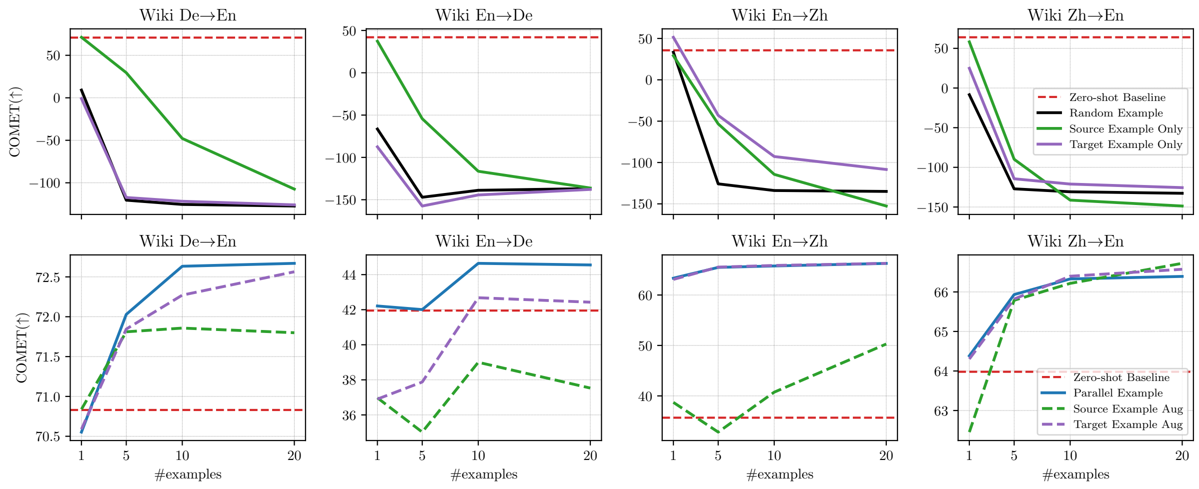


Figure 4: COMET scores for *few-shot* prompting with monolingual data on Wiki Ablation sets. *Random Example*: random sentence pairs; *Source/Target Example Only*: only use source or target data for prompting; *Source/Target Example Aug*: use pseudo-parallel data instead constructed via zero-shot prompting. For each setup, we randomly sample 50 demonstrations and report average performance.

first choose top-11K examples according to SemScore to filter out poor examples, the top-1K of which are also dropped as they tend to be uninformative (see Table 12 in Appendix); then we re-rank the rest with LMScore and retain top-1K examples, upon which we further apply the TLength-based strategy. In Table 3, 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 of 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 and 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. 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 real parallel examples.

5 Transfer Learning for Prompting

After obtaining a performant demonstration, we are interested in to what extent its capability

Setting	Correlation		Δ Quality	
	BLEU	COMET	BLEU	COMET
Source Shared	0.08	0.10	+0.59	+7.03
Target Shared	0.20	0.24	+1.32	+9.67
Reversed	0.15	0.06	+1.41	+11.56

Table 4: Spearman's ρ and relative performance for cross-lingual transfer under *1-shot* prompting on Wiki Ablation sets (among En, De and Zh). When studying transfer from language pair S_1 to S_2, we randomly sample 300 demonstrations from the default pool of S_1, and then evaluate them on the Ablation test sets for S_1 and S_2 respectively, based on which we compute the correlation. The performance is also averaged. Δ *Quality*: relative quality against the zero-shot baseline. Blue cells indicate positive gains. *Source/Target Shared*: average result for transfer settings where the source/target language is shared; *Reversed*: average result for the same language pair but in different directions.

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., 2021), 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 these questions through 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. Unfortunately, the correlations in Table 4 and 5 are very 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 demonstrated by the positive gains in Table 4 and 5, 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 6. Results in Table 19 show that the transfer is unstable and could deliver negative results, i.e. worse than zero-shot prompting, partially resonating with previous findings (Lin et al., 2021). We leave the question of how to select prompt examples in transfer learning setups to future.

Transfer from Wiki to ⇒		WMT	IT	Medical
Correlation	En→De	0.09	0.14	$0.27^{\ddagger}$
	De→En	$0.23^{\ddagger}$	$0.20^{\ddagger}$	0.13
Δ Quality	En→De	+4.00	+19.52	+7.80
	De→En	+0.10	+19.46	+1.24

Table 5: Spearman's ρ and relative performance (in COMET) for cross-domain transfer under *1-shot* prompting. We explore transfer from Wiki to Multi-Domain using the Ablation sets. Correlation and performance are calculated in the same way as in cross-lingual transfer, except that we sample 200 demonstrations. ‡: statistically significant at $p < 0.01$; Gray cells indicate insignificance.

Method	d-BLEU	TC	CP	PT	TCP
Zero-Shot	30.2	47.5	**38.7**	41.6	42.4
SemScore	**30.5**	**53.0**	34.4	**43.2**	42.9
LMScore	**30.5**	**53.0**	36.8	42.9	**43.7**

Table 6: Results for transfer learning from sentence-level demonstration to document-level translation under *1-shot* prompting on PDC Zh→En Full sets. We split each test document in PDC into non-overlapped chunks, each of which contains about 4 sentences. *SemScore/LMScore*: prompt example selection strategy; we apply them to PDC's default pool. We select 3 demonstrations and report average performance. *d-BLEU*: document-level BLEU; *TC/CP/PT/TCP*(↑): document-specific metrics proposed in (Sun et al., 2022).

6 Discussion

Although prompting enables translation with decent performance, it still suffers from many (well-known) problems. Here, we briefly explain the problems we observed from the model's outputs.

Prompting sometimes rejects translating the input. Instead, it emits either empty or off-target outputs, i.e. translating in a wrong target language. This occurs frequently when translating into Chinese, where the model often translates into traditional Chinese with messy codes, causing unstable performance. Besides overly relying on a language

Source	根据三江源国家公园管理局长江源园区可可西里管理处统计，藏羚羊回迁数量总体呈逐年上升态势，2019年藏羚羊回迁数量为4860只，比2018年增加338只。
Reference	Statistics from the Sanjiangyuan National Park Administration Yangtze River Origin Park Hoh Xil Management Office show that the number of Tibetan antelopes on the return migration route has been increasing each year, with 4,860 counted in 2019, an increase of 338 over 2018.
GLM-130B (1-shot)	According to the三江源国家公园管理局长江源园区可可西里管理处, the total number of re-migration of the Tibetan antelope has been on the rise since 2018, with 4,860 re-migrating in 2109, an increase of 338 compared to 2808.
Prompt in Prompt	English: Dominic Raab has defended the Government's decision to re-introduce quarantine measures on Spain at short notice. **Translate from English to Chinese:** Chinese:
Reference	针对政府突然做出重新对西班牙实施隔离措施的决定，Dominic Raab 做出了辩解。从英文翻译成中文：
GLM-130B (zero-shot)	多米尼克·拉布(Dominic Raab)对政府决定重新引入西班牙的检疫措施表示支持。**Translate from English to Chinese:**

Table 7: 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**).

Setting	0-shot		1-shot	
	De→Zh	Zh→De	De→Zh	Zh→De
Direct	2.80	10.05	47.23	11.75
Pivoting	**19.23**	**19.53**	**48.25**	**25.31**

Table 8: 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.

model, prompting tends to under-translate the input, copy source phrases, produce code-switched output, mistranslate entities (e.g. dates) and generate hallucination, as illustrated in Table 7.

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 7, 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 8 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 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 and 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, which, to the best of our knowledge, have never been explored before.

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 are also 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 LLM.

Limitations

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. It's unclear how our findings generalize to other pretrained LLMs. In addition, we mainly work on three languages due to resource constraints, and in experiments, results vary greatly across language pairs. Increasing the coverage of experimental languages would make the results more reliable.

Acknowledgments

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

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. arXiv preprint arXiv:2212.02437.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7747–7763, Online. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In Proceedings of the Sixth Conference on Machine Translation, pages 1–88, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 53–63, Florence, Italy. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. arXiv preprint arXiv:2210.00720.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830, Online. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. arXiv preprint arXiv:2202.11822.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. arXiv preprint arXiv:2209.12356.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search engine guided neural machine translation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. arXiv preprint arXiv:2205.12654.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5(0):339–351.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022. Prompt-driven neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2579–2590, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. arXiv preprint arXiv:2112.10668.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? arXiv preprint arXiv:2202.12837.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics. arXiv preprint arXiv:2212.10297.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2021. Msp: Multi-stage prompting for making pre-trained language models better translators. arXiv preprint arXiv:2110.06609.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. arXiv preprint arXiv:2211.09102.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021. TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2792–2802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1628–1639, Online. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. arXiv preprint arXiv:2211.04486.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 12697–12706. PMLR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625.

A Appendix

Dataset	Language(s)	Test Set	Selection Pool (Default)	Source (#sample)
Wiki	English	100	897	FLORES `eng_Latn.dev` (997)
	German	100	897	FLORES `deu_Latn.dev` (997)
	Chinese	100	897	FLORES `zho_Hans.dev` (997)
WMT	English-German	100	2900	newstest2013 (3000)
IT	German-English	100	1900	Multi-Domain Dev Set (2000)
Medical	German-English	100	1900	Multi-Domain Dev Set (2000)

(a) Ablation Sets

Dataset	Languages	Source	Test Set	High-quality Pool (Default)	Low-quality Pool
Wiki	English	FLORES	`eng_Latn.devtest` (1012)	`eng_Latn.dev` (997)	En-Zh* (0.79M)
	German	FLORES	`deu_Latn.devtest` (1012)	`deu_Latn.dev` (997)	De-En* (1.57M)
	Chinese	FLORES	`zho_Hans.devtest` (1012)	`zho_Hans.dev` (997)	De-Zh* (0.13M)
WMT	English-German	WMT	newstest2021 (1002/1000)	newstest2020 (1418)	
	English-Chinese	WMT	newstest2021 (1002/1948)	newstest2020 (1418)	
IT	German-English	Multi-Domain	Test Set (2000)	-	Train Set (0.22M)
Law	German-English	Multi-Domain	Test Set (2000)	-	Train Set (0.47M)
Medical	German-English	Multi-Domain	Test Set (2000)	-	Train Set (0.25M)
PDC	Chinese-English	News	Test Set (4858/148 Docs)	Dev Set (2881)	-

(b) Full Sets

Table 9: Statistics of Ablation sets and Full sets. Numbers in brackets denote the number of instances. *: data from WikiMatrix.v1 (Schwenk et al., 2021).

ID	BLEU							COMET						
	De ↔ En		De ↔ Zh		En ↔ Zh		Avg	De ↔ En		De ↔ Zh		En ↔ Zh		Avg
	→	←	→	←	→	←		→	←	→	←	→	←	
English Template ***Without*** *Line Break*														
A	**38.00**	**23.10**	23.30	**12.10**	31.50	**27.90**	**25.98**	**70.83**	**41.95**	4.34	**15.92**	35.68	**63.98**	**38.78**
B	8.30	9.00	2.80	2.40	6.60	8.20	6.22	-45.75	-70.27	-140.43	-119.82	-112.38	-43.10	-88.62
C	30.60	2.10	5.50	1.10	1.10	8.30	8.12	29.78	-142.36	-117.20	-117.14	-120.57	-58.32	-87.63
D	26.10	0.00	5.10	0.00	0.20	0.60	5.33	-1.20	-160.59	-124.15	-157.62	-130.51	-108.71	-113.80
E	35.90	18.20	26.10	9.60	16.00	22.30	21.35	68.06	5.41	27.53	-6.46	-5.58	35.93	20.81
F	33.50	5.60	25.10	0.80	0.20	9.10	12.38	61.09	-62.31	22.71	-112.79	-50.84	-20.71	-27.14
English Template ***With*** *Line Break*														
A	36.60	21.80	25.10	11.40	26.90	26.90	24.78	67.97	37.41	7.24	9.46	4.89	60.08	31.17
B	7.70	7.70	5.00	2.70	13.20	10.00	7.72	-85.97	-81.79	-126.58	-113.27	-55.64	-48.82	-85.35
C	28.00	4.40	7.70	0.70	13.30	14.50	11.43	36.10	-99.01	-118.99	-133.39	-74.19	-23.00	-68.75
D	25.20	1.60	4.20	0.10	4.90	5.40	6.90	13.96	-121.58	-125.36	-148.29	-78.78	-74.91	-89.16
E	35.70	20.00	24.40	3.90	28.30	20.30	22.10	66.08	22.21	15.62	-55.41	13.36	38.30	16.69
F	33.60	9.30	23.60	3.00	6.70	17.90	15.68	57.46	-45.84	14.73	-69.69	-30.63	32.68	-6.88
German Template ***Without*** *Line Break*														
A	20.00	15.70	1.60	3.10	0.70	7.10	8.03	23.09	4.61	-70.84	-47.51	-65.61	-0.66	-26.15
B	5.60	2.10	0.10	1.60	0.20	1.10	1.78	-82.99	-152.26	-174.72	-132.06	-162.79	-110.99	-135.97
C	4.60	5.40	0.30	3.70	0.00	4.10	3.02	-57.63	-108.36	-120.99	-125.18	-135.21	-90.42	-106.30
D	3.50	0.10	0.00	0.00	0.00	0.10	0.62	-115.55	-168.13	-166.07	-169.21	-161.27	-142.57	-153.80
E	17.30	19.00	0.20	8.50	2.30	19.60	11.15	14.19	6.47	-100.92	-25.14	-50.42	9.85	-24.33
F	6.30	4.80	0.20	7.30	0.10	11.70	5.07	3.88	-65.86	-44.76	-27.91	-60.31	-11.22	-34.36
German Template ***With*** *Line Break*														
A	25.40	20.20	6.40	3.50	8.00	9.20	12.12	38.47	31.45	-80.14	-47.22	-50.26	8.84	-16.48
B	15.60	7.80	2.60	1.00	0.50	0.80	4.72	-20.65	-81.28	-125.21	-137.02	-125.31	-108.45	-99.65
C	15.40	5.70	5.70	3.00	6.00	6.70	7.08	-23.46	-80.15	-86.27	-104.10	-87.18	-58.23	-73.23
D	2.80	0.50	0.00	0.00	0.10	1.10	0.75	-95.30	-154.76	-140.51	-155.91	-137.36	-100.08	-130.65
E	24.70	19.50	10.40	8.50	11.10	17.20	15.23	35.12	3.95	-62.48	-18.32	-27.61	35.26	-5.68
F	7.60	17.20	0.50	8.60	3.90	11.30	8.18	13.01	9.10	-43.63	-10.88	-46.46	23.54	-9.22
Chinese Template ***Without*** *Line Break*														
A	37.60	15.50	**28.30**	2.10	33.40	15.10	22.00	67.41	-5.40	**45.24**	-74.78	53.71	2.72	14.82
B	23.60	6.30	14.50	0.50	19.30	1.90	11.02	-6.41	-90.63	-12.10	-159.66	-9.24	-121.29	-66.55
C	11.40	3.20	14.30	0.40	20.80	5.00	9.18	-32.55	-114.57	-9.91	-140.54	2.89	-85.58	-63.38
D	17.10	6.40	15.90	0.20	19.60	1.90	10.18	-34.15	-101.69	-24.36	-166.15	-9.20	-125.20	-76.79
E	29.00	8.00	27.00	0.40	**34.90**	16.10	19.23	35.55	-63.09	37.06	-119.13	**54.14**	3.80	-8.61
F	31.70	3.70	24.80	0.10	27.20	11.80	16.55	35.65	-105.74	22.97	-129.71	5.61	-34.09	-34.22
Chinese Template ***With*** *Line Break*														
A	26.80	14.70	24.70	3.30	33.80	22.90	21.03	24.46	-84.74	24.76	-64.07	52.65	40.45	-1.08
B	23.70	6.30	11.90	0.10	14.40	0.60	9.50	-11.65	-102.50	-63.95	-161.96	-46.84	-128.12	-85.84
C	12.10	3.00	13.80	0.80	21.20	9.90	10.13	-36.39	-105.55	-42.16	-151.06	-15.41	-74.90	-70.91
D	14.10	3.20	15.10	0.20	20.00	2.50	9.18	-19.15	-106.69	-19.34	-154.73	-11.51	-94.82	-67.71
E	28.60	8.00	26.50	0.90	32.30	21.40	19.62	8.71	-118.14	15.34	-124.30	21.18	14.91	-30.38
F	26.90	3.40	26.10	0.20	25.80	16.00	16.40	11.58	-120.31	10.33	-129.61	-21.19	-20.52	-44.95

Table 10: Detailed *zero-shot* results for prompting with different templates and different template languages on Wiki Ablation sets. Template Ⓐ 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**.

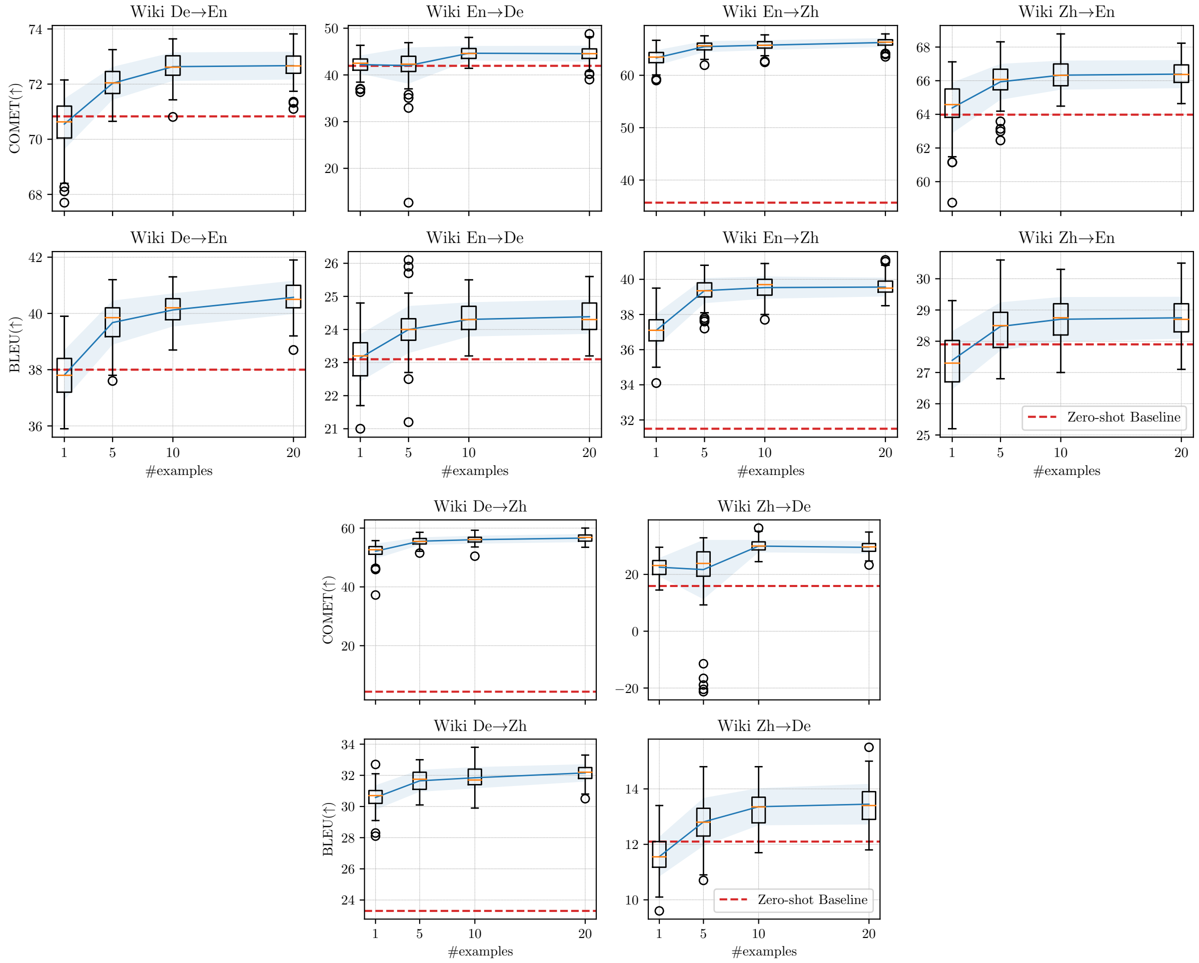


Figure 5: COMET (top) and BLEU (bottom) scores for *few-shot* prompting as a function of the number of prompt examples ($K = 1, 5, 10, 20$) on Wiki Ablation sets. For each setup, we randomly sample 100 times from the example pool and show the performance distribution via box plots. Dashed red line denotes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

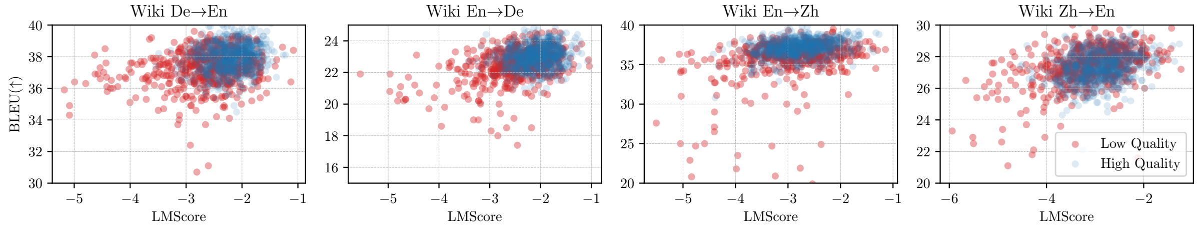
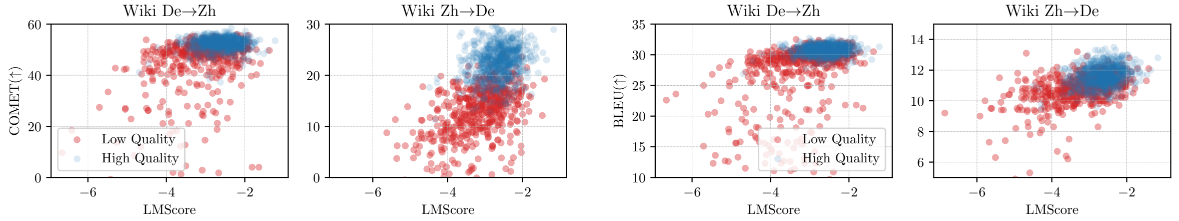


Figure 6: Scatter plotting between BLEU and LMScore for *1-shot* prompting on Wiki De↔En, En↔Zh Ablation sets.



(a) COMET vs. LMScore

(b) BLEU vs. LMScore

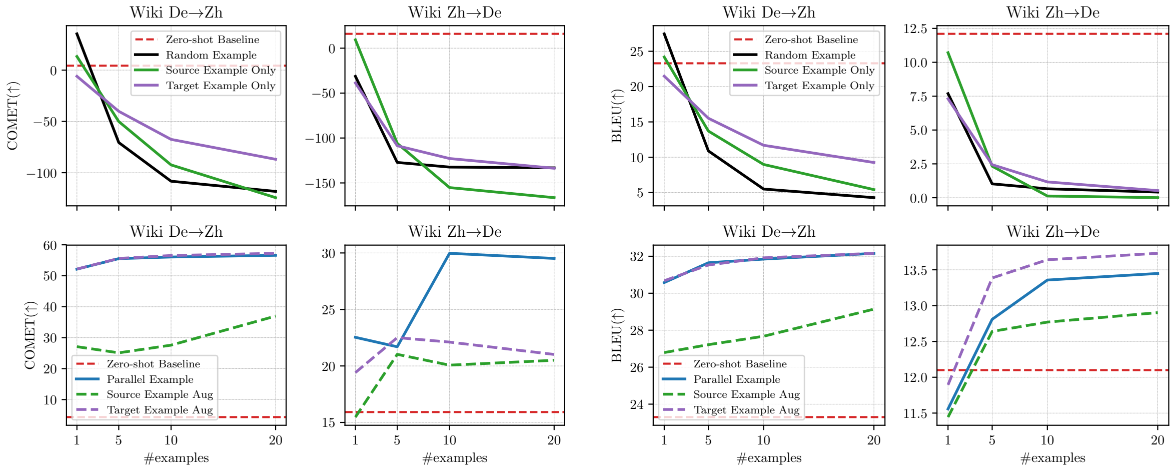
Figure 7: Scatter plotting between COMET/BLEU and LMScore for *1-shot* prompting on Wiki De↔Zh Ablation sets.

Method	High-quality Examples							Plusll Low-quality Examples						
	De ↔ En		De ↔ Zh		En ↔ Zh		Avg	De ↔ En		De ↔ Zh		En ↔ Zh		Avg
	→	←	→	←	→	←		→	←	→	←	→	←	
Correlation with COMET														
SLength	0.02	0.18‡	0.24‡	0.12‡	0.26‡	0.01	0.14	0.09‡	0.20‡	0.52‡	0.44‡	0.24‡	0.10‡	0.26
TLength	-0.01	0.23‡	0.19‡	0.27‡	0.29‡	0.06	**0.17**	0.06†	0.35‡	0.41‡	0.57‡	0.25‡	0.13‡	0.29
LMScore	0.06	0.23‡	0.01	0.20‡	0.12‡	0.21‡	0.14	0.19‡	0.38‡	0.35‡	0.51‡	0.16‡	0.27‡	**0.31**
MTScore	0.01	0.05	0.11‡	0.12‡	0.06	0.28‡	0.11	0.13‡	0.04	0.30‡	0.23‡	0.18‡	0.28‡	0.19
SemScore	0.11‡	0.17‡	0.11‡	0.15‡	0.10‡	0.31‡	0.16	0.12‡	0.24‡	0.42‡	0.50‡	0.17‡	0.33‡	0.30
CaseSemScore-Src	-0.01	0.20‡	0.22‡	0.08†	0.18‡	-0.03	0.11	0.08‡	0.29‡	0.53‡	0.49‡	0.26‡	0.05	0.28
CaseSemScore-Tgt	-0.01	0.22‡	0.25‡	0.14‡	0.21‡	0.05	0.14	0.09‡	0.32‡	0.53‡	0.53‡	0.27‡	0.11‡	**0.31**
Correlation with BLEU														
SLength	0.20‡	0.27‡	0.21‡	0.11‡	0.33‡	0.12‡	0.21	0.23‡	0.30‡	0.51‡	0.35‡	0.29‡	0.18‡	0.31
TLength	0.15‡	0.32‡	0.16‡	0.22‡	0.40‡	0.12‡	**0.23**	0.15‡	0.38‡	0.41‡	0.47‡	0.33‡	0.19‡	0.32
LMScore	0.14‡	0.17‡	0.10‡	0.24‡	0.27‡	0.26‡	0.20	0.23‡	0.30‡	0.39‡	0.46‡	0.27‡	0.32‡	**0.33**
MTScore	0.03	-0.05	0.04	0.09†	0.03	0.12‡	0.04	0.11‡	-0.04	0.26‡	0.19‡	0.17‡	0.14‡	0.14
SemScore	0.13‡	0.11‡	0.15‡	0.20‡	0.25‡	0.29‡	0.19	0.13‡	0.20‡	0.45‡	0.45‡	0.28‡	0.31‡	0.30
CaseSemScore-Src	0.16‡	0.15‡	0.18‡	0.03	0.28‡	0.03	0.14	0.20‡	0.29‡	0.51‡	0.36‡	0.31‡	0.07‡	0.29
CaseSemScore-Tgt	0.14‡	0.17‡	0.16‡	0.05	0.24‡	0.09†	0.14	0.18‡	0.30‡	0.49‡	0.39‡	0.29‡	0.13‡	0.30

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. $^{\dagger}/^{\ddagger}$: statistically significant at $p < 0.05/0.01$. Gray cells indicate insignificance; Red cells indicate $\rho > 0.5$.

En→Zh	Source	Coordinates: 19°43 ′ 10″ S 63°18 ′ 00″ E / 19.71944°S 63.30000°E / -19.71944; 63.30000
	Target	坐标：19°43 ′ 10″ S 63°18 ′ 00″ E / 19.71944°S 63.30000°E / -19.71944; 63.30000
	Source	SAO 40012 is HD 277559.
	Target	SAO 40012是HD 277559。
En→De	Source	2002 and 2004.
	Target	2002 und 2004.
	Source	Brinton, Lauren and Leslie Arnovick.
	Target	Brinton, Lauren und Leslie Arnovick.

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.



(a) COMET

(b) BLEU

Figure 8: Results for *few-shot* prompting with monolingual data on Wiki Ablation sets for De↔Zh.

Method	BLEU							COMET						
	De ↔ En		De ↔ Zh		En ↔ Zh		Avg	De ↔ En		De ↔ Zh		En ↔ Zh		Avg
	→	←	→	←	→	←		→	←	→	←	→	←	
Zero-Shot	37.80	20.50	21.70	9.60	28.60	26.30	24.08	68.30	29.96	2.80	10.05	29.17	63.25	33.92
1-Shot Translation (high-quality pool)														
Random	37.67	21.23	28.70	9.07	34.87	26.30	26.31	68.77	35.56	47.23	11.75	60.69	65.75	48.29
SemScore	38.40	21.37	29.17	9.47	35.50	26.50	26.73	69.04	36.06	48.79	14.63	60.54	66.98	49.34
LMScore	37.80	21.43	28.13	9.40	35.40	26.73	26.48	68.55	35.49	43.54	13.14	59.84	66.98	47.92
TLength	37.00	21.80	28.57	9.47	35.90	26.53	26.54	67.79	37.00	45.66	13.63	61.87	66.45	48.73
5-Shot Translation (high-quality pool)														
Random	**39.03**	22.00	29.37	10.07	**37.07**	**27.20**	**27.46**	**70.30**	36.46	51.77	16.74	63.77	67.62	51.11
SemScore	38.13	21.93	**30.50**	**10.20**	36.87	26.50	27.36	70.12	**38.40**	**52.29**	**16.88**	**64.40**	**67.85**	**51.66**
LMScore	38.87	**22.03**	30.20	9.97	35.83	26.13	27.17	69.74	37.01	51.01	16.63	61.74	67.74	50.65
TLength	38.57	22.00	29.50	10.00	35.90	26.53	27.08	68.94	37.16	50.80	15.80	63.01	67.29	50.50
1-shot Translation (Low-quality Pool)														
Random	36.73	20.53	22.23	8.23	34.63	26.13	24.75	66.82	34.15	10.11	-1.94	57.97	66.08	38.86
Ours	37.90	21.27	20.50	9.37	34.47	26.17	24.94	68.46	33.78	0.19	12.07	58.05	66.75	39.88

Table 13: Detailed test results for *zero-shot and few-shot* prompting on Wiki Full sets with different selection strategies. *Ours*: the proposed combined strategy; *Random*: random sampling; *SemScore, LMScore* and *TLength* denote selecting top-ranked examples based on the corresponding feature values. We select 3 demonstrations for each setup and report the average. *Avg*: average result over language pairs. Underlined results denote the best in each section, while **Bold** results are the overall best.

Method	BLEU					COMET				
	De ↔ En		En ↔ Zh		Avg	De ↔ En		En ↔ Zh		Avg
	→	←	→	←		→	←	→	←	
Zero-Shot	**28.30**	15.70	20.70	16.80	20.38	46.01	13.32	4.63	7.92	17.97
1-Shot Translation (high-quality pool)										
Random	25.63	16.37	26.03	17.03	21.27	45.90	16.89	40.88	19.14	30.70
SemScore	26.90	16.03	26.30	18.07	21.82	46.39	15.13	41.13	22.49	31.28
LMScore	27.53	15.70	25.43	17.70	21.59	47.47	17.53	38.95	19.29	30.81
TLength	25.60	16.33	25.80	17.43	21.29	43.47	18.24	42.17	18.82	30.68
5-Shot Translation (high-quality pool)										
Random	26.40	**17.10**	26.23	17.53	21.82	48.36	20.19	43.97	22.95	33.87
SemScore	27.30	16.57	**26.93**	18.67	**22.37**	**49.33**	18.83	43.49	25.54	34.30
LMScore	25.90	16.87	26.47	18.93	22.04	47.77	**20.83**	44.76	**27.41**	**35.19**
TLength	25.80	17.03	26.55	17.63	21.75	47.34	20.78	**45.17**	23.85	34.29
1-shot Translation (Low-quality Pool)										
Random	27.33	15.53	25.30	20.07	22.06	45.29	14.21	36.83	26.49	30.70
Ours	27.63	15.97	25.23	**20.10**	22.23	47.16	15.01	34.48	26.82	30.87

Table 14: Detailed test results on WMT Full sets.

Method		BLEU						COMET					
		De ↔ En		De ↔ Zh		En ↔ Zh		De ↔ En		De ↔ Zh		En ↔ Zh	
		→	←	→	←	→	←	→	←	→	←	→	←
Prompt Language	De→En	-	0.06	0.08	0.12†	0.13†	0.13†	-	-0.02	0.09	0.12†	-0.01	0.21‡
	En→De	0.07	-	0.14‡	0.19‡	0.17‡	0.11†	0.01	-	0.07	0.21‡	0.14‡	0.17‡
	De→Zh	-0.08	0.06	-	0.14‡	0.24‡	-0.05	0.02	0.15‡	-	0.08	0.40‡	0.02
	Zh→De	0.00	0.26‡	0.26‡	-	0.05	0.01	-0.03	0.21‡	0.22‡	-	0.13†	0.15‡
	En→Zh	0.01	-0.01	0.24‡	0.25‡	-	0.19‡	0.04	-0.01	0.22‡	0.21‡	-	0.03
	Zh→En	0.15‡	-0.16‡	0.14‡	0.34‡	0.15‡	-	0.25‡	0.09	0.14‡	0.21‡	0.03	-

Table 15: Detailed Spearman's ρ for cross-lingual transfer under *1-shot* prompting on Wiki Ablation sets. Gray cells indicate insignificance.

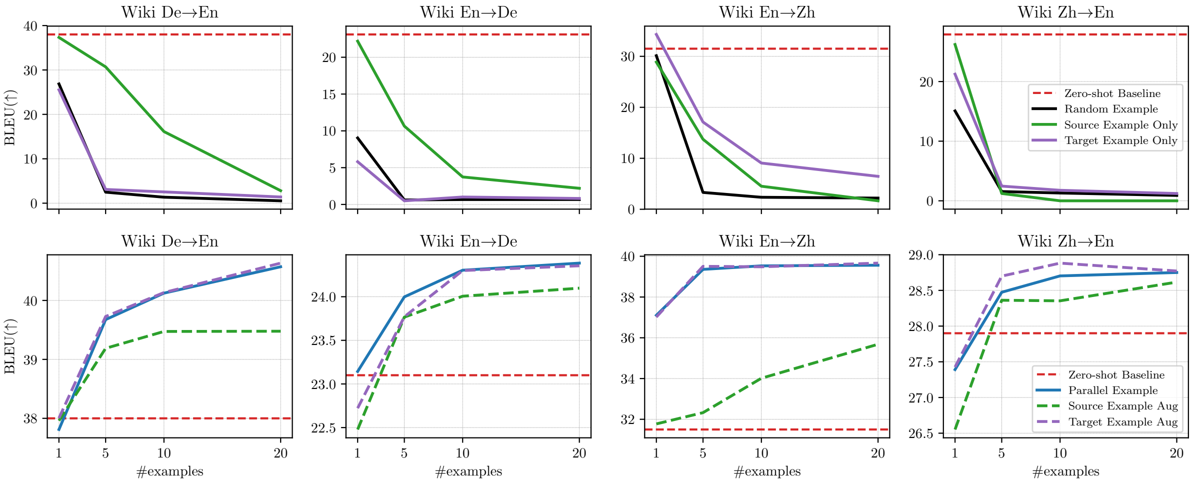


Figure 9: BLEU scores for *few-shot* prompting with monolingual data on Wiki Ablation sets.

Method		BLEU						COMET					
		De ↔ En		De ↔ Zh		En ↔ Zh		De ↔ En		De ↔ Zh		En ↔ Zh	
		→	←	→	←	→	←	→	←	→	←	→	←
Prompt Language	De→En	-	-0.32	5.02	-0.86	1.29	0.00	-	-1.08	35.04	2.71	7.00	-0.01
	En→De	-0.69	-	3.88	-0.69	1.21	-0.41	-0.46	-	26.01	1.56	6.31	-2.40
	De→Zh	-0.63	-0.48	-	-0.65	4.38	0.04	0.92	-3.68	-	4.16	23.51	-0.34
	Zh→De	-0.66	-0.86	6.84	-	3.23	0.19	0.71	-6.15	43.67	-	17.54	0.51
	En→Zh	-1.54	-1.17	6.23	-1.44	-	-1.50	-6.00	-4.47	41.77	-1.79	-	-2.20
	Zh→En	-1.12	-1.00	1.78	-1.11	4.81	-	-2.63	-3.85	15.25	3.90	25.29	-

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.

Transfer from Wiki to ⇒		WMT	IT	Medical
Correlation	En→De	0.05	0.11	0.15[†]
	De→En	-0.25[‡]	0.19[‡]	0.07
Δ Quality	En→De	-0.45	+0.88	-0.21
	De→En	-0.43	+1.00	+0.77

Table 17: Spearman's ρ and relative performance (in BLEU) for cross-domain transfer under *1-shot* prompting.

Setting	0-shot		1-shot	
	De→Zh	Zh→De	De→Zh	Zh→De
Direct	21.70	9.60	28.70	9.07
Pivoting	**24.4**	**11.5**	**29.47**	**11.47**

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

Method	BLEU				COMET			
	IT	Law	Medical	Avg	IT	Law	Medical	Avg
Zero-Shot	32.4	28.5	31.3	30.7	12.39	32.85	33.99	26.41
1-shot Translation (Low-quality Pool)								
Random	33.70	27.33	30.80	30.61	29.12	30.22	34.08	31.14
Ours	32.93	27.60	33.23	31.26	29.95	29.60	41.37	33.64
Cross-domain Transfer								
Wiki⇒Multi-Domain	32.90	26.73	31.87	30.50	25.08	33.27	37.85	32.07
WMT⇒Multi-Domain	30.87	25.37	31.43	29.22	12.98	30.34	34.80	26.04

Table 19: Cross-domain transfer results on Multi-Domain Full sets under *1-shot* prompting. We adopt the SemScore-based strategy for example selection using the default Wiki/WMT Full candidate pool. Results are averaged over 3 different demonstrations.