Prompting PaLM for Translation: Assessing Strategies and Performance

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, George Foster
Google
{vilar, freitag, colincherry, jmluo, vratnakar, fosterg}@google.com

Abstract

Large language models (LLMs) that have been trained on multilingual but not parallel text exhibit a remarkable ability to translate between languages. We probe this ability in an in-depth study of the pathways language model (PaLM), which has demonstrated the strongest machine translation (MT) performance among similarly-trained LLMs to date. We investigate various strategies for choosing translation examples for few-shot prompting, concluding that example quality is the most important factor. Using optimized prompts, we revisit previous assessments of PaLM's MT capabilities with more recent test sets, modern MT metrics, and human evaluation, and find that its performance, while impressive, still lags that of state-of-the-art supervised systems. We conclude by providing an analysis of PaLM's MT output which reveals some interesting properties and prospects for future work.

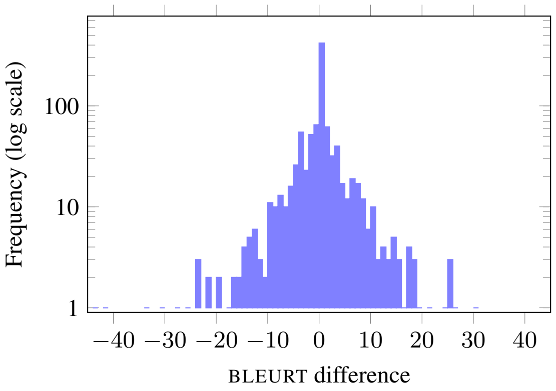

Figure 1: Histogram of the sentence-level BLEURT difference between two different 5-shot PaLM runs using the random prompt selection method from the original paper on a corpus of 1000 sentences. Each bar corresponds to a difference range of 1 BLEURT point. A majority of sentences (516) show a difference of more than 1 BLEURT point, demonstrating that the choice of prompt can strongly affect translation quality.

1 Introduction

Large language models (LLMs) trained to predict the next token from a lengthy context have demonstrated impressive machine translation capabilities, despite being trained on corpora that are overwhelmingly English, with no intentionally-included parallel text. In this paper, we carry out an in-depth investigation into the translation capabilities of LLMs, testing different prompting strategies and carefully assessing the resulting performance. We study the recently-introduced PaLM model (Chowdhery et al., 2022), a 540B-parameter decoder-only language model trained on a heavily English-centric, multilingual corpus. It has achieved the strongest MT results among LLMs trained on non-parallel multilingual corpora.

To ensure a fair assessment of PaLM's MT capability, we begin with an exploration of example selection methods for use with fixed prompt templates. We vary both the pool from which examples are chosen and the method for choosing them, comparing standard random selection to k-nearest-neighbour (kNN) selection that customizes prompts for specific inputs. Figure 1 highlights the importance of example selection by showing that two randomly-selected sets of examples can result in significantly different distributions of sentence-level BLEURT scores.

Although Chowdhery et al. (2022) report interesting results on low-resource and non-English language pairs, their most striking findings concern high-resource pairs. Accordingly, we limit our investigation to French, German, and Chinese translation to and from English. We evaluate sentence-level translation quality using recommended practices for high-quality MT, specifically: (i) we use recent WMT test sets to guard against train/test data leakage, and to facilitate comparison with state-of-the-art (SOTA) MT systems; (ii) we use a SOTA automatic metric (BLEURT) instead of BLEU which has been demonstrated to be suboptimal for high-quality translations (Kocmi et al.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15406–15427
July 9-14, 2023 ©2023 Association for Computational Linguistics

2021; Freitag et al., 2021b); and (iii) we conduct an expert-based human evaluation with detailed categories to characterize the error patterns of the automatically generated translations.

Our contributions are as follows:

- We carry out the first systematic study of LLM prompting for MT, exploring both the example candidate pool and the selection strategy. We find that the *quality* of examples matters more than the domain from which they are drawn or their lexico-semantic proximity to the current input.
- We evaluate the translation capability of LLMs with the procedure currently recommended by the MT community. We find that, although impressive, the sentence-level translation capacity of LLMs still lags behind SOTA MT.

2 Related Work

Inspired by the findings of Radford et al. (2019); Brown et al. (2020), prompting strategies for LLMs have become a topic of intense interest, generating work across a broad spectrum of methods and applications (Liu et al., 2021). A basic distinction can be made between *hard* (explicit text) prompting such as we use, and *soft* prompting that seeks to learn embeddings (Lester et al., 2021), activations (Li and Liang, 2021; Hambardzumyan et al., 2021), or attention weights (Liu et al., 2022a) that condition the model to perform a desired task. The latter approach is more expressive and more efficient at inference time, but performance can be sensitive to initialization (Hou et al., 2022), and some techniques require modifications to the model.

Hard prompts have the advantage of being easy to interpret and modify. Work in this area includes tools to facilitate development of handcrafted prompts (Strobelt et al., 2022; Bach et al., 2022); algorithms to find optimal prompts through gradient-guided search (Shin et al., 2020) or exhaustive search through labels (Schick and Schütze, 2021) or both labels and templates (Gao et al., 2021); as well as studies on the effect of example order (Kumar and Talukdar, 2021; Lu et al., 2022). Hard prompts have also been used to analyze model capabilities (Garg et al., 2022; Li et al., 2022a), the role of data (Singh et al., 2022), and the nature of prompting itself (Min et al., 2022; Wei et al., 2022).

With few exceptions, e.g. (Li et al., 2022b; Liu et al., 2022b; Valvoda et al., 2022), early approaches to hard prompting tended to condition on the task rather than the specific input. Our kNN approach for conditioning on the input was pioneered by Liu et al. (2022b), who used RoBERTa embeddings to identify relevant GPT-3 prompts for sentiment, table-to-text, and QA tasks. They found that kNN works better than a random-selection baseline, and that the advantage grows as the size of the (domain-controlled) example pool increases.

Work on prompting LLMs for MT began with the GPT-3 and PaLM papers (Brown et al., 2020; Chowdhery et al., 2022), which adopted similar approaches, comparing 0, 1, and n-shot[1] random selection of independent sentence pairs from WMT training corpora, and testing on older French, German, and Romanian WMT test sets traditionally used in ML, augmented in PaLM with French→German and Kazakh. For both models, performance increased with number of shots, and n-shot BLEU scores were found to be competitive with previous unsupervised SOTA, and in some settings—particularly into English—supervised SOTA as well.

In other early MT work, Reynolds and McDonell (2021) experimented with prompt templates for GPT-3, and found that 0-shot prompts with carefully-chosen templates can outperform n-shot prompts with sub-optimal templates. Garcia and Firat (2022) explored using prompts with mT5 (Xue et al., 2021) to control output attributes such as formality, and also examine the effect of using prompt-like natural-language tags during fine-tuning. Patel et al. (2022) proposed autoregressive prompting: concatenating only the first predicted word to a prompt and output prefix at each step.

Après nous, le déluge

Since our paper appeared on arXiv in November 2022, there has been a flood of work on using LLMs for MT, which we summarize briefly for completeness. A number of papers (Agrawal et al., 2022; Zhang et al., 2023; Jiao et al., 2023; Hendy et al., 2023) investigate prompt quality and source proximity using methods similar to ours but with different LLMs, notably GPT-3.5, GPT-4 and their instruction-tuned counterparts. Their findings are in line with ours, with the exception of Agrawal et al. (2022), who achieve significant gains using

[1] Where n is 64 for GPT-3 and 5 for PaLM.

lexical matching augmented with a diversity mechanism to select prompts. Apart from differences in model and setting, a potentially salient discrepancy is their emphasis on BLEU rather than neural metrics to measure performance. Other interesting work that conditions prompts on source segments uses dictionaries to supply translations in low-resource settings (Ghazvininejad et al., 2023; Lu et al., 2023), or chain-of-thought inspired prompts that elicit keywords, topic, and related examples from the model itself (He et al., 2023).

Further recent work looks at the role of data, attributing LLM MT capabilities to the presence of incidental bilingual examples (Briakou et al., 2023), or showing that parallel data (Schioppa et al., 2023), dictionaries (Jones et al., 2023), or restriction to bilingual settings (Garcia et al., 2023) can boost performance in smaller LMs. Another popular line aims at controlling various properties of translations such as formality or use of specified terminology, either statically (Garcia et al., 2023; Moslem et al., 2023) or with human interaction (Pilault et al., 2023). Finally, there is extensive work on analyzing the translation output of LLMs, generally finding that it is more fluent than accurate (Hendy et al., 2023; Anonymous, 2023), good at handling document context (Wang et al., 2023; Karpinska and Iyyer, 2023) but also prone to problems such as hallucination (Zhang et al., 2023; Guerreiro et al., 2023), and frequently sub-par in low-resource settings (Zhu et al., 2023; Bawden and Yvon, 2023)

3 Prompting for Machine Translation

For a general task, prompting an LLM to generate a desired output y from an input x can involve many steps (Liu et al., 2021), including template generation, slot filling, answer search, and answer mapping. In MT, the answer search and mapping processes are simplified because the answers generated by the LLM can be used directly; we simplify further by using a fixed template. What we explore in depth is the slot filling portion; in particular, we test a variety of methods to select few-shot examples for the prompt.

In initial experiments we determined that for few-shot prompting the exact form of the template is unimportant, see Appendix A for details. Following this observation, we decided to adopt simple templates where each example if preprended by the corresponding language name. These results in prompts of the form (for n-shot prompting):

```
[source]: [X_1]
[target]: [Y_1]
...
[source]: [X_n]
[target]: [Y_n]
[source]: [ X ]
[target]:
```

where `[source]` and `[target]` are instantiated with the names in English of the source and target languages, e.g. `English` and `German`. Note that this scheme has been found to be present in the training data as a marker for multilingual content (Briakou et al., 2023). Each slot pair (X_i, Y_i) is filled with a translation example for these languages, and the final slot X is filled with the current source text. Our algorithm for n-shot translation from a source text x to a target text y is:

1. Choose translation example pairs (x_1, y_1) ... (x_n, y_n). In general, these can depend on x.

2. Plug the example pairs and x into the template. Condition PaLM on the resulting string.

3. Perform a greedy search,[2] stopping when the model outputs a newline.

4. Output the predicted suffix verbatim as y.

Example selection operates in two phases: first choose a pool containing parallel text, then choose examples from the pool. Choosing the pool lets us control global attributes of examples such as domain and average quality. Our baseline method for choosing examples is to select them randomly from the pool. We also experiment with selecting examples that are "closest" to the source text, on the hypothesis that such examples will help guide the model to produce similar translations.

To find relevant examples, we use k-nearest neighbor (kNN) search on the source side of our parallel pool, inspired by Khandelwal et al. (2021). We carry out the search itself using the method of Guo et al. (2020)[3], and investigate two possible representations of the sentences, with associated distance measures:

[2] We found that using a sampling temperature other than 0 tended to degrade translation quality.

[3] Available at https://github.com/google-research/google-research/tree/master/scann.

LP	Year	#sents	Ref
en → de	2021	1002	C
de → en	2021	1000	B
en → zh	2021	1002	A
zh → en	2021	1948	A
en → fr	2014	3003	N/A
fr → en	2014	3003	N/A

Table 1: Test set information, including the newstest dataset year and, when applicable, the reference we use for scoring.

Bag-of-words (BOW): Each sentence is represented by a (sparse) vector of counts associated with words in the vocabulary. As the associated distance measure we use cosine distance. This representation focuses on the surface form of the words, and thus favors lexical similarity between the examples.

RoBERTa: Sentences are represented as embeddings in the space defined by RoBERTa (Liu et al., 2019), a multilingual transformer-based model, with Euclidean distance used for retrieval. We expect these embeddings to reflect the semantics of the sentence, and thus retrieve prompts that are relevant to their subject matter.[4]

4 Data

We experiment with translation into and out of English for Chinese, French and German. After English (78.0%), German (3.5%) and French (3.3%) are the two largest languages in PaLM's 780B token training corpus; Chinese (0.4%) is the 15th largest, and it also represents an inherently more difficult translation task. To facilitate comparisons with recent SOTA systems, and to minimize the chance of overlap with PaLM's training corpus, we test on news data from the WMT 2021 evaluation campaign (Akhbardeh et al., 2021). Since French was not included in WMT21, we use data from WMT14; apart from being older, these test sets are not purely source-original (Freitag et al., 2019) like the more recent ones. Table 1 shows statistics for our test data.

LP	Pool	Size	
		en → X	X → en
de ↔ en	WMT-full	96M	
	WMT-dev	11 732	13 060
	high-end	152 para.	
zh ↔ en	WMT-full	55M	
	WMT-dev	7 481	5 916
	high-end	170 para.	
fr ↔ en	WMT-full	40M	
	WMT-dev	2 886	2 957
	high-end	98 para.	

Table 2: Size of the different prompt pools, measured in sentences for the WMT sets and in paragraphs for the high-end pool.

For prompt selection, we use three distinct pools: the full WMT training corpus for each language pair (WMT-full), the corresponding WMT development sets (WMT-dev), and a manually-curated "high-end" pool. Sizes are shown in Table 2. The WMT-full pool is largest and offers the highest probability of close kNN matches, but it is crawled text drawn from sources of varying quality. The WMT-dev pool has generally better quality, and is a closer domain match to our test set; to encourage PaLM to produce more natural text, we included only target-original texts.[5] For German ↔ English and Chinese ↔ English we include all the news test sets from 2010 to 2020. As English ↔ French was discontinued after 2015, we used sets from 2010 to 2013, augmented with newsdiscussion2015.

The high-end pool comes from websites containing bilingual articles that we judged to be professionally edited, with native or near-native quality in both languages. The articles are drawn from various domains (biography, business, commentary, culture, fashion, food, news, and obituary), with the news domain of the test sets comprising less than 50% for each language. We treat these articles as symmetrical, and use them as prompt sources in both translation directions. Due to the non-literal nature of the translations, there is frequently no 1-1 correspondence between sentence pairs, so we extract aligned paragraphs for prompting. More detailed information about the high-end pool is provided in Appendix B.

[4] Note that it would be conceivable to use PaLM itself as embedding model, which would provide a representation (and associated similarity measure) closer to the application that we are targeting. However, due to the high computational cost and large amounts of data (for some experiments we embed the totality of the WMT training data) we decided to use a smaller model.

[5] As identified by SACREBLEU.

5 Experiments

For compatibility with Chowdhery et al. (2022), we ran all experiments at the sentence level, translating each test sentence individually and in isolation from its context. This deprives PaLM of the ability to exploit the longer contexts it was exposed to during training, but it matches the operating mode of our baselines (including SOTA baselines), and facilitates evaluation.[6] We leave an exploration of potential gains from conditioning on longer histories to future work.

In preliminary experiments, we varied the number of shots from 0 to 10, and found clear performance gains as we increased the number of shots, with diminishing returns after 5 sentence pairs (see Appendix A). Accordingly we report all results on the WMT pools in the 5-shot setting, where each shot is a single sentence pair, matching the configuration in Chowdhery et al. (2022). For the high-end pool, lacking 1-1 sentence alignments, we use 1-shot examples, where each shot is a single paragraph pair. This provides roughly the same quantity of text as 5-shot with sentences, although it creates a stylistic mismatch with our test setup, as we still translate on a sentence-by-sentene basis, as in the other conditions.

When randomly selecting examples, we observed that there is little variability in automatic scores when selecting different samples[7] (see Appendix C). For the results reported in this section, we let PaLM produce translations with 5 different seeds and we selected the run with the median BLEURT score. Translation time was some orders of magnitude longer than a dedicated translation system.

Following recent recommendations (Kocmi et al., 2021; Freitag et al., 2021a) we favour neural metrics (BLEURT in our case) over BLEU, although we also report BLEU scores for completeness. We use a cased version of BLEURT (Sellam et al., 2020) that is based on RemBERT (Chung et al., 2020). We use BLEU as implemented in SACREBLEU[8] (Post, 2018), with zh tokenization for English-Chinese, and 13a tokenization for all other languages.

To perform human evaluation, we hired professional translators (7 for En→De, 5 for De→En, 4 for Zh→En, and 4 for En→Zh) and measure translation quality with a document-context version of MQM (Lommel et al., 2014) which mimics the setup proposed in Freitag et al. (2021a). This includes using the same error categories, severity levels and error weighting schema. As suggested in the study, we weight each major error with 5 and each minor error with 1, except for minor punctuation errors which get a score of 0.1. We depart from Freitag et al. (2021a) in using only a single annotator per segment, and in not imposing a limit of 5 errors per sentence. Additionally, due to technical restrictions on the length of an evaluation session, we limited the MQM evaluation to the first 12 segments per document.

5.1 Selection strategies and pools

We warm up by comparing example selection strategies on the two WMT pools, using automatic metrics to evaluate quality on English↔German. Results are shown in Table 3. The main observation is that the choice of pool is much more important than the selection method: the results for WMT-dev are notably higher than those for WMT-full across all settings. When comparing kNN selection methods, RoBERTa is more effective than BOW, but it does not provide a consistent advantage over random selection.

We conjecture that the quality of an example is more important than its proximity to the current source sentence. The larger size of the full WMT pool means that the kNN approaches will in general be able to find examples that are closer to each source sentence than those from the dev pool, but any resulting gain is offset by the greater risk that an example from the full pool will be a poor translation (since we match only on the source side). Interestingly, had we relied only on BLEU, we would have concluded that the choice of pool is unimportant, and that random selection consistently outperforms kNN.

5.2 Results on all language pairs

Table 4 contains our main results, for German ↔ English, Chinese ↔ English, and French ↔ English. For each language pair, we ran PaLM with random selection on all three pools and with kNN RoBERTa on the WMT-full pool. We compared these systems to output from the best performing system in the 2021 WMT evaluation campaign for

[6] Evaluation of document-level translations is complicated by potentially non 1-1 sentence correspondences, resulting in long translation units that are truncated by BLEURT and can be difficult for humans to rate reliably.

[7] Note that this holds for document level scores. The effect on single sentences can still be very important, cf. Figure 1.

[8] SACREBLEU signature: `nrefs:1|case:mixed|eff:no|tok:TOK|smooth:exp|version:2.1.0`, where TOK is `13a` or `zh`.

LP	Pool	Selection	BLEURT	BLEU
en → de	full	random	71.8	32.9
		*k*NN BOW	71.7	32.4
		*k*NN RoBERTa	73.0	32.5
	dev	random	74.8	32.8
		*k*NN RoBERTa	74.8	32.3
de → en	full	random	74.8	38.4
		*k*NN BOW	72.7	36.9
		*k*NN RoBERTa	73.8	35.4
	dev	random	75.9	38.0
		*k*NN RoBERTa	75.8	37.2

Table 3: Comparison of example selection strategies on the WMT-full and WMT-dev pools. Values for random selection are averaged over 5 runs.

German and Chinese, and for off-the-shelf Google Translate for all six language pairs. We evaluate with BLEU and BLEURT as in the previous section, augmented with human MQM assessments for German and Chinese. French is a special case, as its evaluation set is eight years old, and it is difficult to ensure that any of the MT systems we evaluate have not been exposed to it during training. We include it mostly for the purposes of comparison to Chowdhery et al. (2022), and do not provide SOTA results or perform human evaluation.

Comparing PaLM results for German and Chinese, the pattern from the previous section holds up: random selection from the WMT-dev pool outperforms selection from the full pool. MQM scores correlate well with BLEURT for these results. Despite domain and style mismatch, results for the high-end pool are very similar to those for WMT-dev—closer than any results on the full pool—adding support to the hypothesis that example quality is the main determinant of PaLM's output quality.

The French results reverse the general pattern. For this language pair, random selection from the WMT-full pool does best, although the results for all methods are fairly similar, with a difference of approximately 0.5 BLEURT between the best and worst. One potential explanation is the age and quality of newstest2014, as WMT test-set creation has dramatically improved since then.

Turning to a comparison between PaLM and conventional MT systems, the specialized SOTA systems have a substantial advantage of between 1 and 3 BLEURT points over the best PaLM results, a gap that is reflected in their much lower MQM scores. The difference is narrower for the general-purpose Google Translate system: less than 1 BLEURT except for Chinese→English (1.8), with French→English at parity. PaLM's performance relative to the best MT system for each language pair is generally better when translating into English, where it is lower by 1.0, 2.3, and 0.0 BLEURT for German, Chinese, and French, compared to drops of 2.1, 2.5, and 0.6 in the reverse direction.

The MQM results show some interesting characteristics of translations produced by PaLM. In all language pairs evaluated, fluency MQM scores for PaLM are generally similar to those for SOTA systems, while accuracy scores are lower. The accuracy gap is dominated by Major Accuracy/Omission errors, followed by inconsistent patterns of other Accuracy/* errors across language pairs. In some languages, the best-performing PaLM systems make fewer Style/Awkward errors than SOTA. Table 5 shows a selection of MQM error counts for PaLM WMT-dev random and SOTA systems; full details are provided in Appendix D.

5.3 Comparison to previous results

Our only results that are directly comparable to the few-shot results from Chowdhery et al. (2022) are the WMT-full BLEU scores in table 4c (WMT14 French test-set). Our result for French→English matches theirs exactly, but our score for English→French is lower by 1.7 (42.3 versus 44.0). We attribute this discrepancy to their use of the SACREBLEU `intl` tokenizer; when we evaluate our output using this version, we obtain matching scores.

Our general finding that PaLM's into-English performance is better than the reverse direction matches the conclusion from Chowdhery et al. (2022), while our comparison with recent SOTA systems on current test sets contrasts with their results indicating that PaLM can rival supervised performance in older settings.

6 Analysis

In this section we delve further into various aspects of PaLM's MT performance.

6.1 *k*NN versus random prompts

To understand the performance difference between *k*NN RoBERTa and randomly-selected examples, we performed a qualitative analysis, choosing sen-

LP	System		MQM ↓	BLEURT ↑	BLEU ↑
en → de	WMT21 Facebook Submission (Tran et al., 2021)		**1.18**†	**76.9**	**42.0**
	Google Trans.		1.59	75.7	39.8
	PaLM	WMT-full random	1.90	73.7	**32.9**
		WMT-full kNN	1.93	73.0	32.5
		WMT-dev random	**1.58**	**74.8**	32.8
		high-end random	1.67	74.7	32.9
de → en	WMT21 Facebook Submission (Tran et al., 2021)		**1.31**†	**76.9**	**41.9**
	Google Trans.		1.71	76.4	40.9
	PaLM	WMT-full random	2.38	74.7	38.3
		WMT-full kNN	3.03	73.8	35.4
		WMT-dev random	1.92	**75.9**	38.0
		high-end random	**1.89**	75.8	**38.8**

(a) German→English (nt2021). All MQM results labelled with † are significantly better than all other systems based on PERM-BOTH pair-wise significance testing (Koehn, 2004) with $p = 0.05$.

LP	System		MQM ↓	BLEURT ↑	BLEU ↑
en → zh	WMT21 WeChat Submission (Zeng et al., 2021)		**2.47**†	**66.6**	**36.9**
	Google Trans.		3.23	65.0	36.2
	PaLM	WMT-full random	4.35	62.2	28.6
		WMT-full kNN	5.06	60.7	28.5
		WMT-dev random	**3.24**	**64.1**	29.2
		high-end random	3.70	63.9	**29.6**
zh → en	WMT21 Borderline Submission (Wang et al., 2021)		**3.11**	**70.0**	**33.4**
	Google Trans.		3.12	69.5	32.2
	PaLM	WMT-full random	3.95	67.2	**25.8**
		WMT-full kNN	4.06	65.8	23.8
		WMT-dev random	**3.60**	67.5	25.3
		high-end random	3.89	**67.7**	25.1

(b) Chinese→English (nt2021). All MQM results labelled with † are significantly better than all other systems based on PERM-BOTH pair-wise significance testing (Koehn, 2004) with p=0.05.

LP	System		BLEURT ↑	BLEU ↑
en → fr	Google Trans.		76.5	45.7
	PaLM	WMT-full random	**75.9**	**42.3**
		WMT-full kNN	75.3	41.8
		WMT-dev random	75.4	41.9
		high-end random	75.2	38.6
fr → en	Google Trans.		77.7	43.2
	PaLM	WMT-full random	**77.7**	**42.7**
		WMT-full kNN	77.3	41.2
		WMT-dev random	77.2	42.1
		high-end random	77.6	40.4

(c) French→English (nt2014).

Table 4: Translation results for all language pairs. Values for random selection are the BLEURT median of 5 runs.

LP	Sev.	Category	PaLM	SOTA
de → en	Major	Omission	51	19
en → de	Major	Omission	26	7
zh → en	Major	Omission	109	42
en → zh	Major	Omission	80	46
de → en	Minor	Awkward	73	81
en → de	Minor	Awkward	166	144
zh → en	Minor	Awkward	205	284
en → zh	Minor	Awkward	115	142

Table 5: Selected MQM error count comparisons between PaLM WMT-dev random and SOTA. Omission is a subcategory of Accuracy errors, and Awkward is a subcategory of Style. Full details are provided in Appendix D.

Year	LP	% Clean
2014	fr → en	**69.2**
	en → fr	93.6
2016	de → en	**80.3**
	en → de	97.3
2021	en → de	99.6
	en → zh	99.7
	de → en	97.9
	zh → en	98.1

Table 6: The size of clean (lacking 15-gram target-side overlap with PaLM training data) versions of test sets for various WMT years and language pairs

tences with the largest BLEURT difference between the two systems. Table 14a in Appendix F shows an example where the kNN system correctly retrieves relevant translation examples in the football domain, guiding PaLM to produce a better translation than the random selection system. This contrasts with the example in Table 14b, where the retrieved source sentences are also from the relevant domain, but all have alignment errors, causing PaLM to generate hallucinated output. In general, random selection is also prone to landing on alignment errors, but as each prompt is selected independently, the odds that *all* examples will be errors are low. An informal analysis of kNN examples indicates that if one non-parallel prompt is selected, the others also tend to be of poor quality, perhaps due to corpus alignment errors that are concentrated in particular documents or topics. Since kNN matches only on the source side, it is not robust to this noise.

6.2 Example Translations

Example translations comparing PaLM and SOTA systems for German→English and English→Chinese are given in Appendix 6.2, in Table 15 and Table 16, respectively. We compared the translations of both systems and chose examples that are short, but include the most frequent patterns that we observed also in longer translations. In general, PaLM's translations are less literal when compared to supervised NMT systems. Even though this is one of the strengths of PaLM, it occasionally misses some important information in the source or hallucinates facts not present in the source sentence. The supervised models on the other hand are faithful to the source; this reduces the risk of omission and addition errors, but occasionally leads to translations that are not natural in the target language (e.g. translating street names or using the wrong time format). These findings are in line with the MQM results presented in section 5.2.

6.3 Overlap of test and training data

One major change with respect to Chowdhery et al. (2022) is our use of more recent WMT test sets, which are unlikely to overlap with PaLM's training data.[9] We test this hypothesis using the technique from Chowdhery et al. (2022), which involves measuring high-order n-gram matches; specifically, we measure 15-gram overlap as tokenized by the mBERT tokenizer (Devlin et al., 2019).[10] For test sequences with fewer than 15 tokens, we consider them overlapping if the complete sequence is found as a subsequence of a training example. We report the degree of overlap by showing the percentage of original test examples that survive in the clean test set after removing overlap in Table 6. This confirms that the older French→English and German→English sets have substantial overlap with PaLM's training data, while the newer test sets, whether into or out of English, have much smaller overlapping portions.

Chowdhery et al. (2022) also measure the effect of test-set overlap on translation quality, comparing scores on the original test set to the clean set with overlapping examples removed. In section H we

[9] Here we measure target-side overlap only; we assume there is no substantial parallel data in PaLM's training corpus, and therefore no substantial parallel overlap.

[10] We selected the mBERT tokenizer, as opposed to the PaLM's sentence-piece tokenizer, because it decouples the measurement of overlap from the model under test.

report similar scores for the older test sets, and extend the analysis to calibrate the effect of overlap on MT evaluation, by comparing to an overlap-free off-the-shelf system.

7 Conclusion

We perform a careful assessment of the sentence-level MT capabilities of PaLM, which we compare to SOTA and a current off the shelf (COTS) MT system for three high-resource languages—German, Chinese, and French—into and out of English, using the latest test sets from WMT. We chose to focus on a small set of high-resource language pairs in order to test the claims of the original PaLM paper, which are most striking for these pairs. The time and expense of performing high-quality human evaluations precluded a broader investigation.

Comparing kNN and random strategies for selecting 5-shot translation examples to instantiate fixed prompt templates, we find that kNN's potential advantage in identifying examples relevant to the source sentence is outweighed by its susceptibility to corpus noise. Choosing examples randomly from small, high-quality pools works well, and performance appears to be independent of the domain and translation style of the pool, suggesting that example quality is the most important factor.

Using both the BLEURT metric and MQM human evaluations, we show that PaLM's performance, while very impressive for a system never deliberately exposed to parallel text, still significantly lags that of competition-grade SOTA systems on recent WMT test sets, and to a lesser extent the performance of COTS systems as well. This contrasts with some of the findings of Chowdhery et al. (2022). As in that work, we find that performance into English is somewhat better than the reverse direction. Finally, we perform an extensive analysis of the characteristics of PaLM's MT output, notably finding that in all languages we tested it tends to be creative and fluent but prone to omissions and other accuracy errors; broadly speaking, it matches the fluency but lags the accuracy of conventional NMT.

In future work we look forward to testing PaLM on document-level translation tasks, unleashing its formidable capacity for leveraging long contexts. We would also like to explore prompt tuning methods that are more sophisticated than the hard-prompt setting we adopted for this paper, particularly to see if these might offer a way to tighten up PaLM's MT accuracy without destroying its impressive ability to generate highly-fluent text.

Limitations

As we use only a small number of language pairs, it is not clear how general our conclusions are; in particular, they pertain only to languages that are well represented in PaLM's training corpus, and only to translation into and out of English. Our restriction to independent sentence-level translations may have caused us to underestimate PaLM's true capabilities, since some of the accuracy problems we observed might be considered less severe in the context of whole-document translation where less literal translations are more typical. Our exploration of prompting barely scratches the surface of the many methods that have been proposed for adapting LLMs to particular tasks, and we may have missed a technique that produces higher-quality translations than we observed. Finally, the human evaluation we rely on to provide our most accurate results is necessarily subjective, and if we were to have carried out the evaluation with different raters and a different methodology, our conclusions might well have been different.

Ethical Considerations

Working with large language models comes with many ethical concerns that are discussed in detail in Brown et al. (2020) and Chowdhery et al. (2022). There, MT is often one task of many, while we focus on the question of proper example selection for few-shot prompting of MT, which adds a few specific concerns. Our conclusion that prompt quality is important could lead one to build a system with prompts drawn from a small set of trusted sources; indeed, our high-end set is one such example of this. In such a scenario, this small source will have an outsized impact on the output of the translation system, and one must be careful to manage issues of attribution and intellectual property. Furthermore, an editorial choice defining high-quality language can potentially reduce quality for groups and topics not typically discussed in this style (Gururangan et al., 2022). Finally, by highlighting the power of few-shot examples, one might be tempted to turn example selection over to the users of a system. There, special steps must be taken to avoid exposing users to biased or toxic outputs, which may be triggered by unconstrained prompting (Gehman et al., 2020; Costa-jussà et al., 2022).

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*.

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.

Anonymous. 2023. Does gpt-3 produces less literal translations? Anonymous preprint under review.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *arXiv preprint arXiv:2305.10266*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint:2010.12821*.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Javier Ferrando, and Carlos Escolano. 2022. Toxicity in multilingual machine translation at scale. *arXiv preprint arXiv:2210.03070*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

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.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

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.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*.

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

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *arXiv preprint arXiv:2208.01066*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. 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 4921–4933, Online. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. Metaprompting: Learning to learn better prompts. *arXiv preprint arXiv:2209.11486*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation. *arXiv preprint arXiv:2303.15265*.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4507–4518.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022a. Probing via prompting. *arXiv preprint arXiv:2207.01736*.

Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Xin Zhao. 2022b. Learning to transfer prompts for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3506–3518, Seattle, United States. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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

4582–4597, Online. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. 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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. 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*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2022. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. *arXiv preprint arXiv:2301.10309*.

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, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. 2023. Cross-lingual supervision improves large language models pre-training. *arXiv preprint arXiv:2305.11778*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Chandan Singh, John X Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. 2022. Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*.

Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics*.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission.

In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Josef Valvoda, Yimai Fang, and David Vandyke. 2022. Prompting for a conversation: How to control a dialog model? *arXiv preprint arXiv:2209.11068*.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent translation system for the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 216–224, Online. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. 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.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. WeChat neural machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 243–254, Online. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Appendices

A Prompt Exploration

As preliminary experiments we tried different prompting templates:

Language This is the prompt template used in the paper (see Section 3). It prepends the examples with the corresponding language name in English.

Codes Like “Language”, but instead of full English names, two-letter languages codes are used (e.g. “en”, “de”).

Header Like “Language”, but the header “Translate following sentences:” is added.

Textual A textual request for translating a sentence: “Translate X_n from English into German: Y_n”, where X_n and Y_n are the translation examples, as in Section 3. The source sentence X is given with the same template, but without specifying any translation.

Deutsch Like “Language”, but the language names are given in German (“Englisch”, “Deutsch”).

None No added text. Source and target examples are just input one after the other.

	# shots				
Prompt	0	1	2	5	10
Language	63.9	69.1	71.7	73.6	74.4
Codes	59.0	68.5	71.2	73.4	74.1
Header	72.4	69.1	70.7	73.4	74.1
Textual	36.9	67.5	71.8	73.0	73.7
Deutsch	72.6	70.8	71.9	73.5	74.1
None	3.2	38.5	59.6	73.0	74.1

Table 7: BLEURT results with different prompt templates and number of prompts for the English → German translation directions. The prompt examples were randomly selected. The median of 5 runs are shown.

As shown in Table 7, the choice of a prompting strategy has a crucial impact when the number of shots is low, but the effect is reduced when we increase the number of examples shown. The number of examples also has a significant impact on translation quality. We chose to work with 5 examples, as there are diminishing returns when increasing the number of prompts, and choosing a higher number has additional practical implications (e.g. possibly exceeding the maximum input length).

B High-end pool

Table 9 describes the high-end pool. All listed articles were manually downloaded in June–August 2022, and semi-automatically divided into bilingual paragraphs. Our high-end pool consists of all paragraphs from all articles. The domain breakdown for each language pair is shown in Table 8.

Genre	Proportion		
	en ↔ de	en ↔ fr	en ↔ zh
biography	31%	20%	–
business	–	–	15%
commentary	25%	10%	16%
culture	–	44%	14%
fashion	16%	–	–
food	–	8%	–
news	4%	18%	43%
obituary	24%	–	13%

Table 8: Genre distributions for the high-end pool.

C Variability of Random Runs

Table 10 shows the automatic scores for random runs for the German→English language pair. It can be observed that the range of scores is quite small, less than 0.5 BLEURT points for all language directions. For both directions, the use of WMT-dev, as opposed to WMT-full, for the random pool reduces the observed range in BLEURT by at least 0.1.

D Detailed MQM Scores

Table 11 presents MQM scores for PaLM WMT-dev random and SOTA systems in the four language pairs evaluated, along with the breakdown of the scores into their Accuracy and Fluency components. Table 12 presents detailed MQM error counts for PaLM WMT-dev random and SOTA systems in en→de and de→en.

E Significance numbers

We calculate pairwise significance numbers based on PERM-BOTH pair-wise significance testing (Koehn, 2004; Deutsch et al., 2021). Results can be seen in Table 13.

F Example Prompts

Tables 14a and 14b show prompt examples where kNN and random selection do better, respectively, as described in section 6.1.

G Example Translations

Tables 15 and 16 show example translations for German→English and English→Chinese as described in section 6.2.

H Overlap Analysis

Chowdhery et al. (2022) show BLEU differences between clean and original test sets, and provide some evidence that differences are not due to memorization, but it still isn't clear how much overlap actually inflates a model's score. We directly quantify the effect of train-test overlap on decision making by comparing 5-shot PaLM to Google Translate (GT)[11] on our two sets with substantial overlap, testing under original, clean and ¬clean (including only overlapping examples) scenarios. BLEU and BLEURT scores for the two systems and three test sets are shown in Table 17.

We can see that directly comparing original and clean results for a single system conflates differences from overlap with those from the increased difficulty of the clean subset. For example, for de→en BLEU, comparing PaLM's original and clean scores gives an overlap gap of 2.6-BLEU, in line with the gaps reported by Chowdhery et al. (2022). However, the non-overlapping GT system also has lower scores on the clean set, indicating that it may simply be more difficult.[12] It's more useful to see that the original test indicated a 1.5-BLEU difference between the two systems, while the clean test indicates a 2.0-BLEU difference, meaning PaLM benefited from overlap by 0.5 BLEU in this comparison. The fully overlapping ¬clean further distorts the difference between the two systems: the true (clean) delta of 2.0 BLEU shrinks to only 0.4. Trends for fr→en are similar: though PaLM and GT are very close according to the original test set, the clean set reveals a delta of 0.8 BLEU. Interestingly, BLEURT may be less sensitive to overlap, with the original-versus-clean deltas hovering around 0 for fr→en regardless of the test subset, and de→en showing that PaLM benefits from an overlap bonus of only 0.3 BLEURT.

In summary, overlap between the target side of the test data and the LLM training data can have an impact on both BLEU and BLEURT scores, altering the delta between two systems where one benefits from overlap and another does not by up to 0.7

[11] We chose Google Translate for comparison because it is non-trivial to build a SOTA baseline for older WMT scenarios. Through personal communication, we understand that Google Translate has no overlap with WMT test sets.

[12] The difference in difficulty between Clean and ¬Clean for systems without overlap is not easily explained. A common difficulty indicator is sentence length, but average lengths, as measured by number of SACREBLEU tokens per sentence, are similar between Clean and ¬Clean for both de→en (23.8 versus 23.0) and fr→en (21.1 versus 22.7).

LP	paras	words	URL
en ↔ de	4	255	www.deutschland.de/en/news/new-supercomputer-in-operation
	4	208	www.deutschland.de/en/news/patents-germany-ranks-second
	11	609	www.deutschland.de/en/news/syrian-swimmer-yusra-mardini-provides-message-of-hope-at-olympics
	24	1787	www.zeit.de/kultur/2019-12/schoenheit-fotografie-aesthetik-rankin-mitch-epstein-roger-ballen-english
	28	2817	www.zeit.de/kultur/2020-07/desinformation-peter-pomerantsev-social-media-regulation-democracy/komplettansicht
	60	2961	www.zeit.de/politik/ausland/2020-11/polarization-us-elections-democrats-republicans-donald-trump-family-division-english
	21	2757	www.zeit.de/politik/deutschland/2015-11/helmut-schmidt-obituary-english/komplettansicht
en ↔ zh	30	1323	cn.nytimes.com/asia-pacific/20220509/taiwan-china-covid/dual
	31	1317	cn.nytimes.com/china/20220427/brownface-barrack-okarma-1968-hong-kong/dual
	6	780	cn.nytimes.com/china/20220401/china-cheng-lei-australia/dual
	13	609	cn.nytimes.com/china/20220421/china-eastern-crash-report/dual
	23	1520	cn.nytimes.com/china/20220412/china-russia-propaganda/dual
	22	1373	cn.nytimes.com/business/20220621/china-housing-real-estate-economy/dual
	13	478	cn.nytimes.com/china/20220415/shanghais-food-crisis-prompts-residents-in-beijing-to-stockpile-supplies/dual
	26	1202	cn.nytimes.com/obits/20220418/peng-ming-min-dead
	6	843	https://cn.nytimes.com/world/20220330/solomon-islands-china/dual
en ↔ fr	6	846	france-amerique.com/a-france-of-many-colors
	10	1177	france-amerique.com/alice-guy-cinema-forgotten-pioneer
	10	1237	france-amerique.com/americanization-is-back-did-it-ever-go-away
	8	666	france-amerique.com/a-propos-a-hard-hitting-french-american-podcast
	3	457	france-amerique.com/camille-laurens-a-womans-life
	8	970	france-amerique.com/football-and-soccer
	6	377	france-amerique.com/france-united-states-naval-battle-and-diplomatic-crisis
	7	615	france-amerique.com/jeanne-damas-all-the-women-in-her-city
	6	631	france-amerique.com/guedelon-building-a-castle-by-hand
	11	811	france-amerique.com/raphael-francois-culinary-director
	12	874	france-amerique.com/thierry-mugler-provocateur
	7	934	france-amerique.com/winds-of-change-over-democracy
	4	255	www.deutschland.de/en/news/new-supercomputer-in-operation

Table 9: Sizes and provenance for articles in the high-end prompt pool. The *words* column contains the number of English words (whitespace-separated character sequences) in each article.

		BLEURT					BLEU				
LP	Pool	Run 1	Run 2	Run 3	Run 4	Run 5	Run 1	Run 2	Run 3	Run 4	Run 5
en → de	full	71.9	71.9	71.6	71.8	71.9	32.4	32.8	32.1	32.9	32.9
	dev	74.7	74.7	74.7	74.9	74.8	32.7	32.6	32.6	32.6	32.8
de → en	full	74.8	75.0	74.8	74.5	74.7	38.4	38.5	38.2	38.0	38.3
	dev	75.9	75.9	76.0	75.7	75.9	38.0	38.0	38.0	38.3	38.2

Table 10: Results for random runs for the German→English translation direction.

	PaLM			SOTA		
	MQM ↓	Accuracy↓	Fluency↓	MQM ↓	Accuracy↓	Fluency↓
en → de	1.58	1.12	0.46	1.18	0.81	0.37
en → zh	3.24	2.69	0.52	2.47	1.96	0.48
de → en	1.92	1.43	0.48	1.31	0.88	0.43
zh → en	3.60	2.97	0.62	3.11	2.43	0.68

Table 11: MQM scores for PaLM WMT-dev random and SOTA systems, split into Accuracy and Fluency. Accuracy scores include "Accuracy/*," "Terminology/*," and "Non-translation!" error categories. Fluency scores include "Fluency/*," "Style/*," and "Locale/*" categories. The "Other" error category is not included in Accuracy or Fluency scores.

	en → de				de → en			
	PaLM		SOTA		PaLM		SOTA	
	Major	minor	Major	minor	Major	minor	Major	minor
Non-translation!	0	0	0	0	2	0	0	0
Acc/Mistrans.	103	89	79	67	73	41	61	49
Acc/Omission	26	6	7	3	51	33	19	11
Acc/Addition	1	6	3	1	10	2	0	3
Acc/Untranslated	12	4	14	0	6	7	5	8
Ter/Inappr	0	7	0	7	17	21	12	15
Ter/Incons	0	4	0	4	1	5	1	7
Fl/Grammar	0	133	0	100	18	41	5	38
Fl/Register	0	2	0	3	0	0	0	0
Fl/Inconsistency	0	2	0	5	0	2	0	2
Fl/Punctuation	0	260	0	31	1	38	2	29
Fl/Spelling	0	12	0	16	0	16	0	17
Fl/Encoding	0	0	0	0	0	0	0	2
St/Awkward	0	166	0	144	13	73	16	81
Locale/Date	0	0	0	0	0	1	0	5
Locale/Name	0	0	0	0	2	8	2	5
Locale/Time	0	0	0	0	0	5	0	5
Source Error	0	0	0	0	1	0	0	1
Other	0	0	1	0	0	3	0	3
Total Errors	142	673	102	362	189	296	123	281

Table 12: MQM error counts for PaLM WMT-dev random and SOTA systems for en→de and de→en. Abbreviations are as follows: “Acc”: Accuracy, “Fl”: Fluency, “St”: Style, “Ter”: Terminology, “Inappr”: Inappropriate for context, “Incons”: Inconsistent.

		SOTA	GTrans.	WMT-dev	high-end	WMT-full	
				random	random	random	kNN
de→en	MQM	1.31	1.71	1.92	1.89	2.38	3.03
	SOTA	-	0.0	0.0	0.0	0.0	0.0
	Google Trans.	-	-	0.073	0.124	0.0	0.0
	WMT-dev random	-	-	-	0.588	0.001	0.0
	high-end random	-	-	-	-	0.001	0.0
	WMT-full random	-	-	-	-	-	0.001
en→de	MQM	1.18	1.59	1.58	1.67	1.90	1.93
	SOTA	-	0.0	0.0	0.0	0.0	0.0
	Google Trans.	-	-	0.512	0.225	0.003	0.003
	WMT-dev random	-	-	-	0.175	0.001	0.0
	high-end random	-	-	-	-	0.021	0.01
	WMT-full random	-	-	-	-	-	0.372
zh→en	MQM	3.11	3.12	3.60	3.89	3.95	4.06
	SOTA	-	0.447	0.0	0.0	0.0	0.0
	Google Trans.	-	-	0.002	0.0	0.0	0.0
	WMT-dev random	-	-	-	0.022	0.006	0.003
	high-end random	-	-	-	-	0.343	0.168
	WMT-full random	-	-	-	-	-	0.281
en→zh	MQM	2.47	3.23	3.24	3.70	4.35	5.06
	SOTA	-	0.0	0.0	0.0	0.0	0.0
	Google Trans.	-	-	0.488	0.004	0.0	0.0
	WMT-dev random	-	-	-	0.002	0.0	0.0
	high-end random	-	-	-	-	0.0	0.0
	WMT-full random	-	-	-	-	-	0.0

Table 13: p-values based on PERM-BOTH pair-wise significance testing (Deutsch et al., 2021). We highlight all numbers with p<0.05.

BLEU or 0.3 BLEURT for a 20-30%-overlap. However, we should emphasize that the differences due to overlap are small overall, and certainly much smaller than expected if one looked only at the difference between original and clean scores.

I Fixed versus random prompts

The results from section 5.2 indicate that random selection from small, high-quality prompt pools can work better than trying to customize prompts for specific inputs. In this section we investigate the effect of using a *single* high-quality prompt for all inputs, chosen using a maximum-likelihood criterion. For convenience, we carried out experiments on the high-end pool with 1-shot paragraph prompts. For each prompt in the pool, we computed the probability of a set of held-out high-end paragraphs when PaLM was conditioned on that prompt. We select the prompt that resulted in the highest probability for each language pair.

Table 18 compares this method to random selection from the high-end pool. For all language pairs except Chinese→English, the fixed prompt does as well or better than the average performance over 5 random runs where a different prompt is selected for each input during each run. In Chinese→English, the prompt that ranked 5th according to the probability criterion also outperformed the random average, suggesting problems with our held-out set for that language pair.

We conclude that using a single high-quality prompt can be a safer strategy than choosing a fresh randomly-selected prompt for each input. Model probability appears to be a reasonable criterion for judging quality, but we look forward to refining this heuristic in future work.

	Source	"Wir haben die Pflichtaufgaben mit Meisterschaft und Pokal einfach hervorragend gemeistert.
	Reference	“Quite simply, we have excellently mastered the necessary tasks for the Championship and the Cup.
kNN RoBERTa	Hyp	"We have simply mastered the tasks of the championship and the cup excellently.
	Prompt 1	**German:** Mit einer verstärkten Mannschaft holte die Mannschaft das Double aus Meisterschaft und Pokal. **English:** The decision paid off as the team achieved a league and cup double.
	Prompt 2	**German:** Darüber hinaus haben wir uns wichtige Meisterschaftspunkte im Kampf um den Vizetitel gesichert.“ **English:** We have furthermore secured some important championship points in the fight about the vice champion’s title.”
	Prompt 3	**German:** Bring deine Mannschaft durch alle Spiele der Europameisterschaft und gewinne den Pokal! **English:** Take your team all the way through the Euro Cup stages and lift the trophy!
	Prompt 4	**German:** So konnte er die französische Meisterschaft, den nationalen Pokal sowie den Supercup gewinnen. **English:** He helped the club to win the national championship and the Supercup.
	Prompt 5	**German:** Roter Stern gewinnt in jener Saison das Double von Meisterschaft und Pokal. **English:** Red Star won their sixth double in this season.
Random	Hyp	"We have the compulsory tasks with mastery and cup simply superbly mastered.
	Prompt 1	**German:** Im November letzten Jahres war ein einzelner Steinadler und auch Bussarde im Blick der Kamera. **English:** In November last year a solitary golden eagle and buzzards too appeared in web camera view.
	Prompt 2	**German:** Teile: Modesto-14th Street, Stanislaus, California, Kalifornien-Luft-Qualitätsindex **English:** Share: Modesto-14th Street, Stanislaus, California, California Air Quality Index
	Prompt 3	**German:** So viel kostet ein Hotel in Chester **English:** How much does a hotel in Chester cost?
	Prompt 4	**German:** "... alle Mitarbeiter, die alles tun, um hilfsbereit zu sein und sehr freundlich zu sein; köstliche Margaritas; Kolibris und Granatäpfel im Garten (sowie eine sehr freundliche Katze); Ein echtes Gefühl von Zuhause. " Aktionsangebot **English:** "... all staff, who go out of their way to be helpful and are extremely welcoming; delicious margaritas; hummingbirds and pomegranates in the garden (as well as a very friendly cat); a real home-from-home feeling. "
	Prompt 5	**German:** Gansevoort Land zum Verkauf **English:** Gansevoort Land for Sale

(a) Example where kNN outperforms random selection.

	Source	Frei von Drogen veröffentlichte Green mit der Peter Green Splinter Group einige Alben, bis sich die Band 2004 auflöste.
	Reference	Free of drugs, Green and the Peter Green Splinter Group released various albums before the band split up in 2004.
kNN RoBERTa	Hyp	The band released their debut album, The Last of the Great Pretenders, in 2003.
	Prompt 1	**German:** Ab 1990 war er Sänger der Gruppe Talisman, mit der er sieben Studioalben veröffentlichte, bis sich die Band 2007 auflöste. **English:** From 1998 until his departure in 2007, he was the lead singer of the group Lonestar, which recorded seven studio albums on BNA Records during his tenure as lead vocalist.
	Prompt 2	**German:** 2001 veröffentlichte die Band unter dem Namen Glass die rockige Single Out Of Nowhere, verpasste die Charts und löste sich im Anschluss auf. **English:** Around this time he wrote and presented the ITV Network productions The Rock that Doesn’t Roll and The Rock That Rolled Away.
	Prompt 3	**German:** Mit ihrer Band Ex Cops veröffentlichte sie zwei Alben, bevor sich die Band 2015 auflöste. **English:** Their new band released two EPs before signing to Apparition Records in 2011.
	Prompt 4	**German:** In seiner Jugend gründete David Haering die Punk-Band Side Effect, mit der er drei Alben und eine EP veröffentlichte. **English:** Peter Hajba and Alexander Brandon used OpenMPT to compose the soundtracks for Bejeweled 2, Bejeweled 3 and other PopCap games.
	Prompt 5	**German:** Nach der Veröffentlichung des Live-Albums Beast from the East 1988 trennten sich die Wege der Musiker, als Don Dokken die Band auflöste. **English:** In 1988, after the Monsters of Rock Tour and a further platinum album, Don Dokken decided to break up the band and they went on their separate ways.
Random	Hyp	Free from drugs, Green released several albums with the Peter Green Splinter Group, until the band broke up in 2004.
	Prompt 1	**German:** 250 gr/m2: eine Reihe merino intermedia, vielseitigkeit und schutz garantiert. **English:** 250 gr/m2: Range merino intermediate, versatility and guaranteed protection.
	Prompt 2	**German:** 127 Moo.3, Choeng Thale, Thalang, Phuket, Strand Bang Tao, Thailand (Karte anzeigen) **English:** 127 Moo.3, Choeng Thale, Thalang, Phuket, Bang Tao Beach (Phuket), Thailand (Show map)
	Prompt 3	**German:** Ich bin stolz, sagen zu können, dass Ihr Produkt mir die Größe verliehen hat, von der ich jahrelang geträumt habe. **English:** I am proud to say that your product has given me the size that I sought for all these years.
	Prompt 4	**German:** In der NBA Saison 2003–2004 schien es endlich besser für Garnett und die Timberwolves zu laufen. **English:** The 2002–03 season seemed to look up for the Timberwolves.
	Prompt 5	**German:** APC-Fhrancis Oliver Lopez extern natürliche Links **English:** APC-Fhrancis Oliver Lopez External Passing Juice

(b) Example where random selection outperforms kNN.

Table 14: Example prompts for random and kNN RoBERTa selection for the German→English translation direction.

Source	Wegen eines ersten offiziellen Corona-Verdachtsfalls im Land hat Nordkorea aus Furcht vor einer Katastrophe eine Großstadt komplett abgeriegelt.
Reference	Due to a first official suspected case of coronavirus in the country, North Korea has completely shut down a major city out of fear of a disaster.
SOTA	Due to a first official suspected case of coronavirus in the country, North Korea has completely sealed off a major city for fear of a catastrophe.
PaLM	North Korea has completely sealed off a city of 300,000 people because of a first suspected case of coronavirus in the country, fearing a catastrophe.

(a) PaLM generated two errors: (i) Makes the assumption that the major city has 300,000 citizens, (ii) Forgot to translate the word official. SOTA generates a perfect translation. Note: this example also exemplifies PaLM's less literal translation style.

Source	Für mindestens drei Menschen endet die Tour mit dem Motorrad tödlich.
Reference	For at least three people, their bike ride ended in death.
SOTA	The motorcycle tour ends fatally for at least three people.
PaLM	At least three people die in motorcycle accidents.

(b) The source mentions a single accident happening on a bike tour. PaLM refers to multiple accidents happening independently.

Source	Ein Zeuge hörte gegen 3.40 Uhr Geräusche in der Talstraße und lief in Richtung des Imbisses.
Reference	One witness heard noises on Talstraße around 3:40 am and ran in the direction of the snack stand.
SOTA	A witness heard noises in the valley road at around 3.40 a.m. and ran towards the snack bar.
PaLM	A witness heard noises in Talstraße at around 3:40 a.m. and ran towards the snack bar.

(c) SOTA generates an overly-literal translation, resulting in copying the street name (Talstrasse) and using the wrong time format.

Table 15: Example translations from newstest2021 German→English. PaLM translations are generated with the high-end prompt pool. These are typical of error patterns observed in the translation output. We also observed the same pattern when using WMT-dev as the prompt pool. In general, SOTA is more faithful to the source while PaLM generates less literal translations that occasionally miss some information from the source.

Source	French World Cup winner Dembele, who has struggled for game time at the Camp Nou, was recently linked with a move to PSG in a swap deal with Neymar.
Reference	在诺坎普球场冲锋陷阵的法国世界杯冠军得主Dembele 最近通过与Neymar 交换转投PSG。
SOTA	法国世界杯冠军登贝莱在诺坎普一直在为比赛时间而挣扎，最近他与内马尔交换转会巴黎圣日尔曼。
PaLM	法国世界杯冠军德容，在诺坎普的出场时间一直不多，最近被传与内马尔进行交换加钱转会到PSG。

(a) PaLM produces two errors: (i) translates a wrong player's name; (ii) adds extra information that the player received a raise in the swap deal. SOTA produces a perfect translation, but is much more literal than PaLM.

Source	. . . in the wake of September 11, ASIO was given power to compulsorily question people for up to seven days in relation to terrorism offences.
Reference	. . . 在911 事件之后，澳安全情报局有权对牵涉恐怖主义行为的人员进行为期最高7 天的强制性询问。
SOTA	. . . 在9月11日之后，安全情报组织被授权对与恐怖主义罪行有关的人进行长达7天的强制性讯问。
PaLM	. . . 澳大利亚安全情报局在9·11恐怖袭击之后获得了强制询问人员的权力，可以在7天内就恐怖主义罪行进行询问。

(b) The source phrase "*September 11*" is translated literally by SOTA into a date, whereas PaLM produces a more appropriate translation by describing it as a terrorist attack.

Table 16: Example translations from newstest2021 English→Chinese. PaLM translations are generated with the WMT-dev prompt pool. We find SOTA to generate more literal translations than PaLM, but PaLM suffers from more omissions and mistranslations.

Data	%Clean	Method	BLEU			BLEURT		
			Orig	Clean	¬Clean	Orig	Clean	¬Clean
de → en 2016	80.3	Google Trans.	47.6	45.5	55.3	78.4	77.7	81.3
		WMT-full Random	46.1	43.5	54.9	77.3	76.3	81.5
		Diff	1.5	2.0	0.4	1.1	1.4	-0.2
fr → en 2014	69.2	Google Trans.	43.1	42.1	44.8	77.7	76.8	79.6
		WMT-dev Random	43.0	41.3	45.4	77.7	76.9	79.5
		Diff	0.1	0.8	-0.6	0.0	-0.1	0.1

Table 17: Comparison between Google Translate and 5-shot PaLM using three test sets: Orig. (original), Clean (overlapping examples removed) and ¬Clean (including only overlapping examples). We use Random instead of WMT-dev Random for de→en to avoid using the WMT 2021 development sets to prompt for the WMT 2016 test ("sampling from the future").

LP	Selection	BLEURT		
		min	avg	max
en → de	fixed		74.7	
	random	74.5	74.7	**75.0**
de → en	fixed		**76.3**	
	random	75.6	75.8	75.9
en → zh	fixed		**64.7**	
	random	63.7	63.9	64.0
zh → en	fixed		67.0	
	random	67.3	67.5	**67.7**
en → fr	fixed		**75.5**	
	random	75.2	75.2	75.3
fr → en	fixed		**77.9**	
	random	77.4	77.6	77.6

Table 18: Fixed (maximum-likelihood) prompts vs random prompts. All prompts are drawn from the high-end pool, and performance is measured on the standard test sets (WMT21 for German and Chinese, WMT14 for French). The scores for random selection are the minimum, average, and maximum over 5 random draws.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Unnumbered Limitations section immediately after Conclusion.

☑ A2. Did you discuss any potential risks of your work?
Unnumbered Ethical Considerations section immediately after Limitations.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Sections 4, 5

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We did not discuss licenses in the paper, but we verified that our use of materials was permitted. We are not distributing artifacts.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We partially address this question in our Ethical Considerations section; as mentioned above, in general we ensured that our use of materials was permitted.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 4, Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 4, Appendix

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We provide partial answers to this question in sections 1 and 5. We are not authorized to provide full details.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5, Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5, 6, Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5, 6, Appendix

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 5, 6, Appendix

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We provide a cite to a paper that reports these instructions (Freitag et al, Experts, Errors, and Context, TACL 2021)

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We use a contractor, and this information is not available to us.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.