In-context Examples Selection for Machine Translation

Sweta Agrawal[1,*], **Chunting Zhou**[2], **Mike Lewis**[2],
Luke Zettlemoyer[2], **Marjan Ghazvininejad**[2]

[1] University of Maryland [2] Meta AI
sweagraw@umd.edu {chuntinz,mikelewis,lsz,ghazvini}@meta.com

Abstract

Large-scale generative models show an impressive ability to perform a wide range of Natural Language Processing (NLP) tasks using in-context learning, where a few examples are used to describe a task to the model. For Machine Translation (MT), these examples are typically randomly sampled from the development dataset with a similar distribution as the evaluation set. However, it is unclear how the choice of these in-context examples and their ordering impacts the output translation quality. In this work, we aim to understand the properties of good in-context examples for MT in both in-domain and out-of-domain settings. We show that the translation quality and the domain of the in-context examples matter and that 1-shot noisy unrelated example can have a catastrophic impact on output quality. While concatenating multiple random examples reduces the effect of noise, a single *good* prompt optimized to maximize translation quality on the development dataset can elicit learned information from the pre-trained language model. Adding similar examples based on an n-gram overlap with the test source significantly and consistently improves the translation quality of the outputs, outperforming a strong *k*NN-MT baseline in 2 out of 4 out-of-domain datasets.

1 Introduction

In-context learning (Brown et al., 2020) has recently received a lot of attention from the NLP research community due to its remarkable ability to utilize only a few input-output examples to perform many NLP tasks (Liu et al., 2021). For example, Lin et al. (2021) demonstrate that a 7.5B multilingual generative model, XGLM, outperforms a supervised sequence-to-sequence baseline in 45 translation directions on the FLORES-101 machine translation benchmark (Goyal et al., 2022) using just 32 randomly sampled translation examples as demonstrations. While these results are compelling, recent work has also shown that the performance and capability of a pre-trained language model (PLM) can be highly sensitive to many factors, such as the choice of in-context examples (Liu et al., 2022b), their ordering (Lu et al., 2022) and the template (Jiang et al., 2020).

Typically, in-context learning for MT uses examples that are randomly sampled from a small development set that resembles the domain of the test dataset. The effect of the aforementioned factors (such as the choice of the examples) on the translation quality of the PLM hence remains unclear and unexplored. Yet another crucial gap in using in-context learning for MT in the current literature is the effect of the domain of in-context examples on translation quality since out-of-domain generalization is a known and important challenge in MT (Koehn and Knowles, 2017).

In this work, we systematically analyze how factors such as the choice and the number of few-shot in-context examples and their ordering impact MT output quality. We show that while noisy unrelated 1-shot example can have a significantly adverse effect on translation quality, a single prompt optimized to maximize the translation quality on a development set can sufficiently elicit task-based information from the PLM. Our analysis thus demonstrates the importance of selecting good examples for MT and raises the question: *What are the properties of good in-context examples for* MT*?* In that direction, our findings suggest that a well-formed meaning-equivalent translation example results in higher quality translation than randomly selected in-context examples.

Motivated by the use of Translation Memory in Computer-Aided Translation (Yamada, 2011) and its usage in computational approaches to Machine Translation (Somers, 1999; Koehn and Senellart, 2010; Khandelwal et al., 2020, *inter alia*), we retrieve similar examples to the test source from

[*] Work done during internship at Meta AI Research.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8857–8873

a datastore that includes pairs of the source text and their corresponding translations via BM25, an unsupervised efficient retriever to provide additional context to the model. We propose a novel in-context example selection and re-ranking strategy to maximize the coverage of the source n-grams in the retrieved examples. Experiments on WMT'19 English↔German and English↔Russian datasets show that our proposed strategy can consistently improve the translation quality over the outputs generated using BM25 retrieved examples. Combining optimized 1-shot task-level with example-specific in-context examples using a simple concatenation strategy further improves translation quality, outperforming state-of-the-art inference-adapted nearest-neighbor MT models (kNN-MT) on two out-of-domain datasets (Medical and IT) while being memory and compute efficient as our approach does not require constructing and querying a dense token-level datastore.

2 Background: In-context Learning

Generating translations from large-scale multilingual language models like mGPT (Shliazhko et al., 2022), XGLM (Lin et al., 2021) or AlexaTM 20B (Soltan et al., 2022) requires conditioning the decoder-only language model with in-context parallel examples. These examples serve two purposes: a) providing the model with the format and knowledge of the task (**task-level**) and b) guiding the output generation via providing useful information about the unseen source sentence (**example-specific**). This is different from the standard sequence-to-sequence models, where the task is always known, and the model learns generalizable patterns from the input-output examples to perform the task (in this case, translation) for the unseen source text.

Source: *Welche Risiken sind mit* **Poulvac FluFend H5N3 RG** *verbunden?*
Template: {Source text} = {Target text}.
Example-Specific: *Welche Risiken sind mit* Sebivo *verbunden?* = What are the risks associated with Sebivo?
Task-Level: Bei PROMESS1 werden drei Hauptziele verfolgt. = PROMESS1 has three main objectives.

Table 1: In-context Examples for Machine Translation.

Formally, given k in-context examples $\{x_i, y_i\}_1^k$ the prefix input or the prompt, x_j^p, is generated by concatenating the demonstration examples $\{(x_i, y_i)\}_1^k$ to the test input, x_j^s according to a *template*, P (see Table 1). The output, $\hat{y}$, is then generated via the PLM with parameters θ via greedy decoding as follows:

$$\hat{y}_{j,t} = \arg\max_{y'_{j,t}} P_{\text{PLM}}(y'_{j,t} | x_j^p, \hat{y}_{j,<t}; \theta) \quad (1)$$

3 Prompt Selection

Ideally, good in-context examples can trigger the pre-trained language model to generate the **desired output** and also elicit the information learned during pre-training (Jiang et al., 2020). Min et al. (2022) show that, for classification tasks, the in-context examples provide information about the task (the distribution of the input text, the label space, and the format of the task) and that the model does not rely on these examples to generate the final output. However, their analysis is limited to a) classification tasks and 2) randomly sampled in-context examples. Prior work has also shown that the order of these in-context examples can also lead to high variance in downstream performance (Zhang et al., 2022). However, less is understood about how these factors impact text generation tasks like MT. *Do we need multiple in-context examples? What makes good in-context examples for* MT*? How sensitive is the model to the order of the prompts?*

In this work, we aim to better understand the impact of prompt selection on the translation quality of the outputs. Given a training dataset consisting of n parallel examples $D = \{x_i, y_i\}_{i=1}^n$, and a test source x_j, we select a subset of m *informative* samples to form a prompt which either provides task-level and/or example-specific information as discussed below.

3.1 Task-level In-context Examples

A good task-level in-context example should be able to elicit information learned during pretraining from the PLM. One way to measure the efficacy of an example as a prompt is via computing the translation quality of the outputs generated when prompting the PLM given an example. Hence, we select the task-level prompt as follow: For a given example sampled from the training dataset, $(x_i, y_i) \in D^S$, we create a prompt, x_i^p by concatenating the example $\{(x_i, y_i)\}$ to each source in the

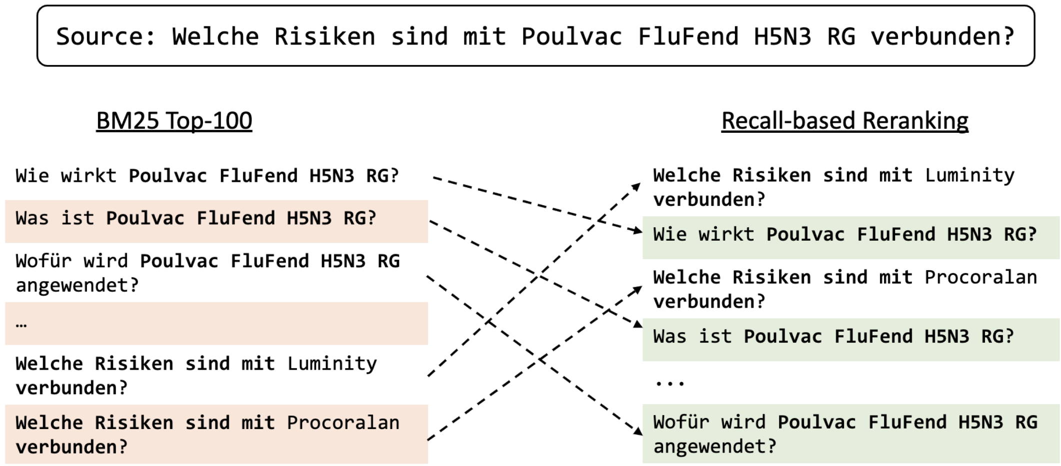

Figure 1: Our proposed strategy can cover all the terms from the input text,"Welche Risiken sind mit Poulvac FluFend H5N3 RG verbunden?", in this case, with just the two examples.

development set. The system outputs are then generated using equation 1. We then rank examples from D^S as task-level prompts based on the BLEU of the generated outputs against the references on this held-out development set, $D^{dev} = \{X, Y\}$:

$$(x_s, y_s) = \underset{(x,y)\in D^S}{\arg\max}\, \text{BLEU}(Y, \hat{Y}) \quad (2)$$

3.2 Example-specific In-context Examples

Prior work on retrieving *good* in-context example-specific prompts for tasks other than MT (like question answering or knowledge retrieval) either trains a dense-retriever (Rubin et al., 2021) or utilizes samples that are closer to the **test source** in the embedding space of a PLM like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or XLNET models (Liu et al., 2022b). While contextual models can generate a global sentence representation, they overlook rare lexicons which can be important for generating translations in unseen domains like medical or IT (Wrzalik and Krechel, 2021).

However, for MT, overlapping n-grams between the source and the retrieved sentences ensures informativeness as the target associated with the retrieved sentence is likely to include partial translations of the source. We can thus use BM25 as an efficient unsupervised retrieval method to retrieve similar examples. However, as the examples are scored independently and BM25 favors rare word matches (Robertson et al., 2009), the top retrieved candidates might not cover all the terms in the source text (Figure 1). Given that the context window of the PLM is usually limited (~ 3096 tokens, $16 - 20$ examples), maximizing the coverage of all the terms found in the test input might be favorable. Hence, we propose to re-rank the top 100 candidates retrieved from BM25 using our algorithm outlined in 1. We extract all the word n-grams, and their counts from the test source, x_j^s and source of the BM25 retrieved examples, $\{P_j(x_i)\}_1^k$ (lines 2-4). Let S and Q denote the set of the source n-grams and the n-grams from a BM25 retrieved example, respectively. We compute a recall-based (R) n-gram overlap score (line 7):

$$R_n = \frac{\sum_{\text{ngram}\in S\cap Q} \text{Count}_{\text{matched}}(\text{ngram})}{\sum_{\text{ngram}\in S} \text{Count}_S(\text{ngram})} \quad (3)$$

$$\text{Score} = \exp\Big(\frac{1}{n}\sum_n \log(R_n)\Big) \quad (4)$$

The example with the maximum score is then added to the set of selected prompts, and the found n-grams from the test source are then down-weighted by a factor, λ, for the next iteration of selection (line 14). For example, setting $\lambda = 0$ will select the example that covers the n-grams from the test source in the subsequent iteration that has not already been encountered. This process is then repeated over the retrieved pool until a set threshold of the score is reached.

Figure 1 shows the top-100 candidates retrieved via BM25 for the input: "Welche Risiken sind mit Poulvac FluFend H5N3 RG verbunden?". The top few candidates provide the same information to the PLM, i.e., translation of the phrase "Poulvac FluFend H5N3 RG". The examples including the other terms ("Welche Risiken sind mit verbunden ?") from the input text, are ranked lower. On the

Algorithm 1: An N-gram Recall-based Strategy to Re-rank In-context Examples

Input: Prompts $\{P_j(x_i, y_i)\}_1^k$ for the test source x_j^s, λ, Threshold
Output :Ordered Selected Prompts $\{T = P_j(x_i, y_i)\}_1^s$, s $\leq$ k
$T \leftarrow$ Empty Ordered List
$S \leftarrow$ EXTRACT_WORD_NGRAMS_WITH_COUNTS (x_j^s)
for $i \in \{1..k\}$ **do**
  $Q[i] \leftarrow$ EXTRACT_WORD_NGRAMS_WITH_COUNTS $(P_j^k(x_i))$
while *True* **do**
  for $i \in \{1..k\}$ **do**
    Score$[i] \leftarrow$ NGRAM_OVERLAP_SCORE $(S, Q[i])$
  if $\max(Score) <$ *Threshold* **then**
    break
  T.append($P_{\arg\max(\text{Score})}$)
  matched_ngrams $\leftarrow S \cap Q[\arg\max(Score)]$
  $Q[\arg\max(\text{Score})] \leftarrow \emptyset$
  for *ngram* $\in$ *matched_ngrams* **do**
    $\text{Count}_S(\text{ngram}) \times = \lambda$
Return T

other hand, our proposed re-ranking strategy can cover all the terms from the input text, in this case, with just the top-2 examples.

4 Evaluation Settings

4.1 Datasets and Evaluation Metric

We perform our in-domain evaluation on the WMT-19 German (de) $\Leftrightarrow$ English (en) and WMT-19 Russian (ru) $\Leftrightarrow$ English (en) datasets (Barrault et al., 2019). For the out-of-domain evaluation, we use the multi-domain dataset from Aharoni and Goldberg (2020) for the following domains: Medical, Law, IT, and Koran. The dataset statistics are reported in the Appendix (Table 8). Following Ng et al. (2019), we normalize punctuation using Moses (Koehn et al., 2007) and remove sentences longer than 250 tokens and sentence pairs with a source/target length ratio exceeding 1.5 from the in-domain datasets. The detokenized length truncated model-generated outputs are evaluated using sacreBLEU (Papineni et al., 2002; Post, 2018).[1] The PLM outputs are truncated to twice the source length, as preliminary analysis suggested degeneration in a few ($\sim$10-20) examples.

4.2 Experimental Conditions

Language Model We use the publicly available checkpoint of the XGLM$_{7.5\text{B}}$, a decoder-only multilingual language model (Lin et al., 2021) for all our experiments, which has 32 layers and a hidden dimension of 4096.

Baselines and Comparisons We consider the following comparisons:

- **Random**: p random few-shot examples sampled from the training dataset (number of trials=3).
- **Task-level**: top-p examples that achieve the highest BLEU on the development set (§ 3.1).
- **Retrieved In-context (BM25)**: q_{max} examples retrieved via BM25, since, unlike task-level examples, there is no guarantee that exactly q similar examples will be found in the training dataset for each input.
- **Retrieved Re-ranked In-context (R-BM25)**: q_{max} re-ranked examples using our proposed approach as detailed in § 3.2.

We also compare our results with the state-of-the-art nearest neighbor-based approach for out-of-domain evaluation, kNN-MT (Khandelwal et al., 2020). We use $\lambda = 0.1$, threshold=1.0 and order the examples according to their similarity to the source, with the most similar examples on the left in all our experiments (Appendix Tables 9,10).

5 Results

Table 2 and 3 summarize the main results for the in-domain and the out-of-domain evaluations.

[1]https://github.com/mjpost/sacrebleu
We also report Comet (Rei et al., 2020) scores for evaluating translation quality in Appendix Tables 14 and 15.

Method	$\mathrm{p} + \mathrm{q_{max}}$	En-De	De-En	Ru-En	En-Ru	Avg.
Task-level	$1+0$	23.35	32.16	30.48	25.04	27.75
BM25	$0+1$	19.17	25.82	24.54	21.51	22.76
R-BM25	$0+1$	20.60	28.19	27.26	21.92	24.49
Random (*Baseline*)	$16+0$	24.48	31.26	30.38	25.67	27.95
Task-level	$16+0$	23.72	31.22	30.89	27.27	28.28
BM25	$0+16$	26.58	32.16	31.44	28.54	29.68
R-BM25	$0+16$	27.07	32.59	31.85	28.90	30.10
R-BM25	$0+17$	27.00	32.68	31.88	28.80	30.09
Task-level + R-BM25	$1+16$	**27.09**	**33.24**	**31.90**	**29.50**	**30.43**

Table 2: Results on WMT'19 test sets: Concatenating task-level prompt to R-BM25 consistently achieves the best BLEU scores across the board. p and $q_{\max}$ are the number of task-level and example-specific prompts respectively.

5.1 In-domain Evaluation

A single task-level prompt is competitive with 16 random few-shot examples. Our experiment suggests that it is possible to elicit the task-level knowledge from the large-scale language model using a single prompt as opposed to using 16 random few-shot examples when translating into English (Table 2). Using a single task-level prompt (optimized on the development set) improves BLEU over using 16 random few-shot examples for 2 out of 4 translation directions (De-En, Ru-En). We hypothesize that when translating out of English, the model still benefits from getting exposed to multiple and diverse random few-shot examples as the target language model is relatively weaker.

Multiple example-specific prompts are required to improve translation quality over a single task-level prompt. Using a single task-level ($p = 1$) prompt attains higher BLEU over using a single example-specific prompt ($q = 1$; BM25, R-BM25) across the board. By contrast, using up to 16 BM25 prompts ($q_{\max} = 16$) significantly improves output quality over using task-level prompts, with an average gain of 1.41 in BLEU.

Re-ranking BM25 retreived examples improves BLEU. Our proposed re-ranking strategy consistently improves BLEU across the board over BM25 for both values of $q_{\max} = \{1, 16\}$ showing that both the order and the choice of the in-context examples matters.

Both task-level and R-BM25 examples provide complementary advantages, as combining them using a simple concatenation strategy improves output quality over task-level or R-BM25 examples. We leave the exploration of optimizing the number and the joint order of task-level and example-specific prompts to future work.

5.2 Out-of-domain Evaluation

As XGLM is trained on monolingual Common Crawl snapshots, translation in any domain and language could be considered an out-of-domain task. However, we hypothesize that translation in specific domains like medical, law, or IT could still be challenging for the PLM as the model is less likely to have observed sufficient monolingual datasets for these specialized domains, in contrast to the news text found in WMT. Examples from these domains will require translating rare terminology and carry domain-specific idiosyncrasies, which is known to pose a challenge even for a well-trained supervised neural MT model (Koehn and Knowles, 2017). Hence, we also evaluate PLM under these specialized out-of-domain scenarios.

Domain of few-shot in-context examples matter. Task-level in-context examples drawn from the domain of evaluation, i.e., domain-specific, obtain on an average higher BLEU scores across the board than using examples from a distant WMT corpus as expected (Table 3) in both 1-shot ($p = 1$: +1.4) and 16-shot ($p = 16$: +2.7) settings.

Example-specific prompts significantly improve translation quality over task-level prompts. Unlike the in-domain evaluation, retrieved and re-ranked example-specific prompts (R-BM25) im-

Method	Corpus	$p + q_{max}$	MEDICAL	LAW	IT	KORAN	Avg.
Task-level	Domain-specific	1 + 0	31.23	32.10	28.70	14.68	26.68
	WMT		30.08	31.10	26.72	13.19	25.27
R-BM25	Domain-specific	0 + 1	52.62	55.46	40.54	13.76	40.60
Task-level	Domain-specific	16 + 0	32.65	33.68	28.81	15.30	27.61
	WMT		30.14	30.76	26.19	12.72	24.95
R-BM25	Domain-specific	0 + 16	56.43	**59.57**	46.57	17.49	45.02
R-BM25	Domain-specific	0 + 17	56.65	59.55	46.64	17.48	45.08
Task-level + R-BM25		1 + 16	**56.76**	59.56	**47.50**	**17.55**	**45.34**
*k*NN-MT	-	-	54.35	61.78	45.82	19.45	45.35

Table 3: Results on the Multi-Domain Test Set: Prompting XGLM with R-BM25 in-context examples outperforms *k*NN-MT on 2 out of 4 domains.

prove the translation quality significantly across the board with up to 23 BLEU gain in the Law domain using just a single example as a prompt over a task-level prompt. This can be attributed to the high lexical overlap in the examples retrieved from the training data for these domains (Table 6).

Task-level and R-BM25 prompts are complementary. Both task-level and R-BM25 provide supporting information for a given test source sentence as concatenating these set of prompts improves output quality over using these methods independently, outperforming a strong *k*NN-MT baseline on 2 out of 4 domains (Medical and IT). Where *k*NN-MT utilizes token-level nearest-neighbor inference with representations extracted for bitext using and in combination with a strong supervised MT model to reach the reported translation quality, our approach only uses a sentence-level unsupervised retrieval (BM25) to provide additional context to the unseen source with a multilingual PLM that has not been trained with any known parallel supervision to reach better or comparable translation quality. Hence, our results provide support for further analysis of the translation abilities of retrieval-augmented PLM on new domains and language pairs.

Our manual analysis suggests that the higher gain obtained in the IT domain (+0.86) with both task-level and example-specific prompts can be explained by the observation that for 100 test source sentences, there are no training examples with any lexical overlap with the test source. The task-level prompt can still elicit learned information from the PLM over using no examples for these inputs.

6 Analysis

6.1 Task-level Example Selection

Choice of Few-shot Examples We show the distribution of output quality as measured by BLEU when using 100 different examples as prompts in Figure 2. Across all four language pairs, there is a large variation in BLEU scores (up to 20 BLEU), where noisy or unrelated prompts can lead to significantly worse output quality. Given that most existing parallel corpora are web-crawled and the quality of bitext can vary significantly across different language pairs (Kreutzer et al., 2022), randomly sampled examples can under-estimate the translation quality attainable by prompting the PLM.

	1-shot Prompts	
	100	**1000**
Max	35.82	36.29
Mean	34.06	29.95
Stdev	0.96	9.55
Random 10 trials of best over 100 1-shot Prompts		
Mean over Max	-	36.08
Stdev over Max	-	0.18

Table 4: Task-level example selection from 1000 1-shot Prompts on the WMT'19 development dataset.

Impact of Pool Size on Task-level Prompt Selection We select the best task-level prompt based on the translation quality on the development set from a random sample of 100 examples (pool) as detailed in Section 3.1. However, one concern regarding selecting the best task-level prompt in this

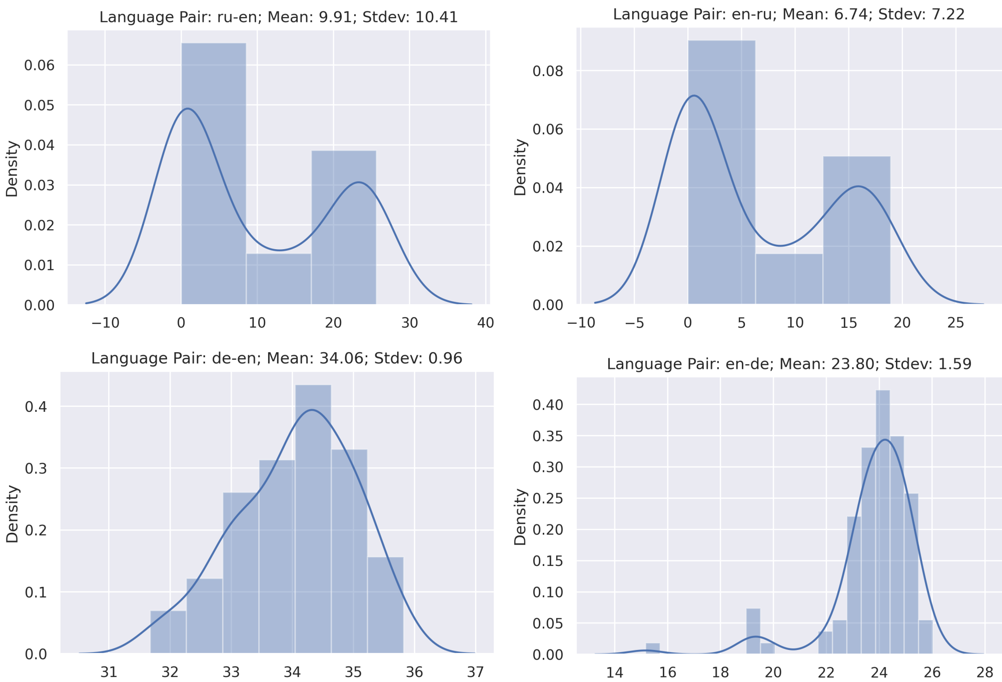

Figure 2: BLEU distribution on the WMT'18 test set for 100 randomly sampled 1-shot prompts from the training dataset. The same set of 100 random 1-shot prompts are used for x→y and y →x translation directions.

fashion could be that we might still be underestimating the PLM (s) performance, as a larger pool size could result in better output quality. We study the impact of using a larger pool size in Table 4 where increasing the number of examples from 100 to 1000 only leads to a gain of 0.5 points in the maximum BLEU. From the same table, we can also observe that for any subset of random 100 few-shot examples, we can extract a task-level prompt (BLEU: 36) with a small standard deviation in overall output quality (0.18).

Features	En-De	De-En	En-Ru	Ru-En
% (Aligned words)				
Random	0.818	0.837	0.594	0.663
Task-level	0.834	0.926	0.773	0.886
Prism-Src				
Random	-1.027	-1.081	-2.214	-1.767
Task-level	-0.843	-0.847	-1.557	-1.206

Table 5: Average scores obtained by top-10 1-best prompts and 10 Random 1-shot prompts (averaged across 3 seeds) on features quantifying semantic equivalence/translation quality (higher is better).

Properties of good Task-level prompts Our manual analysis on the best task-level prompts suggests that any well-formed and meaning-equivalent translation (Vyas et al., 2018; Briakou and Carpuat, 2020) could make a good task-level prompt (see examples in Appendix Table 11). To quantify the meaning equivalence of the 1-best task-level prompt against random 1-shot examples, we report the percentage of aligned words between the source and reference translation ("% Aligned words") using fastAlign (Dyer et al., 2013) and the log probability of generating the reference translation conditioned on the source using a pre-trained multilingual NMT model, Prism-src (Thompson and Post, 2020; Agrawal et al., 2021) in Table 5.[2] Across all language pairs and both metrics, task-level examples achieve higher semantic similarity scores than random 1-shot examples suggesting that task-level examples are relatively more equivalent in meaning than random examples.

Impact of Ordering To investigate the sensitivity of the few-shot prompts ordering on MT quality, we use all possible order permutations of four randomly sampled examples and the top four task-level examples as prompts and report BLEU in Table 7. Task-level prompts are less sensitive to prompt order, as suggested by the lower standard deviation achieved in all settings, and result in higher translation quality than randomly selected examples. Across the three different runs of randomly sampled examples, there is a significant difference in BLEU, further corroborating that the

[2] https://github.com/clab/fast_align, https://github.com/thompsonb/prism

Dataset	Avg. BLEU (I_x, x)	Corr(BLEU ($\hat{y}$, y),BLEU (I_x, x))	Avg. BLEU (I_y, y)	Corr(BLEU ($\hat{y}$, y), BLEU (I_y, y))
Medical	35.785	0.593	32.101	0.777
Law	34.982	0.677	34.349	0.786
IT	25.196	0.497	19.382	0.669
Koran	36.033	-0.016	10.364	0.676

Table 6: Correlation between the degree of overlap as measured by BLEU and the translation quality of the outputs, $\text{BLEU}(\hat{y}, y)$, across different domains when using the top-1 prompt retrieved using BM25. I_x and I_y are the sources and the reference translations in the BM25 examples respectively.

choice of in-context examples and their ordering matters.

	En-De	De-En	En-Ru	Ru-En
	34.43 ±0.25	25.19 ±0.26	12.48 ±5.72	15.56 ±0.50
Random	35.63 ±0.48	25.85 ±0.15	24.99 ±0.21	19.04 ±0.39
	34.73 ±0.30	23.93 ±0.28	10.92 ±4.64	17.91 ±0.07
Optimized	35.95 ±0.24	26.98 ±0.15	25.85 ±0.11	19.96 ±0.24

Table 7: BLEU over all 24 permutations of 3 seeds of 4 randomly selected and top 4 task-level prompts.

6.2 Informativeness of BM25 Examples

To understand the benefit of retrieved examples in the out-of-domain evaluation, we measure the lexical overlap between the test input (x, y) and the prompts (I_x, I_y) using BLEU (Avg. BLEU (I_x, x), Avg. BLEU (I_y, y)), where I_x and I_y are the sources and target translations of the retrieved in-context examples. We also report the correlation against the output translation quality $\text{BLEU}(\hat{y}, y)$. Table 6 shows that the source lexical overlap is a good indicator of the informativeness of a prompt for 3 out of 4 domains, with Koran as an exception. For Koran, while the retrieved sentences have a high overlap with the source (36.03), the target associated with the prompts (I_y) does not get high BLEU with the reference (10.36) compared to other domains. We hypothesize that this might be due to a bias in the reference translations towards a particular output style. We provide examples of this phenomenon in the Appendix Section F.

6.3 Size of the Datastore

Figure 3 shows BLEU when varying the size of the datastore used to retrieve similar in-context examples using BM25 on the Medical dataset. As the size of the datastore increases, the likelihood of retrieving a more similar example increases. However, similar output quality in BLEU can be achieved by using multiple in-context examples when a smaller in-domain datastore is available as multiple examples can provide better coverage of the source terms — BLEU @q=16 with a datastore size of 100k is equivalent to BLEU @q=1 with twice as many examples (200k).

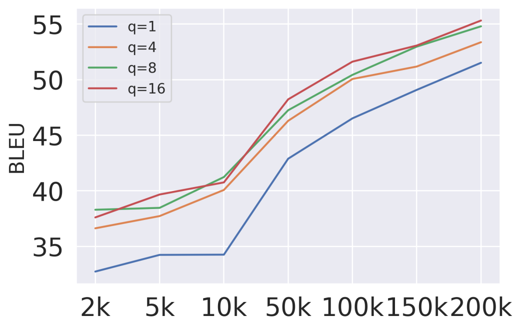

Figure 3: BLEU on the Medical domain when varying the data store size and the number of BM25 examples.

7 Related Work

The selection of in-context examples and their impact on downstream NLP task performance has been studied in prior work for tasks other than MT (Liu et al., 2022b; Lu et al., 2022; Jiang et al., 2020; Min et al., 2022; Zemlyanskiy et al., 2022; Rubin et al., 2021; Liu et al., 2022a). Garcia and Firat (2022) use natural language prompts to control the target language in multilingual MT and investigate the effect of scale, number of languages, and their similarity for this phenomena. Wang et al. (2022) utilize BM25 retrieved training examples in a supervised fashion to learn from similar examples during training. Contrary to prior work, we utilize *similar* examples to form a textual prompt which is used to guide the generation of a translation during inference.

Prior work on domain adaptation for MT uses domain-specific bilingual or monolingual datasets to improve the translation quality of a neural sequence-to-sequence MT model either during training (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Wang et al., 2017) or inference

(Zheng et al., 2021; Khandelwal et al., 2020; Martins et al., 2022). Similar to past work, our work utilizes out-of-domain bitext during inference but instead adapts a PLM on unseen domains. However, our approach does not rely on creating a domain-specific token-level datastore, hence is more compute and memory efficient.

Several concurrent works investigate in-context learning for MT: Zhang et al. (2023) study prompting strategies for MT and examine several factors that could impact translation quality. Garcia et al. (2023) show the effectiveness of using few-shot examples to control translation formality and also corroborates our finding that the quality of the few-shot in-context examples matter. Ghazvininejad et al. (2023) provide control hints to large language models via bilingual dictionaries to improve the translation of rare words. Our work provides both supporting and complementary pieces of evidence to these studies by a) contributing a systematic analysis showing that the impact of the ordering of the demonstration examples on translation quality is dependent upon the nature and the quality of the examples and b) proposing a novel recall-based reranking approach that overcomes the limitations of BM25-based retrieval for in-context examples selection and optimizes for the selection of multiple prompts for MT. To the best of our knowledge, ours is the first work to jointly optimize the selection of multiple prompts for MT either via combining task-level and example-specific prompts or via directly optimizing the joint utility of multiple example-specific prompts by maximizing the coverage of the selected n-grams.

8 Conclusion

We investigate the choice of in-context examples selection for MT in both in-domain and out-of-domain settings. We propose a novel recall-based re-ranking approach to utilize similar training examples as prompts and show their efficacy across multiple datasets and domains. Our findings show that task-level prompts can provide a complementary advantage to example-specific prompts, outperforming a strong *k*NN-MT baseline in 2 out of 4 out-of-domain datasets while being memory and compute efficient. Our manual analysis of the generated outputs reveals that the PLM can mimic the style of the in-context examples provided and can be used for template-based translation synthesis. These results allow future research to evaluate the potential of generating diverse and style-specific outputs for MT.

9 Limitations

We note a few limitations of our work: a) while we systematically investigate the choice of in-context examples for both in- and out-of-domain settings for higher-resource language pairs (English-German, English-Russian), it is unclear how this in-context ability of the PLM varies for the lower-resourced language pairs; b) We only experimented with one pre-trained language model, XGLM. Our preliminary experiments suggested XGLM-7.5B to result in better translation quality than Bloom-7B (Scao et al., 2022) under the same settings. However, further investigation is required to understand how these results vary across different model scales; c) We analyze different orderings for the few-shot task-level prompts but only examine limited sets of ordering (most similar to the left or right) for the example-specific prompts. As the PLM is shown to be sensitive to the ordering of these in-context examples, it remains an open question to study how to best combine the information from multiple example-specific prompts, with prompt ensembling being a viable option, which we leave to future work.

References

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171.

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.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

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.

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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

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*.

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*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. 2022. Towards robust k-nearest-neighbor machine translation. *arXiv preprint arXiv:2210.08808*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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*.

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. *CoRR*, abs/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*.

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.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Pedro Martins, Zita Marinho, and Andre Martins. 2022. Efficient machine translation domain adaptation. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 23–29, 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*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

Harold Somers. 1999. Example-based machine translation. *Machine translation*, 14(2):113–157.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179.

Marco Wrzalik and Dirk Krechel. 2021. CoRT: Complementary rankings from transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4194–4204, Online. Association for Computational Linguistics.

Masaru Yamada. 2011. The effect of translation memory databases on productivity. *Translation research projects*, 3:63–73.

Yury Zemlyanskiy, Michiel de Jong, Joshua Ainslie, Panupong Pasupat, Peter Shaw, Linlu Qiu, Sumit Sanghai, and Fei Sha. 2022. Generate-and-retrieve: Use your predictions to improve retrieval for semantic parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4946–4951.

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

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022. Prompt-based rule discovery and boosting for interactive weakly-supervised learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 745–758, Dublin, Ireland. Association for Computational Linguistics.

Xin Zheng, Zhirui Zhang, Shujian Huang, Boxing Chen, Jun Xie, Weihua Luo, and Jiajun Chen. 2021. Nonparametric unsupervised domain adaptation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4234–4241.

A Statistics of Datasets

Table 8 includes statistics of training, development and test sets used for the experiments discussed in the paper.

Dataset	Train	Dev	Test
WMT-19 (de)	42M	2998	2000
WMT-19 (ru)	10M	3000	2000
Multi-Domain			
Medical	248K	2000	2000
Law	467K	2000	2000
IT	223K	2000	2000
Koran	17K	2000	2000

Table 8: Dataset Statistics.

B Compute Infrastructure & Run time

Each experiment is run on a single Nvidia Tesla V100 Volta GPU machine with 32G Ram. A single inference experiment on 2000 test examples using XGLM with 16 in-context examples takes around 3-4 hrs to complete.

C Results using Second Metric: Comet

We report translation quality using Comet (Rei et al., 2020) in Tables 14 and 15. We use the `eamt22-cometinho-da` model (Rei et al., 2022) to generate the scores as it was shown to achieve higher correlations with human judgments than lexical overlap metrics while being computationally efficient. Our re-ranking strategy (with $q_{max} = 16$) consistently performs the best across the board except for Koran, outperforming strong kNN-MT baselines on the multi-domain test set in 3 out of 4 settings. Adding a task-level prompt to 16 R-BM25 prompts via concatenation further improves quality in 5 out of 8 settings.

D Hyperparameter Search

D.1 Order of BM25 Retrieved Examples

We report BLEU when using two different orderings of example-specific prompts on the development set for the medical domain. Ordering the examples with the most similar examples on the left attains higher BLEU than the right-to-left order. We note that the trend could vary depending on the noise in the training dataset, the degree of similarity, and the number of retrieved examples. We leave the exploration of the ordering of example-specific prompts to future work.

λ	BLEU
Left-to-right	56.84
Right-to-left	54.97

Table 9: BLEU using two different orderings of the top-16 example-specific BM25 prompts on the Medical development Set.

D.2 Choice of λ, Threshold

Table 10 shows the BLEU and the average number of in-context examples selected when varying λ and the threshold described in Section 3.2. We select $\lambda = 0.1$ and threshold value of 1.0 as it achieves the best BLEU on the Medical development set as shown below:

λ	Threshold	BLEU	Avg. # of Examples
0.1	0.1	54.55	14.16
	1.0	**54.56**	12.73
	5.0	53.35	8.83
0.3	0.1	54.47	15.06
	1.0	54.51	14.28
	5.0	53.98	10.32
0.5	0.1	54.44	15.44
	1.0	54.39	15.10
	5.0	54.44	11.85

Table 10: BLEU using different values of λ and threshold on the Medical Development Set ($q_{max} = 16$).

E Example Task-Level Prompts

Table 11 shows the best task-level in-context example selected by our method described in § 3.1 and the respective BLEU scores on the development set for the German-English and Russian-English tasks.

F Output Analysis

We report two interesting findings when prompting PLM with task-level and example-specific prompts:

Stylistic Outputs One advantage of using a single task-level in-context example to prompt the PLM is that it allows us to systematically study how the choice of prompt influences the style of the generated translation. Table 12 illustrates one

German: Beispielsweise der Änderungsantrag zu Artikel 5 in der Stellungnahme des Ausschusses für Landwirtschaft und ländliche Entwicklung weist klar und deutlich darauf hin, dass die Verschlechterung der Qualität des Bodens lokale oder regionale Ursachen und Wirkungen hat und daher unbedingt nationale statt europäischer Maßnahmen ergriffen werden müssen.

English: For example, the amendment to Article 5 in the opinion of the Committee on Agriculture and Rural Development clearly indicates that the degradation of the soil has local or regional causes and effects and it is therefore essential to adopt national as opposed to European measures.

Development BLEU: 35.82

Russian: Если ваш браузер возвращает ранее сохраненный “cookie”, то управляющий им поставщик имеет возможность соединить актуальное посещение пользователя с предыдущими посещениями, но только в отношении своего содержания.

English: If the browser sends back an earlier saved cookie, then the service managing these can connect to the userś earlier visit, but only in respect of their own content.

Development BLEU: 25.63

Table 11: Best task-level prompt For De-En and Ru-En Language Pairs according to the BLEU score on the development set.

such example: we can observe that as the prompt includes a *contraction* (“we are” vs. “we’re”), the outputs generated by the PLM also include contractions and can be incorrectly penalized by BLEU while being meaning equivalent.

Template-based MT Template-based translation in medical, legal, it, or e-commerce domain can be preferable as they reduce the risk of generating errors in automatically generated translations. We present some examples in Table 13 on how PLM can seamlessly use retrieved prompts to synthesize a translation from the template provided.

Prompt: Wegen des heißen Sommers fangen wir erst spät an. = Because of the hot summer, we’re late getting started.
Source: Ja, ich bin sehr zufrieden mit dem Auftritt.
Reference: Yes, I am very happy with the performance.
PLM Output: Yes, I’m very satisfied with the performance.

Source: Es ist eine andere Unternehmenskultur.
Reference: It is a different corporate culture.
PLM Output: It’s a different corporate culture.

Table 12: Outputs mimic the style of the prompt.

Prompt: WIE IST SINGULAIR ANZUWENDEN? = HOW TO TAKE SINGULAIR
Source: WIE IST EVOLTRA ANZUWENDEN?
PLM Output: HOW TO TAKE EVOLTRA

Prompt: Zeigt die aktuelle Datei mit Opera an. = View the current file with Opera.
Source: Zeigt die aktuelle Datei mit Lynx an (Textbasierter Browser).
PLM Output: View the current file with Lynx (Text-based browser).

Table 13: Outputs follow the template of the prompt.

Method	$\mathbf{p + q_{max}}$	**En-De**	**De-En**	**Ru-En**	**En-Ru**
Task-level	$1 + 0$	0.354	0.403	0.428	0.626
BM25	$0 + 1$	0.107	0.149	0.139	0.346
R-BM25	$0 + 1$	0.204	0.249	0.244	0.413
Random-Avg	$16 + 0$	0.387	0.391	0.424	0.636
Task-level	$16 + 0$	0.389	0.381	0.440	0.662
BM25	$0 + 16$	0.423	0.410	0.434	0.673
R-BM25	$0 + 16$	0.438	0.420	0.444	0.677
R-BM25	$0 + 17$	**0.440**	0.421	**0.448**	0.676
Task-level + R-BM25	$1 + 16$	0.434	**0.430**	0.447	**0.694**

Table 14: Comet Scores on WMT'19 test sets.

Method	**Corpus**	$\mathbf{p + q_{max}}$	**MEDICAL**	**LAW**	**IT**	**KORAN**
Results from Jiang et al. (2022)						
Vanilla kNN-MT	-	-	0.548	0.662	0.531	-0.014
Their model	-	-	0.578	0.703	0.585	0.047
Task-level	Domain-specific	$1 + 0$	0.314	0.320	0.240	-0.068
	WMT		0.277	0.345	0.146	-0.113
R-BM25	Domain-specific	$0 + 1$	0.464	0.553	0.389	-0.216
Task-level	Domain-specific	$16 + 0$	0.369	0.365	0.222	**-0.047**
	WMT		0.297	0.399	0.098	-0.131
R-BM25	Domain-specific	$0 + 16$	0.697	0.697	0.666	-0.105
R-BM25	Domain-specific	$0 + 17$	0.699	0.697	0.667	-0.104
Task-level + R-BM25		$1 + 16$	**0.701**	**0.699**	**0.721**	-0.095

Table 15: Comet Scores on the Multi-Domain Test Set.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9

☑ A2. Did you discuss any potential risks of your work?
Section 9 (Limitation 1). In our work, we study and improve the translation ability of large-scale language models for higher resource language pairs only. It still remains an open question on how these abilities transfer to the lower-resourced language pairs. Furthermore, getting reliable and consistent outputs from generative language models is a known problem: https://openreview.net/forum?id=98p5x51L5af. Furthermore,

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, our abstract summarizes the main results and takeaways.

☒ 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?
Section 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☒ 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)?
While large-scale generative models are not directly intended to be used for machine translation and many other downstream NLP tasks, they have been shown to be able to utilize very few examples to perform these tasks. Our work studies this phenomenon and provides analysis and modifications to improve this capability.

☐ 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.?
Section 4

☑ 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.
Appendix Table A, B and Section 4.

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

C ☑ Did you run computational experiments?

Section 4 and Appendix Table B.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2 and Appendix Table B.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix Table D.

☑ 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?
Sections 4, 5 and 6 and Appendix Table C

☑ 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 4.1 (Footnote 1).

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ 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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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.