
Exploring In-Context Learning for LLM Machine Translation for Macedonian

Jana Kuzmanova¹[0009-0009-5015-4183],
Sonja Gievska¹[0000-0002-3411-2399],
Katerina Zdravkova¹[0000-0002-9674-3081]

¹ Ss. Cyril and Methodius University in Skopje,
Faculty of Computer Science and Engineering, North Macedonia,
{jana.kuzmanova|sonja.gievska|katerina.zdravkova}@finki.ukim.mk

Scientific paper

DOI: 10.18485/judig.2026.2.ch22

Abstract

Despite the increasing popularity of large language models for machine translation and their good performance across many language pairs, there is still considerable room for improvement, particularly for low-resource languages and translation into these languages. In-context learning approaches have shown promise in enhancing translation quality in such settings. Providing translations of words from the source text, similar parallel sentences, and grammatical information has resulted in improvements across multiple language pairs. In this paper, we explore whether and how different in-context learning approaches, including parallel sentences and dictionary entries, can improve translation from English into Macedonian. We examine different numbers of examples and different degrees of similarity to the source sentences, phrases, and words to determine which configurations are most effective. Our results show that the effectiveness of in-context learning depends strongly on the model, prompt configuration, and type of additional information provided. Greater example similarity does not consistently improve translation quality, while dictionary-based prompting generally yields poorer results, particularly when out-of-domain entries are used. Although improvements are observed for individual models and translations, all evaluated LLMs perform below the encoder-decoder NLLB baseline across all experimental settings.

Keywords: machine translation, large language models, in-context learning, Macedonian, low-resource languages

1. Introduction

In recent years, LLMs have become increasingly popular for machine translation, sometimes exceeding encoder-decoder models in translation quality, especially for high-resource languages (Zhu et al., 2024). However, they can still struggle with low-resource languages, which are sometimes unseen or other times not as well represented in the training data as higher-resource languages (Pei et al., 2025).

Some approaches to improving low-quality machine translation have emerged in the form of in-context learning strategies that incorporate external knowledge into the prompt, aiming to provide additional information and facilitate more fluent and accurate translation (Pei et al., 2025). Such external knowledge may include dictionaries, parallel examples, grammatical information, or morphological analyses.

In this paper, we apply in-context learning to translation from English into Macedonian to explore which types of information contribute most to translation quality and under which conditions. We use two different prompts, which are subsequently expanded with additional external knowledge. We investigate how the inclusion of dictionary entries and parallel sentences affects the performance of multiple LLMs. Specifically, we examine the effects of varying the number of sentence pairs, the thresholds used for sentence selection, and the ordering of examples; determine which dictionary entries and how many of them yield the greatest improvement; and repeat the experiment using randomly selected sentence pairs as an additional point of comparison alongside the zero-shot baseline. Both in-domain and out-of-domain sources are used for the parallel sentences and dictionary entries.

The remainder of this paper is organised as follows. We first review related work on in-context learning for machine translation, followed by a description of the data, models, and evaluation metrics used in our experiments. We then present the proposed methodology, followed by the experimental results and their discussion.

2. Related Work

Traditional multilingual translation relies on supervised Transformer-based architectures (Vaswani et al., 2017) and cross-lingual representation learning (Conneau et al., 2020; NLLB Team et al., 2022) to handle low-resource and closely related language pairs.

More recently, in-context learning with large autoregressive models has emerged as an alternative paradigm that performs translation via few-shot prompting without model fine-tuning.

One of the early examples of using in-context learning in various NLP tasks, including translation, is given in (Brown et al., 2020), where it’s shown that few-shot performance of LLMs can be improved by scaling them up.

Vilar et al. (2023) study the effectiveness of different prompting strategies for improving machine translation quality using only in-context learning, while Agrawal et al. (2023) focus on example selection.

Moslem et al. (2023) examine the use of in-context learning with in-domain parallel sentence examples across multiple language pairs, encompassing both high- and low-resource languages. Similarly, Zhang et al. (2024a) demonstrate that parallel examples can improve translation quality for unsupported or previously unseen languages.

Other approaches focus on dictionary-level prompt enhancement by providing possible translations of phrases in the text to be translated (Ghazvininejad et al., 2023). This approach also shows improvement over the baseline, being effective for low-resource languages and even in an out-of-domain context.

Some studies have worked on incorporating grammar knowledge in the prompt to help with in-context learning. This includes the addition of grammar books (Hus and Anastasopoulos, 2024), as well as combining the dictionary-based approach with a grammar book and morphological analysis (Zhang et al., 2024b). Miyagawa (2025) also uses a lexicon supplemented with lexical and morphological data to improve translation quality.

Pei et al. (2025) compare these approaches to in-context learning and examine which of these elements lead to the biggest improvement. However, they focus on translation from a low-resource language into English. Our work takes the opposite direction, providing a more granular examination of the most effective in-context learning configurations for machine translation, initially for Macedonian and subsequently for several South Slavic languages.

Our previous research on the machine translation capabilities from English to Macedonian for multiple commercially available tools (Kuzmanova and Zdravkova, 2026) showed there’s room for improve-

ment, and in this paper we examine possible avenues for improvement on smaller models that can later be applied more broadly.

3. Data and Models

The experiments presented in this paper focus primarily on Macedonian. One of the datasets used is MULTEXT-East (Erjavec, 2017), which includes George Orwell’s novel 1984 and its translations into multiple languages, including Macedonian. The dataset contains more than 6,000 English–Macedonian parallel sentence pairs, of which 670 are used for evaluation and 670 for validation, while the remainder serve as a source of examples added to the prompts. MULTEXT-East constitutes the main dataset used to evaluate the approaches presented in this paper.

We also use an additional dataset from the news domain, obtained from the MaCoCu project (Bañón et al., 2022), to evaluate the system under more realistic conditions in which the text to be translated is not available as a source of in-context examples. The dataset contains almost 500,000 parallel pairs of news headlines, which were filtered and aligned to construct a dictionary for use in the experiments. The dictionary construction process is described in greater detail below. In the present experiments, the dataset is used exclusively for this purpose.

Experiments were conducted using multiple LLMs, including models that do not officially support Macedonian. These include Llama-3.1-8B (Meta Llama, 2024), EuroLLM-9B (UTTER, 2024), Qwen2.5-7B (Qwen Team, 2024), and Mistral-Nemo-2407 (12B parameters) (Mistral AI Team, 2024). The models were chosen for their relatively small size and free availability. We considered the possibility of expanding towards larger and even commercial models in our future experiments, however, that is beyond the scope of this paper.

3.1. Dictionary construction

As no freely available, non-commercial, open-access dictionary was available for the experiments involving dictionary-based hints rather than complete sentences, we constructed two dictionaries from the available datasets.

For the MaCoCu news dataset, we first assessed the quality of the translations using COMETKiwi, a reference-free quality estimation metric. Since the dataset consists of news headlines, some parallel

pairs convey different meanings despite referring to reports on the same topic. We therefore filtered the dataset, retaining only those pairs for which COMETKiwi assigned a score greater than 0.8. This reduced the original dataset from 478,059 to 359,235 sentence pairs.

These sentences were then aligned at the phrase level using awesome-align (Neuraz, 2022). The model automatically identifies corresponding units in the source and target sentences and aligns them to produce a parallel phrase dictionary. To convert these alignments into plain-text pairs, we construct a graph connecting the indices of the aligned units in the source and target sentences. This enables efficient many-to-many mapping and accommodates all types of alignments: single word to single word, single word to multi-word phrase, multi-word phrase to single word, and multi-word phrase to multi-word phrase. The initial set of parallel dictionary entries contained 5,568,820 pairs, which was reduced to 1,718,490 after duplicates were removed. Finally, we assessed the quality of the remaining entries using COMETKiwi and retained only those with a score above 0.75. A lower threshold than that used at the sentence level was selected based on empirical observations that some correct phrase-level translations received scores below 0.8. The resulting dictionary constructed from the news dataset contains 263,351 phrase pairs.

We used the same approach to construct a dictionary from the MULTEXT-East dataset. Since this is a manually verified parallel dataset, the initial filtering step was not required. We applied awesome-align to the training split of the original MULTEXT-East Macedonian dataset and aligned the words and phrases in each sentence as described above. The resulting dictionary contains 8,968 phrase pairs.

4. Methodology

We use two different base prompts, which are then enhanced with various in-context examples. The two base prompts are a simple translation request with a completion template, as well as a prompt following the example provided for the WMT (Conference on Machine Translation) Shared Translation Task (Kocmi et al., 2025), adapted to the used languages and domain, and with an added completion template at the end. The prompts were chosen from a larger pool based on a zero-shot experiment using the validation set. The final

two prompts we used for the baseline and further extended in the experiments are the following:

- Simple completion prompt:

Translate the following sentence from English to Macedonian.

English: {SOURCE SENTENCE}

Macedonian:

- WMT-based hybrid prompt:

You are a professional English-to-Macedonian translator, tasked with providing translations suitable for use in Macedonia (mk_MK).

Your goal is to accurately convey the meaning and nuances of the original English text while adhering to Macedonian grammar, vocabulary, and cultural sensitivities. The original English sentence is from a literary text.

Aim to maintain the original tone and register, retaining the emotional depth of the story.

Dialogues should sound natural and follow the conventions of the target language. Produce only the Macedonian translation, without any additional explanations or commentary.

Please translate the following English sentence into Macedonian (mk_MK):

English: {SOURCE SENTENCE}

Macedonian:

The experiments focus on determining the models' ability to learn from in-context examples and achieve a higher translation quality as a result. The examples are added to the prompt before requesting the final translation of the sentence. We experimented with the following possible enhancements for in-context learning:

- Randomly selected parallel sentences. These sentences are drawn from the corpus and added to the prompt, which is then used to evaluate the entire test set. The same randomly selected examples are used for all test sentences to avoid variation attributable to example selection. We vary the number of examples provided to the model from one to five. The examples are added incrementally, such that the sentences used in the one-shot setting form a subset of those used in the subsequent settings, up to the five-shot experiment. The prompts described above are adapted by inserting the examples before the final question and completion template, with each example explicitly delimited. The following illustrates the resulting prompt for the two-shot setting:

Examples:

English: From her breathing it was evident that she was going off to sleep again.

Macedonian: Според ритамот на дишењето беше јасно дека таа повторно тоне во сон.

English: It was very heavy in his pocket, but fortunately it did not make much of a bulge.

Macedonian: Беше многу тежок во неговиот џеб, но за среќа не правеше големо испупчение.

- Parallel examples selected using embedding-based similarity. For each source sentence, we select the K most similar sentences from the training set using multilingual MPNet embeddings (Song et al., 2020). The selected examples are ordered from least to most similar, placing the most similar example closest to the final prompt to account for the models' potential recency bias. Unlike the random selection approach, each source sentence is therefore paired with a different set of examples. As in the random experiments, we vary the number of examples from one to five while retaining the same prompt format. The examples are inserted into the base prompts before the final completion prompt and separated using the same delimiter as above. We restrict the candidate examples to source sentences containing more than five words to reduce the likelihood that the exact sentence to be translated is provided as an example.
- Parallel examples selected using an embedding-based similarity threshold. We use the same multilingual embedding model to compute the cosine similarity between each test sentence and the potential example sentences. In this setting, we do not consider the ordering of the examples but include all sentence pairs for which the cosine similarity between the source test sentence and the source training sentence exceeds a specified threshold. We evaluate four similarity thresholds: 0.6, 0.7, 0.8, and 0.9. As above, we use the filtered training dataset containing only source sentences with more than five words. If no sentences exceed the specified threshold, no examples are added and the resulting prompt is equivalent to the base prompt described above.
- Dictionary entries for words in the sentence being translated. We use the dictionaries constructed from the MULTEXT-East and MaCoCu datasets, as described in the previous section. The experiments vary both the source dictionary (MULTEXT-East or

MaCoCu) and the method used to select candidate dictionary hints. The prompts are constructed by replacing the parallel sentence examples used in the previous experiments with the selected dictionary hints, while retaining the final question and completion prompt. The general format of these prompts is as follows:

```
{BASE PROMPT}
  Use the following dictionary hints:
  {English phrase 1}={Macedonian phrase 1}
  {English phrase 2}={Macedonian phrase 2}
{FINAL PROMPT}
```

We compared two search strategies for the dictionary entries to include - embedding-based versus string-matching based search strategies.

- For the string-matching strategy, we first extract all n-grams containing one to three words from the test sentence. We then search for these n-grams in the dictionary and include the corresponding entries as dictionary hints only when an exact match is found.
- The embedding-based strategy uses the same MPNet embeddings as those employed for the sentence similarity calculations. We first extract the n-grams from the test sentence and then compute their embeddings. We experiment with different inclusion thresholds, using cosine similarity values of 0.6, 0.7, 0.8, and 0.9. If the cosine similarity between an n-gram from the test sentence and a phrase from the dictionary exceeds the specified threshold, the corresponding dictionary entry is included as a hint for that sentence.

All in-context learning strategies were compared with one another and against two baselines: zero-shot prompting and the encoder-decoder model NLLB (No Language Left Behind) ([Facebook, 2022](#)). This comparison was used to assess the effectiveness of each strategy under different parameter settings. We used the same generation parameters across all experiments to ensure a consistent setup across the different models. The maximum number of new tokens was set to 100 and the temperature to 0.2, while the `no_repeat_ngram_size` and `top_p` parameters were set to 3 and 0.9, respectively. We used the AutoTokenizer associated with each model, with left padding.

We postprocessed the outputs using a simple rule-based approach, looking for the first output that’s not in English following

the prompt and that does not contain any additional commentary or artifacts. However, this approach still let through a considerable amount of clutter and is subject to improvement in future work.

5. Results and Discussion

In this section, we present the results of the experiments described above and discuss possible reasons why, in most cases, the proposed approaches do not outperform the baseline methods. Although the hybrid prompt generally achieves higher scores than the simple prompt, all tested LLMs across all experimental settings perform below the NLLB baseline. The translation quality is evaluated using multiple evaluation metrics, including lexical metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015) using their implementation in SacreBLEU (Post, 2018), as well as the learned metrics BERTscore (Zhang et al., 2020), and COMET (Rei et al., 2022a), and CometKiwi (Rei et al., 2022b).

5.1. Baseline

The baseline results for the completion prompt are presented in Table 1, while those for the hybrid prompt are shown in Table 2. In most cases, the hybrid prompt outperforms the completion prompt for the same model. A notable exception is Mistral, whose scores decrease substantially when the hybrid prompt is used. Further inspection reveals that this decrease is primarily attributable to the large number of empty responses generated by Mistral with this prompt. Excluding empty responses and other invalid outputs, such as unrelated text, explanations, notes and comments, and punctuation-only responses, has little effect on the baseline scores of most models. In most cases, no more than 50 responses are excluded, resulting in negligible changes to the final scores. With the hybrid prompt, however, Mistral produces more empty than non-empty responses. When these invalid responses are excluded, its scores increase substantially, reaching a BLEU score of 20.13, a chrF score of 47.28, and a COMET score of 0.803.

Looking at the outputs, it’s notable that the responses by EuroLLM and often Qwen are not in Macedonian. These models don’t officially support the language, and in most cases output responses in Bulgarian, with Qwen occasionally also returning other Slavic languages as well as Chinese, which can be noticed by the lower metric scores, especially for the lexical metrics. Llama, while also

Table 1: Baseline results for the completion prompt.

<i>Model</i>	<i>BLEU</i>	<i>chrF</i>	<i>BERTScore</i>	<i>COMET</i>	COMETKiwi
EuroLLM	8.6565	28.9125	0.7712	0.6387	0.6417
Llama	13.3992	39.1519	0.7975	0.7327	0.7226
Mistral	14.9326	37.9522	0.7799	0.7293	0.7341
NLLB	22.1728	50.3946	0.8463	0.7964	0.7785
Qwen	9.3149	29.2926	0.7573	0.5534	0.5623

Table 2: Baseline results for the hybrid prompt.

<i>Model</i>	<i>BLEU</i>	<i>chrF</i>	<i>BERTScore</i>	<i>COMET</i>	COMETKiwi
EuroLLM	9.1908	31.3964	0.7954	0.7507	0.7906
Llama	13.7898	39.0105	0.7839	0.7158	0.7097
Mistral	9.3539	21.0759	0.3726	0.5368	0.5338
NLLB	22.1728	50.3946	0.8463	0.7964	0.7785
Qwen	9.3269	30.8134	0.7646	0.5779	0.5772

not officially supporting Macedonian, does mostly return Macedonian translations of the given source sentences, and so does Mistral.

5.2. Random Examples

The experiment using random examples reveals differences in the behavior depending on the model and prompt used. While EuroLLM and Qwen improve with the addition of random examples for the completion prompt, their scores decline or stay flat for the hybrid prompt. On the other hand, Llama in particular declines sharply, and Mistral only shows improvement when using a single example.

Table 3 shows the results for the completion prompt. In order to improve readability, we present only the results for one, three, and five examples and only for a selected number of metrics; however, the rest of the results follow similar trends.

As mentioned previously, scores for EuroLLM and Qwen improve under these conditions, while Mistral stays flat. Llama, on the other hand, shows a significant decline in all metric scores. Inspecting the results shows that this is due to an increased number of empty or invalid responses. These are increased ten times across all five experiments using random sentences. While a closer analysis of the translations where both the baseline and the random experiment

Table 3: chrF, COMET, and CometKiwi scores for one, three, and five random examples for the completion prompt.

<i>model</i>	<i>examples</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	1	31.8528	0.7066	0.7438
	3	32.0792	0.7007	0.7379
	5	31.8434	0.7027	0.7335
Llama	1	33.1588	0.6449	0.6377
	3	39.9238	0.7196	0.7051
	5	38.8531	0.7056	0.6876
Mistral	1	44.8487	0.7918	0.7791
	3	45.7523	0.7949	0.7798
	5	45.2762	0.7923	0.7795
Qwen	1	30.2313	0.586	0.5829
	3	30.9666	0.5818	0.5791
	5	29.8860	0.5724	0.5719

returned a valid output shows small improvements in grammar or meaning in select sentences, this is also true in the other direction. The opposite is true for Qwen, where adding the random examples often stabilizes the responses and the language thereof. While other Slavic languages are still common, it does get closer to Macedonian grammar and vocabulary. In the case of EuroLLM, the change in score comes down to actual changes in the translation and its quality for most of the test sentences.

However, these improvements disappear for the hybrid prompt. These results are presented in Table 4, again, only showing the scores for one, three, and five examples and select metrics. This table also excludes the Llama scores, since they follow the same trends as the results for the completion prompt.

An analysis of the translations under this configuration reveals an increased number of non-relevant responses for EuroLLM, as well as actual changes in the translations. A notable case here is Mistral, which shows a significant improvement over the baseline for the hybrid prompt, in large part owed to filling in a significant number of empty responses. However, there are occasional improvements in the translation quality as well. This advantage does not hold for a larger number of examples. As the number of examples increases, the number of filled-in empty examples drops and sometimes new empty responses are introduced, which keeps the metric scores flat or

Table 4: chrF, COMET, and CometKiwi scores for one, three, and five random examples for the hybrid prompt.

<i>model</i>	<i>examples</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	1	32.7646	0.6979	0.7311
	3	32.377	0.6882	0.7213
	5	32.612	0.6961	0.7287
Mistral	1	38.1339	0.7241	0.7148
	3	12.1235	0.4475	0.4449
	5	15.0146	0.4788	0.4769
Qwen	1	30.5758	0.5843	0.5833
	3	30.7558	0.5822	0.5806
	5	31.264	0.5835	0.5841

decreased in comparison with the baseline. Looking at only the clean, non-empty responses, shows that Mistral does show improvement, with the hybrid prompt reaching as far as 0.823 for COMET and 50.63 for chrF.

5.3. Similar Examples - K-shot Selection

In this subsection we present the results of the similarity-based K-shot experiment. We again only present the results for the same numbers of examples and the same metrics as in the previous sections. We only present the results for the hybrid prompt, as in this case, the results for both prompts follow the same trends.

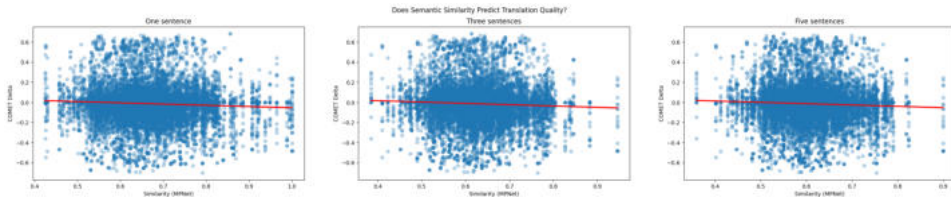
The results are shown in Table 5. Surprisingly, they show the same, and sometimes worse, difference in relation to the baseline when compared to the randomly-chosen examples. In this case as well, EuroLLM and Qwen show improvement over the completion baseline, but fail to improve on the hybrid prompt baseline.

We also examined the relationship between the change in COMET score relative to the baseline and the average cosine similarity between the source sentence and the selected examples for the one-, three-, and five-shot settings. The results are shown in Figure 1. No strong relationship between translation quality and example similarity is observed. The slight negative trend may be attributable to the smaller number of examples available at higher similarity levels, as the shortest candidate sentences were removed from the training set. Nevertheless, owing to the style and repetitive nature of the novel,

Table 5: chrF, COMET, and CometKiwi scores for one, three, and five most similar examples for the hybrid prompt.

<i>model</i>	<i>examples</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	1	33.7883	0.7128	0.7432
	3	35.0321	0.6956	0.7192
	5	34.6251	0.6841	0.7089
Llama	1	32.9343	0.6499	0.6435
	3	38.0608	0.6973	0.6839
	5	37.6595	0.6813	0.6703
Mistral	1	30.037	0.6309	0.6206
	3	21.351	0.5457	0.5369
	5	17.0467	0.498	0.494
Qwen	1	32.0882	0.5755	0.5765
	3	32.7297	0.5705	0.567
	5	32.052	0.5708	0.5719

some selected examples remain highly similar or even identical to the source test sentence. Interestingly, such high similarity does not necessarily result in the corresponding translation being reproduced, and these translations often exhibit issues similar to those observed in the baseline outputs.

Figure 1: Relationship between the COMET score difference and average similarity across one, three, and five example sentences.¹

The largest decreases in COMET score at high similarity levels are observed for Mistral, which frequently produces no response in cases where the baseline produces a valid output. Llama occasionally exhibits similar behaviour. However, the largest increases in COMET score after excluding blank or otherwise invalid responses are distributed across different levels of example similarity, although they generally occur at similarities above 0.5. Manual analysis likewise reveals no clear relationship between the cosine similarity of the examples and translation quality. This observation should, however, be

interpreted with caution, as most candidate sentences exhibit moderate similarity to the test sentences, while relatively few are highly similar.

5.4. Similar Examples - Threshold Selection

Subsequently, we focused on selecting sentences similar to the source sentence as in-context examples, using a similarity threshold to determine whether an example should be included. If no sentence in the training set exceeded the specified threshold, no examples were added to the prompt. As the similarity threshold increased, the number of test sentences for which suitable examples were identified decreased substantially: 523 sentences at a threshold of 0.6, 217 at 0.7, 50 at 0.8, and only 14 at 0.9.

The evaluation results for this experiment are presented in Table 6. As expected, the results obtained with a similarity threshold of 0.9 remain largely unchanged compared with the baseline, since only 14 test sentences received additional examples. The only notable exception is Mistral, which achieves higher scores, apparently owing to the smaller number of blank responses produced in this experimental setting.

However, the results at lower similarity thresholds suggest that the inclusion of examples is more likely to introduce noise than to improve translation quality. This effect is particularly evident at the lowest threshold of 0.6, at which nearly 80% of the test sentences receive at least one example. At this threshold, scores decrease substantially for most models, with the exception of Mistral, which improves, and Qwen, whose performance remains largely unchanged. Nevertheless, some of the largest sentence-level improvements in COMET score are also observed at the 0.6 threshold, both when invalid responses are included and when they are excluded. As in the previous experiments, the largest decreases are primarily associated with blank or otherwise invalid responses and with less predictable output formatting, which increases the likelihood of post-processing errors. Although such cases occur across all similarity thresholds, they are more frequent at thresholds of 0.6 and 0.7.

5.5. Dictionary examples

Finally, the last experiment involved trying to insert dictionary hints instead of full sentences. These experiments were the least

Table 6: chrF, COMET, and CometKiwi scores for similar examples above different thresholds for the hybrid prompt.

<i>model</i>	<i>threshold</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	0.6	33.0196	0.6946	0.7234
	0.7	31.8650	0.7288	0.763
	0.8	31.1883	0.7419	0.7795
	0.9	30.7656	0.7493	0.7881
Llama	0.6	34.1631	0.6381	0.6269
	0.7	36.0115	0.6715	0.665
	0.8	37.114	0.6984	0.6948
	0.9	39.2967	0.718	0.7167
Mistral	0.6	25.7477	0.5876	0.5808
	0.7	33.7045	0.6682	0.6631
	0.8	34.9579	0.6872	0.6828
	0.9	34.9184	0.685	0.6801
Qwen	0.6	31.8799	0.5767	0.5768
	0.7	31.3195	0.5722	0.5791
	0.8	30.846	0.5712	0.5749
	0.9	30.584	0.5722	0.5766

successful, leading to a decrease in translation quality across the board.

Table 7 presents the results obtained using the string-based matching strategy, in which dictionary entries were included only when an exact match for a word or phrase was found in the corresponding dictionary. We evaluated the two dictionaries separately. The dictionary constructed from the MULTEXT-East corpus resulted in either a smaller decrease or little change in translation quality compared with the baseline. However, this setting does not fully reflect realistic translation conditions, particularly for literary texts, as a domain-specific dictionary containing vocabulary from the text being translated may not be available. In contrast, adding entries from the dictionary constructed from the MaCoCu news corpus leads to a marked decrease in translation quality.

This approach is most likely to introduce noise and lead to nonsensical translations, especially when the news dictionary is used, even when the dictionary hints are appropriate and accurately translated. However, using a dictionary also introduces different word forms, which often leads to a decline in the grammatical proficiency in

Table 7: chrF, COMET, and CometKiwi scores for dictionary hints for string-based matching for the hybrid prompt.

<i>model</i>	<i>threshold</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	MTE	33.0042	0.7274	0.7649
	news	32.189	0.679	0.7138
Llama	MTE	39.2196	0.694	0.6912
	news	35.5304	0.6258	0.6221
Mistral	MTE	17.0394	0.4984	0.4902
	news	15.6766	0.4871	0.4828
Qwen	MTE	31.5857	0.5667	0.5692
	news	28.6973	0.5021	0.4981

the machine translation if the model decides to use the word as-is. The biggest wins are surprisingly not sentences containing book-specific words as expected, but improved word choice in shorter sentences. However, these improvements aren’t enough to offset the decline caused by the confusion that happens in other cases.

The experiment using an embedding-based strategy to select words showed an even bigger decline in translation quality, likely due to introducing even more noise. Table 8 shows these results for the Multext-East dictionary, only for 0.6 and 0.9 as cosine similarity thresholds for inclusion of the selected words and phrases. We omitted the results for the news dictionary and the other thresholds for readability, as they follow the same trends. In both cases, translation quality decreases compared to the baseline, with little difference in estimated quality between the different thresholds.

There are rare wins in translation quality here resulting in improved word choice, but especially for the 0.6 threshold, some dictionary hints lead to inadequate words being chosen. One such example is the word "snap" being translated as "експлозија" (explosion) in the version including the dictionary hints. There are also many case of increased noise, nonsensical and ungrammatical translations, and a larger number of letters outside of the Macedonian alphabet appearing in the translations, the latter particularly being the case for Qwen.

Table 8: chrF, COMET, and CometKiwi scores for dictionary hints for embedding-based matching for the Multext-East dictionary and the hybrid prompt.

<i>model</i>	<i>threshold</i>	<i>chrF</i>	<i>COMET</i>	<i>CometKiwi</i>
EuroLLM	0.6	34.9268	0.6562	0.6703
	0.9	35.1305	0.6612	0.6771
Llama	0.6	36.6363	0.6243	0.6215
	0.9	36.7682	0.6252	0.6176
Mistral	0.6	21.121	0.5348	0.5266
	0.9	21.4067	0.5359	0.5264
Qwen	0.6	31.2155	0.5241	0.5175
	0.9	31.2293	0.5218	0.5188

6. Limitations and Future Work

One limitation identified during the experiments concerns the robustness of the post-processing procedure, which was not always able to accommodate variations in model output across different experimental settings. This occasionally resulted in valid responses being processed incorrectly and consequently affected the evaluation scores. Future work will therefore focus on developing a more robust post-processing approach that extends beyond the use of simple rules.

Additionally, owing to time and hardware constraints, we were unable to evaluate the proposed approaches using larger models, which may respond differently to the introduced modifications. Future work will therefore extend the evaluation to larger and potentially more capable models.

Manual analysis also revealed cases in which higher-quality translations received relatively low scores, even when evaluated using established metrics such as COMET. Future work will therefore complement automatic evaluation with a more systematic and rigorous manual assessment of translation quality, as well as additional prompting techniques and using LLM as a judge.

7. Conclusion

This study investigated the effectiveness of several in-context learning strategies for machine translation from English into Macedonian, focusing on relatively small LLMs and a language that is not officially supported by several of the evaluated models. We examined

two base prompts and enriched them with randomly selected parallel sentences, similarity-based examples selected using either a fixed number of examples or similarity thresholds, and dictionary entries obtained through string- and embedding-based matching. The results demonstrate that the effectiveness of in-context learning in this setting is highly dependent on the model, prompt configuration, and type of external knowledge provided.

Overall, the hybrid prompt generally performed better than the simple completion prompt, although all evaluated LLMs remained below the NLLB baseline across the experimental settings. Adding in-context examples occasionally improved individual translations and, for some models, helped stabilise the generated language. However, these improvements were not consistent across models or experimental configurations. In particular, selecting examples with greater semantic similarity did not systematically lead to higher translation quality. At higher similarity thresholds, suitable examples became increasingly scarce, whereas lower thresholds increased coverage but also tended to introduce more noise. These findings indicate a trade-off between the relevance and availability of in-context examples rather than a straightforward relationship between example similarity and translation quality.

Dictionary-based prompting produced the least favourable overall results. Entries derived from the in-domain MULTEXT-East corpus generally caused smaller performance decreases than those obtained from the out-of-domain MaCoCu news corpus, suggesting that domain compatibility is important when lexical information is incorporated into the prompt. Nevertheless, even appropriate dictionary entries could introduce unsuitable word forms or lexical choices, resulting in less grammatical or semantically inappropriate translations. Embedding-based dictionary selection further increased this problem by introducing additional potentially irrelevant lexical information.

More broadly, the experiments show that in-context information is not inherently beneficial for machine translation with smaller LLMs, particularly when translating into an unsupported or low-resource language. Additional context can improve particular translations, but it can also increase output instability, invalid responses, formatting inconsistencies, and sensitivity to irrelevant information. The results therefore suggest that the effectiveness of in-context learn-

ing depends not only on the amount or similarity of the information provided, but also on its quality, domain relevance, and the model’s ability to use it reliably. Future work involving larger models, more robust post-processing, and systematic human evaluation will help determine whether these limitations are primarily associated with model capacity or represent broader challenges for in-context machine translation into low-resource languages.

Language Resources

Bañón, Marta and Esplà-Gomis, Miquel and Forcada, Mikel L. and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Ortiz Rojas, Sergio and Ramírez-Sánchez, Gema and Rupnik, Peter and Santi, Vít and Steingrímsson, Steinþór and Toral, Antonio. 2022. *MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages*. European Association for Machine Translation. PID <https://macocu.eu/>.

Tomaž Erjavec. 2017. Multext-east. *Handbook of linguistic annotation*, page 441–462.

Facebook. 2022. *NLLB*. Hugging Face. PID <https://huggingface.co/facebook/nllb-200-3.3B>. Hugging Face model card.

Meta Llama. 2024. *Llama-3.1-8B-Instruct*. Meta. Hugging Face. PID <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Hugging Face model card.

Mistral AI Team. 2024. *Mistral-Nemo-Instruct-2407*. Mistral AI. Hugging Face. PID <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>. Hugging Face model card.

Antoine Neuraz. 2022. *AWESOME: Aligning Word Embedding Spaces of Multilingual Encoders*. Hugging Face. PID <https://huggingface.co/aneuraz/awesome-align-with-co>. Hugging Face model card.

Post, Matt. 2018. *A Call for Clarity in Reporting BLEU Scores*. Association for Computational Linguistics. PID <https://github.com/mj-post/sacrebleu>.

Qwen Team. 2024. *Qwen2.5-7B-Instruct*. Qwen. Hugging Face. PID <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>. Hugging Face model card.

- Rei, Ricardo and C. de Souza, José G. and Alves, Duarte and Zerva, Chrysoula and Farinha, Ana C and Glushkova, Taisiya and Lavie, Alon and Coheur, Luisa and Martins, André F. T. 2022a. *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*. Association for Computational Linguistics. PID <https://huggingface.co/Unbabel/wmt22-comet-da>.
- Rei, Ricardo and Treviso, Marcos and Guerreiro, Nuno M. and Zerva, Chrysoula and Farinha, Ana C and Maroti, Christine and C. de Souza, José G. and Glushkova, Taisiya and Alves, Duarte and Coheur, Luisa and Lavie, Alon and Martins, André F. T. 2022b. *CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task*. Association for Computational Linguistics. PID <https://huggingface.co/Unbabel/wmt22-cometkiwi-da>.
- Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu. 2020. *MPNet: Masked and Permuted Pre-training for Language Understanding*. PID <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>. Hugging Face model card.
- UTTER. 2024. *EuroLLM-9B-Instruct*. UTTER. Hugging Face. PID <https://huggingface.co/utter-project/EuroLLM-9B-Instruct>. Hugging Face model card.
- Tianyi Zhang and Varsha Kishore and Felix Wu and Kilian Q. Weinberger and Yoav Artzi. 2020. *BERTScore: Evaluating Text Generation with BERT*. PID https://github.com/Tiiiger/bert_score.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. *In-context Examples Selection for Machine Translation*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christo-

- pher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#). pages 8440–8451.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation](#).
- Jonathan Hus and Antonios Anastasopoulos. 2024. [Back to School: Translation Using Grammar Books](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouagna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinhór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. [Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Jana Kuzmanova and Katerina Zdravkova. 2026. [A Pilot Study of Multi-method Evaluation of Machine Translation in Macedonian](#). *Comput. Sci. Inf. Syst.*, 23(2):827–860.
- So Miyagawa. 2025. [RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI](#). In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.

- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive Machine Translation with Large Language Models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No Language Left Behind: Scaling Human-Centered Machine Translation](#).
- 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.
- Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. [Understanding In-Context Machine Translation for Low-Resource Languages: A Case Study on Manchu](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.
- Maja Popović. 2015. [ChrF: Character n-gram F-score for Automatic MT Evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you Need](#). *Advances in neural information processing systems*, 30.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. [Prompting PaLM for Translation:](#)

- Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.
- Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. [Teaching Large Language Models an Unseen Language on the Fly](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.
- Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. [Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Istraživanje učenja u kontekstu za mašinsko prevođenje na makedonski jezik pomoću velikih jezičkih modela

Jana Kuzmanova, Sonja Gievska, Katerina Zdravkova

Sažetak

Uprkos sve većoj popularnosti velikih jezičkih modela u mašinskom prevođenju i njihovim dobrim rezultatima na brojnim jezičkim parovima, i dalje postoji značajan prostor za poboljšanje, posebno kada je reč o jezicima sa ograničenim resursima i prevođenju na te jezike. Pristupi zasnovani na učenju u kontekstu pokazali su potencijal za unapređenje kvaliteta prevoda u takvim uslovima. Uključivanje prevoda reči iz izvornog teksta, sličnih paralelnih rečenica i gramatičkih informacija dovelo je do poboljšanja rezultata na više jezičkih parova. U ovom radu istražujemo da li i na koji način različiti pristupi učenja u kontekstu, uključujući paralelne rečenice i rečničke odrednice, mogu poboljšati prevođenje sa engleskog na makedonski jezik. Ispitali smo različit broj primera i različite stepene njihove sličnosti sa izvornim rečenicama, frazama i rečima kako bismo utvrdili koje su konfiguracije najefikasnije. Naši rezultati pokazuju da efikasnost učenja u kontekstu u velikoj meri zavisi od modela, konfiguracije upita i vrste dodatnih informacija. Veća sličnost primera ne dovodi dosledno do boljeg kvaliteta prevoda, dok upiti zasnovani na rečničkim odrednicama uglavnom daju slabije rezultate, naročito kada se koriste odrednice iz drugog domena. Iako su kod pojedinačnih modela i prevoda zabeležena poboljšanja, svi testirani veliki jezički modeli postigli su rezultate ispod NLLB referentnog sistema u svim eksperimentalnim postavkama.

Ključne reči: mašinski prevod, veliki jezički modeli, učenja u kontekstu, jezici sa ograničenim resursima

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

