Every Word Presented in Context: Syntactic Coverage as Objective for Low-Resource Machine Translation with Large Language Models

Samuel Frontull, Thomas Ströhle
Department of Computer Science / University of Innsbruck
Technikerstraße 21a, 6020 Innsbruck, Austria
{samuel.frontull,thomas.stroehle}@uibk.ac.at

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities in multilingual machine translation. However, they underperform for low-resource languages, indicating the need for more explicit instructional guidance. In this work, we introduce Fragment-Shot Prompting, a novel few-shot prompting method that aims to retrieve examples for every word occurring in the sentence to be translated, illustrating their use and meaning in context. We evaluate our method on translation between Italian, Ladin (Val Badia) and Ladin (Gherdëina) and compare its performance with zero-shot prompting, random few-shot prompting, as well as established lexical and semantic retrieval strategies. We conduct these experiments using state-of-the-art LLMs, including GPT-3.5, GPT-4o, o1-mini, LlaMA-3.3, and DeepSeek-R1. Our results demonstrate that LLMs can extract substantial value from limited data when translating from a low- to the high-resource language. However, this does not apply to translations into the low-resource languages, where the prompting method plays a much more important role. In particular, our method consistently delivers the best results and enables significant gains. Even though translation performance into Ladin remains limited with the available resources, our results highlight the importance of syntactic coverage for improving translation accuracy and variant-specific adaptation in low-resource scenarios.

Keywords: Less-Resourced/Endangered Languages, Machine Translation, Natural Language Generation

1. Introduction

In recent years, Large Language Models (LLMs) have made significant advancements in machine translation (MT), delivering state-of-the-art performance for high-resource languages (Zhang et al., 2023). However, effective language modeling by LLMs can only be enabled with sufficient (training) data. For low-resource languages, the lack of such data makes this data-intensive process impractical. In particular, LLMs have limited awareness of (different variants of) smaller languages and struggle to produce coherent output (Court and Elsner, 2024; Ondrejová and Šuppa, 2024). This is also evidenced by the poor translations they produce into these languages (Robinson et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023). In low-resource scenarios, the fine-tuning of pre-trained multilingual machine translation models with the available training data has proven to be an effective approach (Scalvini et al., 2025). However, fewer than 13,000 sentence pairs are not enough to train a neural machine translation model to an acceptable quality (Pei et al., 2025; Gu et al., 2018). LLMs, on the other hand, can (through carefully designed prompts) rapidly be adapted to new tasks, thanks to their *in-context learning* (ICL) capabilities (Rubin et al., 2022; Cahyawijaya et al., 2024; Dong et al., 2024). Therefore, they offer the potential to make more efficient use of the available data.

This work aims to evaluate the effectiveness of different few-shot prompting methods for machine translation between Italian and two variants of Ladin, a low-resource language, on different LLMs. Rather than fine-tuning LLMs (Yong et al., 2023; Zhu et al., 2024; Stap et al., 2024; Toraman, 2024; Vieira et al., 2024), our approach aims to stimulate the generalization capabilities of LLMs through ICL using a single prompt. In particular, we investigate what can be achieved with a small retrieval corpus of approximately 18,000 parallel sentences and compare the results with a neural MT system trained on the same data.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8824–8837
11–16 May 2026.

Our key contributions:

- We introduce *Fragment-Shot* prompting (FS), a novel few-shot prompting method that aims to retrieve examples for all individual words that occur in the input sentence.
- We measure the performance of state-of-the-art LLMs (GPT-3.5, GPT-4o, o1-mini, Llama-3.3, and DeepSeek-R1) on translation between two variants of Ladin and Italian using FS and compare it with established prompting paradigms.
- We investigate the role of syntactic features in low-resource translation by analysing the *syntactic coverage*, i.e. the proportion of the words in the source/target sentence that are exemplified in the examples included in the prompt.

These contributions aim to explore the potential of LLMs in low-resource settings, to better understand the importance of the retrieved examples as well as other factors that can enable more accurate translations.

2. Related Work

The use of LLMs for MT has emerged as an active research area at the latest since the release of ChatGPT (Zhang et al., 2023). Researchers have increasingly explored LLMs as an alternative to traditional neural MT (NMT), showing that in some cases human annotators preferred ChatGPT over NMT systems (Manakhimova et al., 2023). However, the way LLMs are prompted plays a critical role and affects translation quality (Zhang et al., 2023; Agrawal et al., 2023; Vilar et al., 2023). Moreover, there is experimental evidence showing that GPT-models underperform on low-resource and African languages (Robinson et al., 2023). This has motivated research into strategies to improve this. In the following, we discuss related work in this area.

Prompt Engineering for MT with LLMs The *zero-shot* approach (Robinson et al., 2023) is the simplest way to prompt a LLM for translation, relying solely on the model's inherent language understanding (without providing task-specific examples). Various strategies that enrich the prompt with supplementary information have been shown to improve translation quality. These include providing randomly selected translation pairs in the prompt (Zhang et al., 2023) incorporating dictionary entries, word definitions, or grammar rules (Elsner and Needle, 2023; Court and Elsner, 2024; Guo et al., 2024; Merx et al., 2024; Zhang et al., 2024; Marmonier et al., 2025), translating individual words and retrieving related sentences (Shu et al., 2024; Zhang et al., 2024), or including translations of semantically related sentences in the prompt (Merx et al., 2024; Zebaze et al., 2025b; Tang et al., 2024; Kumar et al., 2023).

Recent work on Manchu (Pei et al., 2025) shows that, high-quality dictionaries and properly retrieved parallel examples are the most influential factors for translation quality. In our preliminary experiments, we observed that providing dictionary entries for individual words to the models do not seem to be very useful for Ladin, which is partly due to the high ambiguity of this language (many words can function as both adjectives and nouns). We therefore did not make use of dictionaries in our main experiments.

Example Sentence Retrieval Strategies While previous work has selected examples based on sentence-level similarity using metrics like BLEU (Agrawal et al., 2023), BM25 (Robertson et al., 1995) or semantic embeddings (Merx et al., 2024; Zebaze et al., 2025b), our FS approach aims to maximise the syntactic coverage with the selected examples. This strategy is particularly valuable in settings where reliable embedding models are not available. The importance of lexical overlap was previously emphasized by Agrawal et al. (2023); we measure the coverage more directly by counting the number of complete source words for which examples can be retrieved and reinforce this finding with the correlation results observed between this fragment coverage and BLEU scores in the FS setting.

Our approach combines syntactical overlap with context-aware retrieval, a combination identified by Tang et al. (2024) and Kumar et al. (2023) as particularly effective. Tang

et al. (2024) propose selecting full sentences by combining BM25 (Robertson et al., 1995) for lexical similarity with syntactic similarity at the example level to enhance translation performance. In contrast, our method promotes high syntactic overlap by increasing the fragment size, which enables the retrieval of sentences with greater syntactic relevance. It was inspired by DecoMT (Puduppully et al., 2023), a decomposed prompting approach that significantly outperforms standard few-shot methods, especially for translation between related low-resource languages. Unlike this multi-stage approach, which segments the text and translates each part with added context, our method is a single-prompt approach.

A similar multi-stage idea was introduced in Zebaze et al. (2025a), where the authors presented a method that first decomposes a sentence into simpler phrases, translates those phrases (with ICL examples) and then use these translations as ICL examples for the full sentence translation. In contrast, our work focuses on retrieval strategies for ICL examples, prioritizing high syntactic overlap, which we show to be particularly beneficial in our low-resource setting.

Machine Translation for Ladin To date, only a few studies have explicitly focused on Ladin in the context of machine translation with LLMs. Frontull and Moser (2024) explored the impact of synthetic data generated by different models, including `GPT-3.5`, on the performance of downstream NMT systems. Similarly, Valer et al. (2024) introduced a bidirectional MT system for Fassa Ladin, highlighting the benefits of multilingual training and knowledge transfer from related languages like Friulian and compared the results to the ones produced by `GPT-4o`. Recent work (Frontull et al., 2025) introduced the FLORES+ benchmark for two Ladin variants, Val Badia and Gherdëina, and compared the performance of a fine-tuned NLLB model with few-shot prompting using BM25 retrieval. The present work can be seen as a natural extension of these experiments, further exploring the capabilities of LLMs in this low-resource setting, leveraging the same datasets while introducing a novel strategy to improve example selection.

3. Prompting Techniques and Syntactic Coverage

This section details the prompting methods applied in our experiments and provides a precise definition of the concept we refer to as *syntactic coverage*.

3.1. Prompting Techniques

Our experiments include zero-shot prompting and four few-shot prompting techniques, each of which is detailed in this section.

Zero-Shot (`ZS`) The *zero-shot* prompting method (Robinson et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023; Gao et al., 2024) relies solely on the model's pre-existing knowledge. The prompt directly instructs the model to translate sentence into the target, without providing explicit translation examples or lexical guidance. This baseline approach tests the model's inherent understanding of Ladin syntax and vocabulary.

Random Few-Shot (`RS`) The *random few-shot* technique (Agrawal et al., 2023; Robinson et al., 2023; Bawden and Yvon, 2023) represents the most basic form of few-shot prompting, in which randomly selected source–target translation examples are included in the prompt, followed by the sentence to translate. Although these examples are not necessarily contextually related, they can help a model to infer translation patterns and linguistic structures. In our experiments, we provided 16 such examples.

Semantic Retrieval (`SR`) In the `SR`-method, source–target translation pairs are selected based on their semantic similarity to the input sentence. To identify these examples, we employed the multilingual sentence embedding model by Duquenne et al. (2023), which supports over 200 languages. Although the model does not officially support Ladin, it is reasonable to expect that the embeddings (and thus the retrieved examples) remain useful, given that Ladin shares linguistic features with several of the supported languages (e.g. Italian). In our experiments, we provided 30 examples in `SR` few-shot prompting.

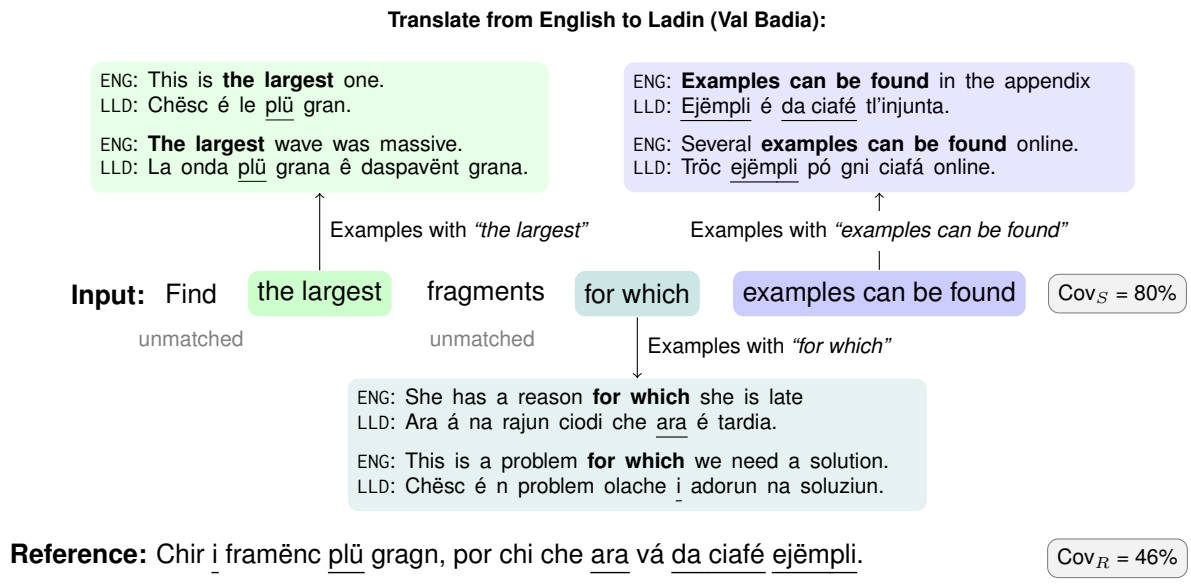

Figure 1: Fragment-Shot Prompting

Lexical Retrieval (BM) The BM-method includes translation pairs selected using the BM25 algorithm (Robertson et al., 1995). This retrieval method ranks examples based on lexical-level overlap with the input sentence, specifically prioritizing informative key terms. In our experiments, we provided 30 examples in BM few-shot prompting.

Fragment-Shot (FS) In contrast to the previous methods, which are already well established, *Fragment-Shot* is a novel technique developed in this work. This method employs a hierarchical sliding window search over contiguous word sequences, which we call *fragments*, to retrieve examples that ensure high syntactic coverage, prioritizing long-phrase matches and progressively falling back to smaller ones. This way, the examples illustrate both the use of the fragments and their plausible translations in (different) contexts. We illustrate the concept in Figure 1. In our implementation, we start with a window size of seven[1]. If examples are found, we randomly select up to six sentence pairs and include them in the prompt. If no matches are found, the window size is progressively reduced until we arrive at one-word units. Also, the fragments are chosen to avoid overlap, ideally ensuring a complete but non-redundant coverage of the sentence. To allow reproducibility and further experimentation, we released a Python package[2] that implements this method.

	VB→IT	GH→IT	IT→VB	IT→GH	VB→GH	GH→VB
1	79.64	81.33	81.68	82.14	80.54	82.13
2	15.75	13.12	15.94	15.36	15.50	12.13
3	3.93	4.75	2.20	2.41	3.32	4.95
4	0.53	0.69	0.17	0.09	0.55	0.76
5	0.11	0.03	-	-	0.04	-
6	0.04	0.07	-	-	0.04	0.04
>6	-	-	-	-	-	-

Table 1: Distribution of fragment window sizes retrieved for FS across language pairs.

Table 1 provides an overview of the fragment window sizes obtained for the different language pairs and reports the proportion of windows of size 1–6 (no fragments larger than size 6). Most retrieved fragments consist of single words, accounting for approximately 80% of all fragments.

3.2. Syntactic Coverage

We define *syntactic coverage* as the proportion of words in a sentence that appear in at least one of the retrieved examples. More formally, let I be the input sentence in the source

[1]Due to the short length of sentences in our corpora with less than 5 words on average, we could not retrieve any examples for window sizes larger than six; however, in principle, the initial window size could be set arbitrarily high.

[2]https://pypi.org/project/fragmentshot/

language and R the reference translation in the target language. We use $w(X)$ to define the set of words $\{x_1, x_2, \ldots, x_n\}$ in a sentence X. Let E_I denote the set of retrieved few-shot examples for I, each example $(S_k, T_k) \in E_I$ being a pair of a source-language sentence S_k with the corresponding translation T_k in the target language. The set of words in I covered in some example in E_I is defined as:

$$W_E(I) = \{\, x_i \in w(I) \mid \exists (S_k, T_k) \in E_I : x_i \in w(S_k) \,\}$$

We treat words in $w(I) \cap w(R)$ as non-translational units. Since non-translational units do not require translation, they are counted as covered and contribute to the overall coverage score. Thus, we define the syntactic coverage of the input sentence as:

$$\text{Cov}_I = \min\left(1, \frac{|W_E(I)| + |w(I) \cap w(R)|}{|w(I)|}\right)$$

This can be defined analogously for the reference translation R. Note that we use these metrics strictly for evaluation; they do not influence the selection process (after all, reference translations are not available at inference time). This analysis is intended solely to quantify the coverage of relevant content and lexical forms covered by the selected examples. In Figure 1 we provide the values for Cov_I and Cov_R for the example illustrated.

4. Experimental Setup

In our experiments, we focus on assessing how different LLMs perform when translating between Italian and the written standards of Val Badia and Gherdëina (two of the five main regional variants of Ladin). Due to the very limited amount of machine-readable data available for Ladin and the fact that these variants are not distinguishable in ISO 639-3[3] it is likely that LLMs have minimal exposure to Ladin and are unaware of its distinct varieties.

Retrieval Corpora As retrieval corpora, we have included the following datasets: 18,140 sentences for Val Badia–Italian[4], 19,971 sentences for Gherdëina–Italian[5] and 14,953 sentences for Val Badia–Gherdëina[6]. All of these datasets are publicly available under the `CC BY-NC-SA 4.0` license. These sentences, originally created as language reference material, are relatively short and simple. In particular, the average sentence lengths observed in our corpora are: (i) Val Badia (`VB`): 22.83 characters, 4.90 words; (ii) Gherdëina (`GH`): 22.69 characters, 4.91 words; (iii) Italian (`IT`): 25.69 characters, 4.36 words.

Test Data For evaluation, we used 175 sentences from the FLORES+ dev split (Frontull et al., 2025). To give an intuition on the similarity between the two variants and Italian and on the difficulty of the translation task, we computed the BLEU score obtained by leaving the text untranslated, which resulted in a score of 12.9 for Val Badia–Gherdëina, 5.0 for Italian–Val Badia and 4.3 for Italian–Gherdëina.

Large Language Models We selected the following five state-of-the-art LLMs: (i) `GPT-3.5`, a general-purpose language model from OpenAI's GPT-3 series with 175B parameters, released in 2022 (Brown et al., 2020); (ii) `GPT-4o`, a model by OpenAI, introduced in 2023, with 200B parameters and enhanced reasoning capabilities (Hurst et al., 2024); (iii) `o1-mini`, a model by OpenAI optimised for reasoning with 50B parameters, launched in September 2024 (Jaech et al., 2024); (iv) `Llama-3.3`, a text-only model by Meta AI, released in December 2024, featuring 70B parameters (Touvron et al., 2023); and (v) `DeepSeek-R1`, is a reasoning-focused model with 658B parameters by DeepSeek AI, introduced in January 2025 (DeepSeek-AI et al., 2025). The models were prompted using the API services: OpenAI API[7] for `GPT-3.5`, `GPT-4o`, and `o1-mini`, DeepSeek API[8] for `DeepSeek-R1`, and Together Inference API[9] for `Llama-3.3`. The hyperparameters (e.g., temperature) were kept at the default values of the respective API services.

[3] https://iso639-3.sil.org
[4] https://doi.org/10.57967/hf/1878
[5] https://doi.org/10.57967/hf/6726
[6] https://doi.org/10.57967/hf/6727
[7] https://platform.openai.com
[8] https://www.deepseek.ai/api
[9] https://www.together.ai

$t_{\text{inference}}$ [s]	ZS	RS	SR	BM	FS
GPT-3.5	0.82	0.91	0.66	0.75	0.77
GPT-4o	1.43	1.54	1.19	1.56	1.53
o1-mini	6.34	8.20	5.73	6.00	7.78
Llama-3.3	1.44	1.61	1.01	1.26	1.33
DeepSeek-R1	18.73	15.66	15.30	11.88	9.82
t_{corpus} [ms]	–	0.05	725k	200	324
t_{prompt} [ms]	0.12	15	1037	6	310
avg. #shots	0	16	30	30	33
avg. #chars	247	1882	3316	2913	3903

Table 2: Runtime and prompt statistics.

5. Results

To evaluate the different prompting techniques, we focused on several aspects. We measured the "costs" associated with each method, the time needed to initialise the retrieval corpus (t_{corpus}), including prompt creation time (t_{prompt}) and the median model inference time ($t_{\text{inference}}$). We also calculated the average prompt sizes in characters (#chars) and the average number of shots (#shots) included. Table 2 reports these statistics.

We evaluated translation quality using BLEU. Table 3 reports mean BLEU (Post, 2018) scores (computed with `sacrebleu`[10]) and standard deviations for the selected LLMs and the different prompting methods across translations between Val Badia, Gherdëina, and Italian. To determine whether differences between prompting strategies are statistically meaningful, we performed pairwise significance tests with `sacrebleu`, using FS as the baseline. The underlined values for FS denote statistically significant differences ($p < 0.05$). To further examine the influence of the different methods on example selection, we calculated the average syntactic coverage Cov_I and Cov_R and report their means and standard deviations in Table 3. As shown in the table, the examples selected by FS consistently achieve the highest coverage values.

Moreover, for translations into Italian, we conducted a qualitative evaluation using an LLM as judge (Kocmi and Federmann, 2023), in order to obtain deeper insights into the errors produced. We analyzed all 8,750 translations from both Ladin variants into Italian (for each model and prompting method) using Rubric-MQM (Kim, 2025).[11] This analysis revealed that mistranslations are the dominant error category, accounting for approximately 70% of the errors identified. As this category is the main target of improvement through few-shot prompting, we focus our analysis on mistranslations rather than on other error types such as stylistic, grammatical, or omissions errors. Figure 2 presents these results.

We also manually reviewed a subset of the Ladin translations and found that performing a similar qualitative error analysis would have been difficult. The outputs exhibited a high error density (consistent with the low BLEU scores) and were characterized by a lack of orthographic coherence, with the models failing to maintain a consistent spelling standard. Instead, we measured the proportion of words passing spellcheck.[12] These results are reported in Table 4. To contextualize: roughly 80% of the words in the Ladin reference translations pass spellcheck verification.

6. Discussion

Our results reveal several findings that can be summarized as follows:

1. syntactic coverage appears to be a crucial factor for low-resource translation: FS consistently outperforms the other methods across all models in both BLEU and lexical accuracy of the generated translations;

2. few-shot prompting yields only partial gains for Ladin to Italian translation. However, in this direction, LLMs outperform the NMT approach, positioning them as a prominent model for bootstrapping low-resource datasets;

3. despite the gains achieved via FS prompting, the fine-tuned NMT model remain superior for translation into Ladin, achieving similar/higher BLEU scores and better orthographic accuracy.

In the following, we draw conclusions from these findings.

[10] https://github.com/mjpost/sacrebleu

[11] We used the default configuration provided in the repository, with `gpt-4.1-mini-2025-04-14` as model and a temperature of 0.7

[12] These spellcheckers are also available as LibreOffice extensions (Frontull et al., 2025).

			Syntactic Coverage		GPT-3.5	GPT-4o	o1-mini	Llama-3.3	DeepSeek-R1
			Cov_I	Cov_R	BLEU	BLEU	BLEU	BLEU	BLEU
low- to high-resource	VB→IT	ZS	0.18 ± 0.11	0.19 ± 0.11	19.35 ± 2.09	22.90 ± 2.11	18.36 ± 2.31	20.31 ± 2.10	22.47 ± 2.04
		RS	0.41 ± 0.11	0.35 ± 0.12	20.75 ± 2.07	22.83 ± 2.11	18.29 ± 2.09	20.44 ± 2.27	22.80 ± 2.05
		SR	0.56 ± 0.11	0.46 ± 0.11	24.81 ± 2.04	29.36 ± 2.24	23.70 ± 2.10	26.46 ± 1.95	28.76 ± 2.13
		BM	0.73 ± 0.10	0.49 ± 0.12	**26.68** ± 2.03	**29.47** ± 2.25	23.87 ± 2.16	27.20 ± 2.04	**28.91** ± 1.93
		FS	**0.84** ± 0.08	**0.56** ± 0.12	25.20 ± 2.06	29.22 ± 2.32	**25.05** ± 2.05	**27.89** ± 2.21	27.98 ± 1.96
	GH→IT	ZS	0.18 ± 0.11	0.19 ± 0.11	18.52 ± 2.04	23.03 ± 2.09	17.26 ± 1.99	21.02 ± 2.12	**23.02** ± 1.96
		RS	0.42 ± 0.11	0.35 ± 0.11	19.35 ± 1.98	22.15 ± 2.32	18.63 ± 2.04	20.59 ± 2.11	22.45 ± 2.00
		SR	0.58 ± 0.10	0.44 ± 0.12	20.28 ± 1.96	23.12 ± 2.10	18.70 ± 2.08	21.24 ± 1.97	22.29 ± 1.75
		BM	0.74 ± 0.12	0.45 ± 0.13	**20.95** ± 1.69	**23.51** ± 2.16	19.96 ± 2.10	**22.33** ± 1.98	22.30 ± 1.83
		FS	**0.86** ± 0.07	**0.54** ± 0.12	18.49 ± 1.77	21.31 ± 2.11	**20.35** ± 2.05	21.68 ± 1.99	20.57 ± 2.00
hight- to low-resource	IT→VB	ZS	0.19 ± 0.11	0.18 ± 0.11	4.91 ± 1.23	5.70 ± 1.40	4.67 ± 1.22	6.46 ± 1.29	6.31 ± 1.24
		RS	0.36 ± 0.11	0.40 ± 0.11	5.01 ± 1.18	5.70 ± 1.45	5.22 ± 1.30	7.94 ± 1.47	6.91 ± 1.38
		SR	0.49 ± 0.13	0.53 ± 0.13	9.41 ± 1.25	10.89 ± 1.37	9.53 ± 1.30	13.94 ± 1.54	14.49 ± 1.62
		BM	0.67 ± 0.13	0.54 ± 0.13	9.64 ± 1.27	14.49 ± 1.55	11.99 ± 1.30	15.50 ± 1.77	16.06 ± 1.66
		FS	**0.77** ± 0.10	**0.61** ± 0.12	**11.63** ± 1.39	**16.91** ± 1.46	**14.23** ± 1.35	**17.36** ± 1.48	**18.32** ± 1.49
	IT→GH	ZS	0.19 ± 0.11	0.18 ± 0.11	6.73 ± 1.18	5.53 ± 1.20	6.69 ± 1.15	9.50 ± 1.35	8.77 ± 1.30
		RS	0.36 ± 0.12	0.43 ± 0.10	6.65 ± 1.15	7.56 ± 1.26	6.39 ± 1.13	9.50 ± 1.28	9.07 ± 1.23
		SR	0.48 ± 0.13	0.54 ± 0.12	7.44 ± 1.23	9.98 ± 1.37	8.28 ± 1.28	12.12 ± 1.65	12.52 ± 1.40
		BM	0.66 ± 0.12	0.54 ± 0.12	8.22 ± 1.19	11.62 ± 1.41	9.34 ± 1.26	13.20 ± 1.51	14.07 ± 1.46
		FS	**0.76** ± 0.10	**0.60** ± 0.11	**10.19** ± 1.42	**13.10** ± 1.47	**11.62** ± 1.43	**14.23** ± 1.46	**14.62** ± 1.40
low- to low-resource	VB→GH	ZS	0.40 ± 0.14	0.40 ± 0.13	10.52 ± 1.41	11.15 ± 1.61	8.94 ± 1.25	15.01 ± 1.69	13.11 ± 1.52
		RS	0.52 ± 0.12	0.54 ± 0.12	10.39 ± 1.34	12.21 ± 1.49	12.07 ± 1.64	16.61 ± 1.83	15.28 ± 1.64
		SR	0.63 ± 0.12	0.63 ± 0.11	12.14 ± 1.50	16.67 ± 1.92	14.91 ± 1.60	20.22 ± 1.89	19.27 ± 1.92
		BM	0.77 ± 0.10	0.66 ± 0.12	15.95 ± 1.49	22.72 ± 1.97	18.92 ± 1.75	24.35 ± 1.89	24.81 ± 1.78
		FS	**0.86** ± 0.08	**0.71** ± 0.11	**17.90** ± 1.94	**23.94** ± 2.18	**21.70** ± 1.87	**26.46** ± 2.17	**26.43** ± 2.21
	GH→VB	ZS	0.40 ± 0.13	0.40 ± 0.14	10.73 ± 1.53	11.17 ± 1.37	8.10 ± 1.20	12.46 ± 1.53	10.19 ± 1.34
		RS	0.55 ± 0.11	0.52 ± 0.12	11.25 ± 1.50	12.36 ± 1.45	10.83 ± 1.42	13.14 ± 1.53	13.75 ± 1.60
		SR	0.66 ± 0.10	0.60 ± 0.12	12.60 ± 1.61	16.40 ± 1.80	13.46 ± 1.56	17.39 ± 1.79	17.31 ± 2.21
		BM	0.79 ± 0.10	0.64 ± 0.13	16.00 ± 1.68	19.58 ± 2.00	16.90 ± 1.78	21.22 ± 1.98	22.49 ± 2.08
		FS	**0.88** ± 0.07	**0.71** ± 0.12	**17.19** ± 2.04	**21.92** ± 2.08	**19.52** ± 1.90	**22.81** ± 1.52	**24.44** ± 2.41

Table 3: BLEU mean scores and confidence intervals for the selected LLMs across various Ladin and Italian translation pairs using different prompting methods as reported by `sacrebleu`.

Syntactic Coverage Is a Key Objective for Low-Resource Translation The LLMs do not appear to have prior knowledge of Ladin or awareness of its variants, as demonstrated by their (rather poor) performance in ZS. For translation into Ladin, the choice of prompting method plays a particularly important role and can have a strong impact on the results.

Our method consistently achieves the highest BLEU scores, while also yielding the highest syntactic coverage values (both Cov_I and Cov_R). Moreover, it always leads to the highest proportion of words in the generated Ladin translations passing spellcheck verification. We also observe a significant correlation between coverage values and BLEU score for translation between Val Badia and Gherdëina. The importance of syntactic overlap is further emphasized by the fact that BM consistently performs better than SR in these translation directions. Thus, selecting few-shot examples with syntactic coverage as objective, as in FS, appears justified and effective in this scenario. This contrasts with previous results reported in Zebaze et al. (2025b), where semantic retrieval using SONAR embeddings outperformed BM25 retrieval for Swahili and Wolof. A direct comparison is not possible here, as the experimental setup differs and we did not use the same embedding model. Nevertheless, this difference would be worthwhile to investigate in more detail.

In addition to syntactic coverage, the size of retrieved fragments is also an important factor, as longer fragments generally capture more context and, therefore, should lead to better results. Our distribution of fragment sizes, shown in Table 1, indicates that the majority of retrieved fragments are single-word units, which reflects a limitation of what the current corpus allows us to retrieve. Experimenting with corpora that yield more multi-word fragments could provide deeper insights into how fragment granularity influences the performance.

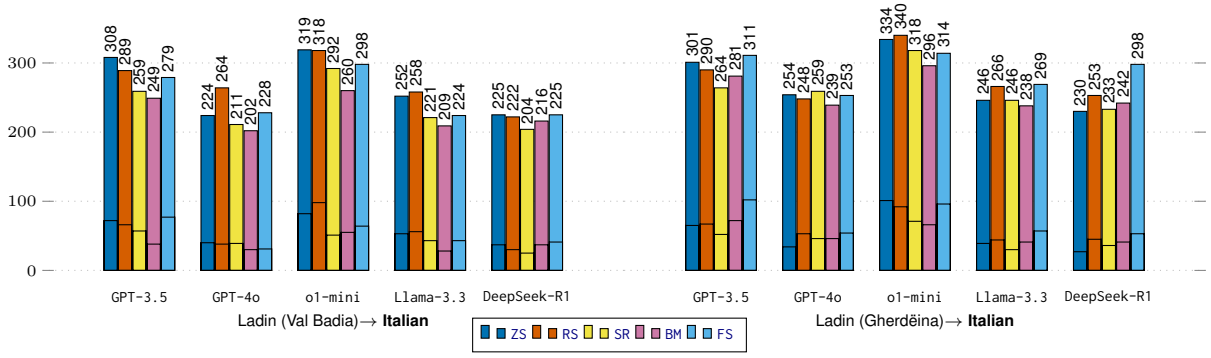

Figure 2: Number of mistranslations in the two low- to high-resource translation directions, evaluated across all 175 translations for each model and method. The horizontal line in the bars indicates the total number of mistranslations with severity ≥ 4.

	Italian→**Val Badia**					Italian→**Gherdëina**					Val Badia→**Gherdëina**					Gherdëina→**Val Badia**				
	①	②	③	④	⑤	①	②	③	④	⑤	①	②	③	④	⑤	①	②	③	④	⑤
ZS	41	43	38	44	44	40	37	39	52	50	44	45	44	55	53	44	48	44	50	48
RS	44	44	42	49	47	40	44	40	54	50	44	48	46	57	53	46	51	49	53	54
SR	47	54	49	57	62	46	53	48	59	62	48	56	53	62	61	52	58	55	60	61
BM	53	58	54	60	63	52	58	53	62	64	55	64	60	68	67	58	65	61	66	68
FS	**56**	**64**	**62**	**66**	**68**	**56**	**64**	**62**	**66**	**68**	**61**	**70**	**68**	**72**	**72**	**65**	**71**	**69**	**73**	**73**

Table 4: Percentage (%) of words in the generated translations in the Ladin target variant passing spellcheck verification for each model across methods and language pairs: ① GPT-3.5, ② GPT-4o, ③ o1-mini, ④ Llama-3.3, ⑤ DeepSeek-R1.

Few-Shot Prompting with Partial Gains from Ladin to Italian For the Ladin → Italian translation direction, neither the tested prompting strategies nor a higher coverage of retrieved information resulted in significant improvements in BLEU. To contextualize these scores, we also computed the BLEU score for English to Italian translation using GPT-4o, obtaining approximately 30 BLEU with ZS. This represents an approximate upper bound for the BLEU scores achievable in Ladin → Italian translation.

For translations to Italian, syntactic coverage seems less important. Correlation values between Cov_I and BLEU are low, even when syntactic coverage is high. In the cases where FS achieves statistically significant BLEU improvements (o1-mini), these gains are not confirmed by qualitative analysis. Rather than substantial differences between the (more sophisticated) prompting methods, the largest variations arise from the choice of LLM, with GPT-4o, Llama-3.3 and DeepSeek-R1 outperforming GPT-3.5 and o1-mini, as can be seen in both BLEU scores and number of mistranslations. This suggests that in this direction, the LLMs rely more on their inherent understanding of the target language.

Interestingly, in several cases, RS produced more mistranslations than ZS which suggests that (random) examples confused the models. This observation is consistent with Robinson et al. (2023); Alves et al. (2023); DeepSeek-AI et al. (2025); Reynolds and McDonell (2021), who reported that providing additional information can in some cases even degrade performance (when the examples are of low utility). Nevertheless, more careful or targeted example selection as in BM and SR consistently improves performance compared to RS, highlighting the importance of example selection.

For the Val Badia-to-Italian direction, FS, SR and BM yield substantial gains, bringing performance close to the 30 BLEU observed for English to Italian translation. However, in the qualitative analysis, this improvement is less pronounced and is clearly observable only for Llama-3.3. The absolute frequency of errors remains high across all models, suggesting that the outputs would still require significant human post-editing Even tough the syntactic coverage values and spellcheck statistics

	BLEU	Spell	Δ_{BLEU}	Δ_{Spell}
VB → IT	21.73		−7.74	
GH → IT	17.80		−5.71	
IT → VB	18.06	76	−0.26	+8
IT → GH	14.48	73	−0.14	+5
VB → GH	29.63	76	+3.17	+4
GH → VB	31.15	79	+6.71	+6

Table 5: Performance of the fine-tuned NMT model (BLEU and spelling accuracy in %) as reported in Frontull et al. (2025) in comparison with the best performing LLM and prompting method combination.

are similar across the two variants, there appears to be limited room for improvement for Gherdëina→Italian. Further investigation is needed to understand the underlying reasons for this discrepancy.

Few-Shot Prompting Falls Short of Neural MT in Low-Resource Translation As also reflected in the spell-checking statistics, the LLMs struggle with basic orthographic correctness of ladin texts and none of the evaluated methods yet reaches an acceptable level of quality for Italian to Ladin translation.

The performance of a fine-tuned NMT model trained and evaluated on the same datasets, as reported in Frontull et al. (2025), alongside the differences (Δ) relative to the best-performing LLM and prompting combination is given in Table 5. The results show that the NMT model underperforms compared to the best LLM-based approach for translation into Italian from both VB and GH, highlighting the potential of LLMs for *back-translation* (Sennrich et al., 2016) in low-resource languages, where they can provide high-quality initial translations (Ondrejová and Šuppa, 2024; Pei et al., 2025).

Conversely, for translations between into Ladin and between the variants, the NMT model performs comparably or even surpasses the LLM FS prompting in both BLEU and spelling accuracy. This shows that while the FS method allows for significant improvements, specialized NMT models remain superior for this task, aligning with previous findings (Robinson et al., 2023; Aycock et al., 2024; Scalvini et al., 2025).

7. Conclusion

In this work, we investigated the role of example selection strategies for few-shot machine translation of low-resource languages using LLMs. Our results demonstrate that syntactic coverage can be a key factor for effective in-context learning for translation into low-resource languages. However, fine-tuned NMT models remain the most effective approach, achieving comparable or higher BLEU scores and superior orthographic accuracy. When translating from the low-resource language to the high-resource language studied, LLMs are able to leverage the limited corpora more effectively than traditional NMT models, highlighting their potential as a tool for bootstrapping low-resource datasets.

Future Work Our results show that the in-context learning capabilities of LLMs can allow to improve translation quality for low-resource languages, but the mechanisms underlying this effectiveness still need to be understood in more detail (Chitale et al., 2024; Alves et al., 2023). It is still not entirely clear which specific characteristics of ICL examples drive performance in different translation directions. For instance, integrating semantic similarity metrics or contextual diversity could help explain why certain examples yield stronger correlations with BLEU than others, especially in low to high scenarios. Additionally, combining syntactic coverage with other example selection criteria, as for example done in Kumar et al. (2023), might yield even more effective few-shot prompting strategies.

Moving beyond single-query prompting, approaches that actively guide the reasoning process through mechanisms such as query languages like LMQL (Beurer-Kellner et al., 2023) could offer a way to make more effective use of the available data and to trigger specific reasoning. These insights also point toward the emerging paradigm of agentic machine translation (Briakou et al., 2024; Wu et al., 2025). Our findings suggest that understanding the role of syntactic coverage could provide valuable guidance for designing such agentic systems in low-resource MT scenarios. These are promising directions to explore in future work.

8. Acknowledgements

This work was carried out as part of the research project *Intelligent Writing Assistant for Ladin* at the University of Innsbruck, in collaboration with the Ladin Cultural Institute “Micurá de Rü”. We thank the reviewers for their comments which have greatly helped to improve the clarity and quality of our presentation.

Ethical Statement

We see a particular community consciousness in low-resource languages that are perceived as more trustworthy in certain contexts. Speakers of smaller languages have so far hardly been affected by phishing or similar attacks in their native language. With technological advances, these methods could be abused to exploit precisely this "greater trust" that these languages retain in the digital world. Researchers should consider this potential risk when developing technologies for low-resource languages. Even when data and models are published, risk mitigation measures can be taken, such as carefully documenting the intended use, providing usage guidelines, and developing tools with built-in safeguards to prevent misuse.

Limitations

We only conducted experiments on the translation between Ladin and Italian. As Italian belongs to the same language family as Ladin and shares structural and lexical similarities, our methods may have benefited from these similarities. For more distant languages, translation may be more challenging and further evaluation is needed to assess the generalizability of our findings.

Not all prompts included an instruction on in which format the result should be returned, and even when they did, the automatic readout of the translations was not always possible. Moreover, the models occasionally generated additional content or information that went beyond the translations to be generated. Since the amount of generated translations was manageable, we parsed them manually. However, we recommend considering this for the efficient scaling of the experiments.

9. Bibliographical 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.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima’an. 2024. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?

Rachel Bawden and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting Is Programming: A Query Language for Large Language Models. *Proc. ACM Program. Lang.*, 7(PLDI).

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An Empirical Study of In-context Learning in LLMs for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding Are Both the Problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and et. al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Samuel Frontull and Georg Moser. 2024. Rule-Based, Neural and LLM Back-Translation: Comparative Insights from a Variant of Ladin. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 128–138, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfarei. 2025. Bringing Ladin to FLORES+. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.

Yuan Gao, Ruili Wang, and Feng Hou. 2024. How to Design Translation Prompts for ChatGPT: An Empirical Study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. 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 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching Large Language Models to Translate on Low-resource Languages with Textbook Prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ahrii Kim. 2025. RUBRIC-MQM : Span-Level LLM-as-judge in Machine Translation For High-End Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining Multiple Features for In-context Example Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit Learning and the LLM in Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Viktória Ondrejová and Marek Šuppa. 2024. Can LLMs Handle Low-Resource Dialects? A Case Study on Translation and Common Sense Reasoning in Šariš. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. 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.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy Chen. 2023. DecoMT: Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4602, Singapore. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending Language Boundaries: Harnessing LLMs for Low-Resource Language Translation.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand. Association for Computational Linguistics.

Chenming Tang, Zhixiang Wang, and Yunfang Wu. 2024. SCOI: Syntax-augmented Coverage-based In-context Example Selection for Machine Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9956–9971, Miami, Florida, USA. Association for Computational Linguistics.

Cagri Toraman. 2024. Adapting Open-Source Generative Large Language Models for Low-Resource Languages: A Case Study for Turkish. In *Proceedings of the Fourth*

Workshop on Multilingual Representation Learning (MRL 2024), pages 30–44, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Giovanni Valer, Nicolò Penzo, and Jacopo Staiano. 2024. Nesciun lengaz lascià endò: Machine translation for Fassa Ladin. In *Proceedings of the 10th Italian Conference on Computational Linguistics*, Pisa, Italy. CEUR-ws.org.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA. Association for Machine Translation in the Americas.

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.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025a. Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22328–22357, Suzhou, China. Association for Computational Linguistics.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025b. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. 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.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.