Compensating for Data with Reasoning: Low-Resource Machine Translation with LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities in multilingual machine translation, sometimes even outperforming traditional neural systems. However, previous research has highlighted the challenges of using LLMs — particularly with prompt engineering — for low-resource languages. In this work, we introduce Fragment-Shot Prompting, a novel in-context learning method that segments input and retrieves translation examples based on syntactic coverage, along with Pivoted Fragment-Shot, an extension that enables translation without direct parallel data. We evaluate these methods using GPT-3.5, GPT-4o, o1-mini, LLaMA-3.3, and DeepSeek-R1 for translation between Italian and two Ladin variants, revealing three key findings: (1) Fragment-Shot Prompting is effective for translating into and between the studied low-resource languages, with syntactic coverage positively correlating with translation quality; (2) Models with stronger reasoning abilities make more effective use of retrieved knowledge, generally produce better translations, and enable Pivoted Fragment-Shot to significantly improve translation quality between the Ladin variants; and (3) prompt engineering offers limited, if any, improvements when translating from a low-resource to a high-resource language, where zero-shot prompting already yields satisfactory results. We publicly release our code and the retrieval corpora on https://github.com/XXX.

1 Introduction

In recent years, LLMs have made significant advancements in machine translation, gaining widespread attention and achieving promising results, especially for high-resource languages (Zhang et al., 2023). However, LLMs often face challenges when applied to low-resource scenarios where limited training data and resources are available, leading to poor translation quality (Robinson et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023). In particular, LLMs have limited awareness of (different variants of) smaller languages and struggle to distinguish between them and in producing coherent output (Court and Elsner, 2024; Ondrejová and Šuppa, 2024). In contrast, fine-tuning specialised models can be a more effective approach. However, Gu et al. (2018) found that fewer than 13,000 sentence pairs are not enough to train a neural machine translation model to an acceptable quality. Therefore, methods that can better exploit the potential of limited data are particularly in demand, and LLMs are a promising solution due to their strong generalisation capability. Previous studies have explored different techniques to improve LLM performance for low-resource languages (Elsner and Needle, 2023; Zhang et al., 2024; Merx et al., 2024; Guo et al., 2024; Gao et al., 2024; Shu et al., 2024; Moslem et al., 2023; Bawden and Yvon, 2023). These approaches typically enhance LLM output by incorporating additional information, such as dictionaries, grammatical rules, or example sentences via Retrieval Augmented Generation (RAG) (Lewis et al., 2020). Although LLMs still lag behind traditional neural translation systems in the translation of low-resource languages (Robinson et al., 2023), they hold significant potential for further exploration, especially as they continue to evolve. Recent advances in the multi-hop reasoning capabilities of LLMs have opened up new possibilities, especially in low-resource scenarios.

This work aims to evaluate the effectiveness of different prompting techniques for machine translation to and from the low-resource language Ladin using LLMs, in the case of Italian and the two standard variants of Ladin: Val Badia and Gherdëina. Specifically, it explores what can be achieved with a small set of parallel sentences available for as retrieval corpus. Rather than fine-tuning LLMs (Yong

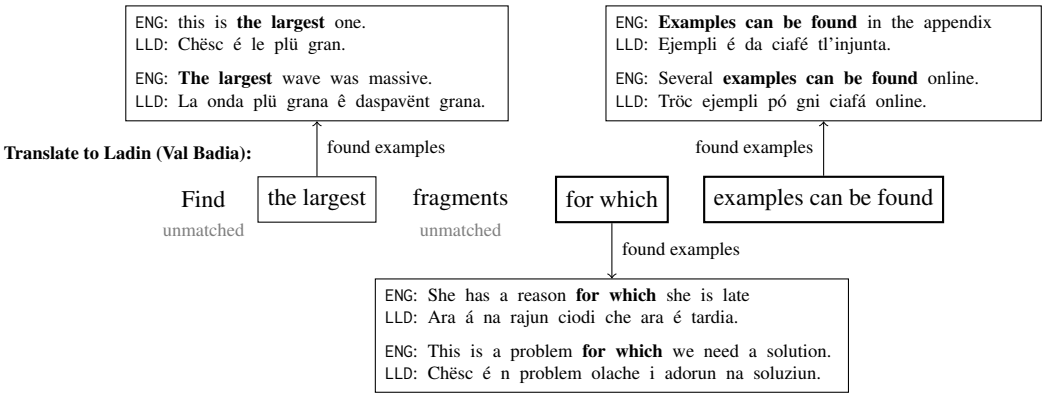

Figure 1: Fragment-Shot Prompting

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 *In-Context Learning* (ICL) (Rubin et al., 2022; Cahyawijaya et al., 2024; Dong et al., 2024), using a single RAG-augmented prompt.

Our key contributions: (i) We introduce the *Fragment-Shot* prompting technique, a novel prompting method that offers exemplary translations for individual fragments of the input sentence, selected to ensure broad syntactic coverage (Figure 1). Furthermore, we extend this approach with the *Pivoted Fragment-Shot* method, which enables translation between two languages that lack direct parallel data by leveraging a pivot language. (ii) We evaluate the performance of `GPT-3.5`, `GPT-4o`, `o1-mini`, `Llama-3.3`, and `DeepSeek-R1` on translation tasks between two variants of Ladin and Italian using four prompting methods: zero-shot, random-shot, Fragment-Shot, and Pivoted Fragment-Shot. We further examine the role of LLM reasoning capabilities in low-resource language translation through a coverage correlation analysis, assessing the relationship between translation quality and retrieved reference data, as well as a qualitative evaluation of the results. (iii) We publicly release our code along with the retrieval datasets containing parallel sentences for Gherdëina–Italian and for Val Badia–Gherdëina, to support further research on Ladin and low-resource languages in general. These contributions seek to illustrate how LLMs can be leveraged in low-resource settings and deepen our understanding of their reasoning capabilities.

2 Related Work

The use of LLMs for machine translation has emerged as an active research area at the latest since the release of ChatGPT (Zhang et al., 2023). Researchers have increasingly explored the potential of LLMs in comparison to traditional neural machine translation (NMT) systems, showing that human annotators, in some cases, preferred ChatGPT over mainstream 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 for low-resource and African languages (Robinson et al., 2023).

LLM-MT with RAG and Prompt Engineering The *zero-shot* approach (Robinson et al., 2023) is the simplest way to prompt an LLM for translation, relying solely on the model's inherent language understanding without task-specific examples. However, translating low-resource languages requires more sophisticated prompt engineering techniques: Some of the most effective strategies include Few-Shot Prompting (Brown et al., 2020), which improves output quality by providing a few illustrative examples; RAG (Lewis et al., 2020), which enriches prompts with relevant external information to reduce hallucinations and enhance translation quality; and Chain-of-Thought Prompting (Wei et al., 2022), which enables models to tackle complex problems by decomposing them into smaller, sequential reasoning steps.

Several studies have recently focused on improving LLM performance for low-resource languages through prompt engineering and/or RAG (Elsner and Needle, 2023; Zhang et al., 2024; Merx et al., 2024; Guo et al., 2024). Various strategies that enrich the prompt with supplementary information have been shown to improve translation quality. Examples include (i) *random-shot*, where randomly selected translation pairs are provided in the

prompt (Zhang et al., 2023), (*ii*) *dictionary-prompting*, where dictionary entries or word definitions are included (Merx et al., 2024; Zhang et al., 2024; Elsner and Needle, 2023; Guo et al., 2024), as well as (*iii*) the inclusion of translations of similar sentences in the prompt (Merx et al., 2024). The Fragment-Shot and Pivoted Fragment-Shot we present in this work build on these ideas. In contrast to Merx et al. (2024), we base our method on the highest word overlap, as semantic similarity via embeddings is not feasible due to the unavailability of suitable models.

Machine Translation and Multi-Hop Reasoning There is strong evidence of latent multi-hop reasoning in LLMs (Yang et al., 2024), a capability that enables models to draw on multiple pieces of information — potentially from different parts of a prompt or even from external knowledge — to arrive at a final answer. Puduppully et al. (2023) applied this idea in DecoMT, a decomposed prompting method that significantly outperformed standard few-shot approaches, particularly for translation between related low-resource languages. This method shares similarities with our Fragment-Shot approach. However, unlike the two-stage process that first segments the text and then translates each part with added context, our method translates the full text in a single prompt. Furthermore, it raises important questions about the performance of newer reasoning models on low-resource languages and the effective evaluation of the reasoning process involved in translation.

Machine Translation for Ladin To date, only two studies have explicitly focused on Ladin in the context of machine translation, both relying on basic zero-shot prompting without exploring more advanced prompting strategies: Frontull and Moser (2024) explored the effect of different models used for back-translation, including `GPT-3.5`. Similarly, Valer et al. (2024) introduced a bidirectional machine translation 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`.

3 Prompting Techniques

This section details the four prompting methodologies applied in our experiments to enhance machine translation performance for Ladin.

Zero-Shot (`ZS`) The *zero-shot* method (Robinson et al., 2023; Gao et al., 2024; Hendy et al., 2023; Bawden and Yvon, 2023) 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 intrinsic understanding of Ladin syntax and vocabulary.

Random-Shot (`RS`) In the *random few-shot* technique (Agrawal et al., 2023; Robinson et al., 2023; Bawden and Yvon, 2023) we provided 16 randomly selected source–target language translation pairs followed by the sentence to translate. These examples, even if not necessarily related, serve as in-context references, encouraging the model to infer translation and language patterns.

Fragment-Shot (`FS`) In the *Fragment-Shot* method, the sentence to be translated is partitioned into contiguous word sequences which we call *fragments*. For each fragment, we retrieve validated translation examples from the training corpus in which the fragment appears in the source sentence (see Figure 1). These fragments are constructed based on their occurrences in the retrieval data. We start with a sliding window of seven contiguous words and check whether any sentence in the retrieval data contains an exact match for this fragment on the source side. If such examples are found, we randomly select up to six sentence pairs containing the full fragment 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. We prioritize examples that explicitly contain the fragment in the source sentence, rather than selecting on the basis of a global sentence similarity. The examples serve to illustrate plausible translations of the fragment in context.

Pivoted Fragment-Shot (`PF`) The *pivoted* Fragment-Shot approach extends the `FS`-approach by enabling translation between two languages for which no direct parallel data is unavailable, leveraging a pivot language. This method applies the `FS` method in a nested manner across two different bilingual corpora: source-pivot and pivot-target. In our case we implemented this method for translating between Ladin (Gherdëina)

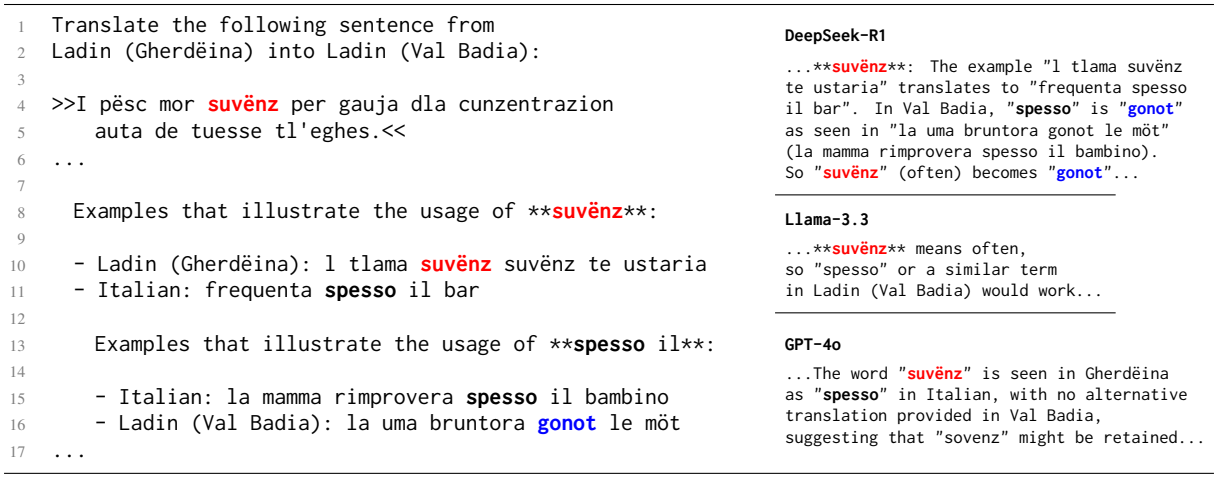

Figure 2: Example of Pivoted-Fragments Prompting and the corresponding reasoning employed by different LLMs.

and Ladin (Val Badia) via Italian.

Specifically, for a given sentence in the source language, we extract fragments as in the FS-method. For each fragment, we search the source-pivot corpus for up to three sentence pairs that contain the fragment on the source side and the corresponding pivot translations. We reduced to 2 for that exceeded the context size of the model. We then treat the pivot translations as new source texts and perform a second round of this search in the pivot-target corpus for (again, up to three) sentence pairs in the pivot language that contain fragments of the pivot translation, along with their translations into the target language. Given the substantial overlap between the Italian sentences in both corpora, we deliberately excluded exact pivot-sentence matches in order to force reasoning about fragments and thus simulate a more difficult translation problem. Figure 2 illustrates this approach by showing, on the left, the examples provided for the fragment **suvënz** occurring in the Ladin (Gherdëina) sentence, which connects via Italian to the corresponding Ladin (Val Badia) translation **gonot**. The right side illustrates the reasonings employed by `DeepSeek-R1`, `Llama-3.3`, and `GPT-4o`.

4 The Ladin Language

Ladin is a Rhaeto-Romance language spoken in the Dolomite region of Northern Italy. Ladin is characterised by its internal linguistic diversity, with five main regional variants: *Val Badia*, *Gherdëina*, *Fassa*, *Fodom*, and *Anpezo*. Each variant has its own orthographic conventions, vocabulary, and grammatical structures, making it a compelling case for machine translation research. This study analyses the performance of various LLMs in translating texts between Italian and the written standards of Val Badia and Gherdëina. These standards represent the Ladin varieties spoken by around 20,000[1] people in these two South Tyrolean valleys. Due to the very limited amount of machine-readable data available for Ladin and the fact that ita variants are not distinguishable in ISO 639-3[2] it is likely that LLMs have minimal exposure to Ladin and are unaware of its internal variation. To give an intution 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.

Retrieval Corpora As retrieval corpora, we have included the following datasets: 18,140 sentences for Val Badia–Italian, which have already been published[3], and 19,971 sentences for Gherdëina–Italian, extracted from the dictionary Ladin (Gherdëina)–Italiano (Forni, 2013). Since the Italian sentences of both datasets largely overlap, we aligned them to construct an additional parallel dataset for Val Badia–Gherdëina with 14,953 sentences. We make both the Gherdëina–Italian[4] and Val Badia–Gherdëina[5] datasets publicly avail-

[1]This number is based on data from the 2024 South Tyrolean Language Group Census, 2023 published by ASTAT at https://astat.provinz.bz.it/de/publikationen/ergebnisse-sprachgruppenzahlung-2024.

[2]https://iso639-3.sil.org

[3]https://doi.org/10.57967/hf/1878

[4]https://doi.org/XXX/YYY

[5]https://doi.org/XXX/YYY

able under a CC BY-NC-SA 4.0 license. These sentences, originally created as language reference material, are relatively short and simple with an average length of ≈ 25 characters.

Test Data For this study, we had 175 sentences from the FLORES+ (NLLB Team et al., 2024) dataset (dev split) translated into Val Badia and Gherdëina. These translations were produced by native Ladin speakers and professional translators affiliated with the Ladin Cultural Institute "Micurá de Rü"[6]. All translators followed the guidelines provided by OLDI[7], ensuring consistency and accuracy in the translation process.

5 Large Language Models

To evaluate the efficacy of the prompting techniques, we selected the following five state-of-the-art LLMs: (i) `GPT-3.5` is 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 designed for complex problem–solving (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` is a text-only model by Meta AI, released in December 2024, featuring 70B parameters (Touvron et al., 2023). (v) `DeepSeek-R1` is a reasoning-focused model by DeepSeek AI, introduced in January 2025, with 658B parameters (DeepSeek-AI et al., 2025). The models were prompted using the API services: OpenAI API[8] for `GPT-3.5`, `GPT-4o`, and `o1-mini`, DeepSeek API[9] for `DeepSeek-R1`, and Together Inference API[10] for `Llama-3.3`. The hyperparameters were configured according to the default settings provided by each service.

6 Results

Table 1 shows, for the selected LLMs `GPT-3.5`, `GPT-4o`, `o1-mini`, `Llama-3.3` and `DeepSeek-R1`, the mean BLEU (Post, 2018) scores computed with `sacrebleu`[11] as well as the standard deviation observed for `ZS`, `RS`, `FS` and `PF` between Val Badia, Gherdëina and Italian. We perform pairwise statistical significance tests to assess whether differences in BLEU scores between models and prompting strategies are meaningful or attributable to chance. Therefore, we examined three aspects, using `sacrebleu` pairwise tests: (1. underlined) we used the `FS` approach as the baseline in the significance test to determine whether it yields the best results for each model and translation direction; (2. dashed underlined) we used the `PF` approach as a baseline to assess whether it outperforms `ZS` and `RS` methods for each model and for translations between low-resource languages; (3. **bold**) for each prompting approach and translation direction, we used the `DeepSeek-R1` model as the baseline to assess whether it outperforms the others. We underlined the `FS` approach if it was statistically significantly better than all others, and we highlighted `DeepSeek-R1` in bold if it outperformed all other models.

Table 1 highlights three key findings: (1) In translations from low-resource to high-resource languages, the `FS` approach yields significant improvements only for `o1-mini` and `Llama-3.3`. For all other models, and for Gherdëina to Italian translations in particular, more sophisticated prompting techniques show little to no benefit; in some cases, `ZS` prompting even delivers the best results. (2) In translations from high-resource to low-resource languages, as well as between two low-resource languages, our `FS` approach consistently achieves the highest performance. Additionally, the Pivoted-Fragments prompting method yielded significant translation improvements compared to `ZS` and `RS`, but only for reasoning models. (3) `DeepSeek-R1` consistently outperforms all other models, especially in translations from high-resource to low-resource languages and between low-resource language pairs.

Table 2 presents statistics for the different methods. We compare the prompt creation times, the average prompt length (in characters), and the inference times of the various models. Our findings indicate that reasoning-oriented models generally require longer inference times. For all models, inference time tends to increase with prompt size, though not always in direct proportion. Notably, `PF`-prompts are significantly larger – approximately ten times the size of `RS`-prompts – and can quickly approach the model's context limits when processing longer sentences. Table 3 presents the syntactic coverage analysis for the different language com-

[6] https://www.micura.it
[7] https://oldi.org/guidelines
[8] https://platform.openai.com
[9] https://www.deepseek.ai/api
[10] https://www.together.ai
[11] https://github.com/mjpost/sacrebleu

Translation direction / BLEU		GPT-3.5	GPT-4o	o1-mini	Llama-3.3	DeepSeek-R1
Val Badia → Italian	ZS	19.35±2.09	22.90±2.11	18.36±2.31	20.31±2.10	22.47±2.04
	RS	20.75±2.07	22.83±2.11	18.29±2.09	20.44±2.27	22.80±2.05
	FS	19.77±2.01	22.49±1.99	21.07±2.16	21.98±2.17	22.46±1.94
Gherdëina → Italian	ZS	18.52±2.04	23.03±2.09	17.26±1.99	21.02±2.12	23.02±1.96
	RS	19.35±1.98	22.15±2.32	18.63±2.04	20.59±2.11	22.45±2.00
	FS	19.43±1.82	21.18±1.95	19.78±1.86	20.96±1.90	21.79±1.75
Italian → Val Badia	ZS	4.91±1.23	5.70±1.40	4.67±1.22	6.46±1.29	6.31±1.24
	RS	5.01±1.18	5.70±1.45	5.22±1.30	7.94±1.47	6.91±1.38
	FS	7.28±1.31	13.31±1.63	11.37±1.59	12.85±1.55	**14.22**±1.61
Italian → Gherdëina	ZS	6.73±1.18	5.53±1.20	6.69±1.15	9.50±1.35	8.77±1.30
	RS	6.65±1.15	7.56±1.26	6.39±1.13	9.50±1.28	9.07±1.23
	FS	8.58±1.27	12.58±1.50	11.51±1.46	13.32±1.58	**14.63**±1.48
Val Badia → Gherdëina	ZS	10.52±1.41	11.15±1.61	8.94±1.25	15.01±1.69	13.11±1.52
	RS	10.39±1.34	12.21±1.49	12.07±1.64	16.61±1.83	15.28±1.64
	FS	13.91±1.68	25.46±2.13	23.81±1.75	24.16±1.99	**28.60**±2.23
	PF	10.78±1.48	16.19±1.62	14.52±1.70	15.52±1.65	20.94±1.89
Gherdëina → Val Badia	ZS	10.73±1.53	11.17±1.37	8.10±1.20	12.46±1.53	10.19±1.34
	RS	11.25±1.50	12.36±1.45	10.83±1.42	13.14±1.53	13.75±1.60
	FS	13.99±1.70	23.10±2.02	22.21±2.00	21.70±2.01	**26.27**±2.16
	PF	10.08±1.39	15.67±1.69	13.32±1.54	14.07±1.67	19.33±1.70

Table 1: BLEU mean scores and confidence intervals of GPT-3.5, GPT-4o, o1-mini, Llama-3.3, and DeepSeek-R1 across various Ladin and Italian translation pairs using different prompting methods as reported by `sacrebleu`.

avg. duration [s]	ZS	RS	FS	PF
GPT-3.5	1.01	0.99	1.18	1.24
GPT-4o	1.63	1.78	6.85	7.34
Llama-3.3	1.74	2.04	6.49	9.54
o1-mini	7.10	9.54	15.27	27.56
DeepSeek-R1	19.42	20.87	34.22	30.97
creation time [s]	0.00	0.02	1.08	1.90
avg. # chars	247	2232	8974	24852

Table 2: Prompt statistics and average inference times.

binations for the FS-method. We evaluated the input sentence coverage by counting, for each sentence, how many words could be exemplified or assumed to be non-translatable (e.g., proper names). Moreover, we list the total number of fragments of size 1,2,3 and 4 we found in the retrieval corpus. We observe a significant correlation between coverage and BLEU score in translations from high-resource to low-resource languages, as well as between two low-resource languages–indicating that higher coverage of relevant information leads to better translation quality. However, this correlation does not hold in translations from low-resource to high-resource languages.

To give an insight into the translations generated with the different prompting methods, we have included an example sentence in Gherdëina in Table 4, along with the different outputs by selected models, as well as the reference translation in Val Badia. We have highlighted the input fragments for which we could retrieve examples, as well as the correctly translated segments in the generated translations.

7 Discussion

In the following, we discuss our main findings based on the results presented above.

Syntactic Coverage Correlates with Translation Quality into and between Low-Resource Languages In contrast to the Ladin to Italian direction, we observe substantial performance improvements when moving from ZS or RS to FS or PF. While previous work has selected examples based on sentence-level similarity—using metrics like BLEU (Agrawal et al., 2023) or semantic embeddings (Merx et al., 2024; Shu et al., 2024), our FS and PF approaches prioritise examples that maximise syntactic coverage, i.e., the inclusion of source words present in the input sentence. This strategy is particularly valuable in settings where building reliable embedding models is challenging due to limited data availability. 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 example translations are available and reinforce this finding with the correlation results observed between this fragment coverage and BLEU scores in the FS setting. There is a clear trend in all models that the FS-method

	Coverage	fragment size 1	2	3	4	pearson correlation GPT-3.5	GPT-4o	o1-mini	Llama-3.3	DeepSeek-R1
VB→ IT	0.88±0.08	1559	646	122	12	0.04	0.01	0.02	0.04	0.07
GH→ IT	0.89±0.06	1538	730	165	22	0.10	0.00	0.06	0.01	0.03
IT→ VB	0.70±0.12	1629	491	50	3	0.28*	0.24*	0.29*	0.29*	0.35*
IT→ GH	0.71±0.11	1628	480	54	1	0.42*	0.34*	0.41*	0.35*	0.42*
VB→ GH	0.85±0.13	1607	621	97	13	0.46*	0.48*	0.46*	0.45*	0.53*
GH→ VB	0.83±0.08	1602	706	146	21	0.40*	0.47*	0.48*	0.43*	0.50*

Table 3: FS-coverage and pearson correlation FS-coverage–BLEU statistics.

help the models to produce more accurate translations and promotes adherence to standard spelling conventions. However, the degree of improvement varies between models. For example, compared to ZS, the gain is approximately +3.3 BLEU points for GPT-3.5 and +15.5 for DeepSeek-R1. DeepSeek-R1 not only delivers the best performance gains, but also shows the highest correlation between syntactic coverage and translation quality, highlighting the importance of reasoning for effectively processing the prompts.

Reasoning Can Compensate for Data In the absence of direct parallel data, the Pivoted FS method—which translates between Val Badia and Gherdëina—offers a promising approach for low-resource language translation by leveraging nested FS prompting with Italian as the pivot language. Although the results achieved with this method are clearly inferior to those achieved with FS on direct parallel data, the leap in comparison to RS is significant for GPT-4o, o1-mini and DeepSeek-R1. DeepSeek-R1 demonstrates notable strengths in processing the structured PF prompts, significantly outperforming the other models. These findings suggest that robust reasoning capabilities are crucial for effectively applying the PF approach. PF requires models to perform *multi-hop* reasoning (Yang et al., 2024) across retrieved examples and pivot language rather than rely solely on language modeling or memorized translation patterns. We thus conclude that such capabilities can, at least to some extent, compensate for the lack of extensive training corpora. However, this is related to longer inference times (see Table 2). Further qualitative analysis revealed that DeepSeek-R1 is the model with the highest proportion of syntactically valid words in the generated translations for the specific target variant, also leaving the smallest proportion of words untranslated. Nevertheless, there are still challenges in fully capturing the vocabulary and morphology of the target variant. On average—compared to the reference translations—7–9% more words in the generated translations are syntactically not valid in the target language, indicating substantial room for improvement. The observed weaknesses may be explained by several factors: (*i*) the models do not appear to have prior knowledge of Ladin or a built-in distinction between its variants, as demonstrated by their performance in ZS; (*ii*) the retrieval corpus used in prompt construction is relatively small; based on the statistics shown in Table 3, we could expect coverage of only around 55% for Gherdëina—-Val Badia prompts; (*iii*) the reasoning is highly sensitive to the prompt. Figure 2 illustrates this issue: despite the prompt containing an example for the word *suvënz* with its Italian and Val Badia translations, some models fail to correctly infer the translation. In this case, the prompt explicitly refers to the fragment *spesso il*, which may have caused GPT-4o to claim that no example for *spesso* exists. Llama-3.3 understands the meaning of *suvënz* but fails to generate the correct translation, while DeepSeek-R1 succeeds, highlighting its comparatively stronger ability to perform *multi-hop* reasoning (Yang et al., 2024). Addressing these weaknesses could improve the overall effectiveness of the approach. Further improvements can be expected if language-specific resources are incorporated during the pre-training of the model, as discussed in Ye et al. (2025).

Limited Adaptation in High-Resource Language Translation Our findings indicate that, in the Ladin to Italian translation direction, neither the tested prompting strategies nor higher coverage of retrieved information lead to noticeable improvements in translation quality. This observation aligns with Robinson et al. (2023), who noted that few-shot prompting offers limited benefits. In some cases, providing additional information even led to degraded performance, an effect also reported by Alves et al. (2023); DeepSeek-AI et al. (2025); Reynolds and McDonell (2021). An exception is o1-mini and Llama-3.3, which achieve the best

	Gherdëina: I pësc mor suvënz per gauja dla cunzentrazion auta de tuesse tl'eghes.
GPT-4o	ZS: I pësc moro suvent per ieie dla cunsentrazion ota de tuesc tl'ega. FS: I pësc mör gonot por gauja dles conzentrazion alta da tosser tl'ecas. PF: I pësc ne mör sovenz por gauja dl cunzentrazion eveata de töies tl'egnes.
Llama-3.3	ZS: I pesc se à ziplà de spess per la gauja dla concentrazion auta de tuess te eghes. FS: I pësc mör gonot por gauja dles cunzentrazions altas de tosser te eglises. PF: I pesc muion spesso per la ragion de na cuncentrazion auta de velen tl'ega.
DeepSeek-R1	ZS: I pesc muer sovenz per via de la cunzentrazion auta de tossich tla ega. FS: I pësc mör gonot por gauja dla cunzentrazion alta de tosser tles eghes. PF: I pësc mör gonot por gauja dla cunzentrazion auta de tuesse tles eghes.
	Val Badia: I pësc mör gonot porvia dles conzentraziuns altes dla tossina tles eghes.

Table 4: Example translation Gherdëina to Val Badia.

results with the FS method for Ladin (Val Badia) to Italian translation. To contextualise the results, we computed the BLEU score for English to Italian translation using GPT-4o and obtained a value of $\approx$ 30 BLEU, demonstrating strong ZS performance, while still highlighting room for improvement, considering that the Ladin texts are expert translations of the original English. We observed notable performance differences across the models where o1-mini often retained Ladin words. In contrast, GPT-4o and DeepSeek-R1 consistently generated acceptable results without source-language interference. GPT-3.5 and Llama-3.3 sometimes adhering too closely to Ladin sentence structure. These results primarily reflect the underlying multilingual capabilities of the models.

Specialised NMT Models Are Superior for Translation into Low-Resource Languages For Ladin, NMT models have so far only been published for the language pair Ladin (Val Badia)–Italian (Frontull and Moser, 2024). We have evaluated them on our test data in order to obtain a performance comparison. For Italian to Ladin (Val Badia), the best performing NMT model (L4) achieved a BLEU score of 16.77, which is significantly higher than the best performing LLM. This confirms that while the FS method allows for significant improvements, specialised NMT models remain superior for this task (Scalvini et al., 2025; Robinson et al., 2023; Aycock et al., 2024). However, these NMT models were trained also with monolingual texts, giving them access to language-specific information that was not leveraged in our prompting experiments. Methods that aim to incorporate such monolingual data have, for instance, been explored in Guo et al. (2024). In the opposite direction, the best performing NMT model R4 achieved a BLEU score of 17.04, which is lower than the one achieved by the different LLMs with ZS. This highlights the potential of LLMs for low-resource languages. For example, they can be used to produce higher-quality initial translations (Ondrejová and Šuppa, 2024), which is a particularly relevant aspect of the widely used *back-translation* (Sennrich et al., 2016).

8 Conclusion

In this work, we introduced Fragment-Shot Prompting, a novel in-context learning method that improves the quality of translations into and between low-resource languages with LLMs by retrieving examples based on syntactic coverage. Building on this idea, we proposed Pivoted Fragment-Shot: an extension that enables translation without the need for direct parallel data, leveraging a pivot language instead. While prompt engineering only offers marginal improvements when translating into high-resource languages, it becomes significantly more impactful in low-resource scenarios. Our experiments emphasise the importance of syntactic coverage in example selection. However, selecting examples solely based on syntactic overlap, without access to semantic information, makes it difficult to capture connections between source and target language, as fragments may have multiple meanings and uses. As a result, effective translation in such cases requires models with strong reasoning capabilities. At the same time, reasoning enables models to generalise beyond the examples, reducing the need for large amounts of data.

Future Work Our results show that in-context learning combined with the reasoning capabilities of LLMs can 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), highlighting the need for further research. Future work could also consider approaches that go beyond a single-query approach that guide the reasoning process, such as through the use of query languages like LMQL (Beurer-Kellner et al., 2023).

Ethical Statement

We see a particular community consciousness in low-resource languages that are perceived as more trustworthy in certain contexts. For example, speakers of such 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.

Limitations

Our approach augments prompts with data from an external corpus. While this method has the potential to improve machine translation for low-resource languages, we recognise several risks associated with its use. LLMs often lack robust safety mechanisms for low-resource languages, making them more prone to generating inaccurate, inappropriate or harmful content (Yong et al., 2024; Deng et al., 2024; Zeng et al., 2024). Augmenting prompts with external data could exacerbate this problem by increasing hallucinations, potentially leading to the production of offensive or misleading translations and texts. In addition, our approach retrieves relevant ICL-examples based on syntactic similarity to fragments of the input sentence. However, this purely syntactic selection does not take into account potential biases in the training data or in the model itself. As a result, our method may inadvertently propagate or even amplify existing biases, raising concerns about fairness and ethical use. Future work should explore strategies to mitigate these risks, such as refining selection criteria beyond syntax or incorporating bias-aware filtering mechanisms.

We only conducted experiments on 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 generalisability 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, e.g. to explain the generated translations or to warn the user of a lack of knowledge of the Ladin language. Since the amount of generated translations was manageable, we parsed the translations manually from the generated output. However, we note that this should be considered for efficient scaling of the experiments.

Although the PF-method shows promising results when translating between variants of Ladin, the size of the prompts generated by this nesting is a point of criticism. The size of the prompts increases rapidly depending on the length of the input and the translations found, as there are no assumptions about how the fragments in the source sentence correspond to those in the pivot language. The result is a search for example translations for all the fragments in the intermediate sentences. In our case, the translations in the corpus were simple, short sentences and so we have not yet reached the limits here, but this could look different with other training data. Introducing an alignment for these fragments could reduce the size of the prompts and improve efficiency by allowing less relevant examples to be omitted.

One limitation of our work is the relatively small test data set, which consists of only 175 sentences. A more comprehensive evaluation using the full FLORES+ dataset would provide more robust and representative results. To ensure compliance with the original access conditions, any release of the dataset should be complete and formally submitted to OLDI, thereby preserving its integrity as an evaluation benchmark. Given that our dataset is incomplete, we have chosen not to release it to prevent unintended uses, as it may end up in the training data of models - an outcome that would compromise FLORES+'s role as a benchmark for assessing translation quality.

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? *Preprint*, arXiv:2409.19151.

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).

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. *Preprint*, arXiv:2501.12948.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. *Preprint*, arXiv:2310.06474.

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.

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.

Marco Forni. 2013. *Dizionario italiano – ladino gardenese / Dizioner ladino gardenese – italiano*. Istitut Ladin Micurá de Rü, San Martin de Tor.

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.

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. *Preprint*, arXiv:2302.09210.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

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.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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.

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.

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. *Preprint*, arXiv:2411.11295.

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.

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. *Preprint*, arXiv:2302.13971.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4. *Preprint*, arXiv:2310.02446.

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.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. 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.