Understanding In-Context Machine Translation for Low-Resource Languages: A Case Study on Manchu

Renhao Pei[1,*], **Yihong Liu**[1,2,*], **Peiqin Lin**[1,2], **François Yvon**[3,†], **and Hinrich Schütze**[1,2,†]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]Sorbonne Université, CNRS, ISIR, France
renhaopei@gmail.com yihong@cis.lmu.de

Abstract

In-context machine translation (MT) with large language models (LLMs) is a promising approach for low-resource MT, as it can readily take advantage of linguistic resources such as grammar books and dictionaries. Such resources are usually selectively integrated into the prompt so that LLMs can directly perform translation without any specific training, via their in-context learning capability (ICL). However, the relative importance of each type of resource, e.g., dictionary, grammar book, and retrieved parallel examples, is not entirely clear. To address this gap, this study systematically investigates how each resource and its quality affect the translation performance, with the **Manchu** language as our case study. To remove any prior knowledge of Manchu encoded in the LLM parameters and single out the effect of ICL, we also experiment with an enciphered version of Manchu texts. Our results indicate that high-quality dictionaries and good parallel examples are very helpful, while grammars hardly help. In a follow-up study, we showcase a promising application of in-context MT: parallel data augmentation as a way to bootstrap a conventional MT model. When monolingual data abound, generating synthetic parallel data through in-context MT offers a pathway to mitigate data scarcity and build effective and efficient low-resource neural MT systems.[1]

1 Introduction

Neural machine translation (NMT) systems have achieved remarkable performance in high-resource language pairs for which parallel sentence-level or document-level data are abundant (Bahdanau et al., 2015; Vaswani et al., 2017; Tiedemann and Scherrer, 2017; Läubli et al., 2018). However, parallel data is scarce or even unavailable for many low-resource or endangered languages (Haddow et al., 2022), which prevents the training of dedicated MT systems for these languages. While multilingual models partly mitigate this issue (Costa-jussà et al., 2024), they only cover a small fraction of the world's languages and their performance remains unsatisfactory for many language pairs.

On the other hand, owing to the work of field linguists, grammatical descriptions or dictionaries are available for more than 60% of the world's languages (Nordhoff and Hammarström, 2011; Zhang et al., 2024b).[2] Some low-resource languages are well-documented, with rich linguistic resources gathered over decades of meticulous fieldwork and analysis by linguists: this is for instance the case of Japhug, a minority Sino-Tibetan language, for which a comprehensive grammar, including plentiful glossed and translated examples, has been released by Jacques (2021). The situation of Manchu is even more favorable, as multiple grammar books, dictionaries, and textbooks are readily available. Yet, all these languages are still considered low-resource in the context of data-driven MT, simply due to the scarcity of parallel data. A natural question is then to explore whether such linguistic knowledge can make up for the lack of parallel data, and help develop MT systems.

The recent emergence of LLMs seems to offer new promising ways to address this question, based on their in-context learning ability (Tanzer et al., 2024; Zhang et al., 2024b; Hus and Anastasopoulos, 2024; Merx et al., 2024). In these studies, linguistic resources such as dictionaries, parallel examples, and grammar books are integrated into the prompt and encoded together with the sentence to be translated. We continue this line of work, trying to better analyze the role and impact of each

[*]Equal contribution.
[†]Equal advising.
[1]We make our code and data publicly available at: https://github.com/cisnlp/manchu-in-context-mt.
[2]This includes long grammatical books (24%), short grammatical books (13%), and grammatical sketches (25%), according to https://glottolog.org/langdoc/status.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8767–8788
July 27 - August 1, 2025

type of linguistic knowledge that can be put to use in LLM-based machine translation systems.

For this, we perform a systematic investigation of how each component affects the in-context MT performance, with the translation from **Manchu** into English[3] as a case study. Specifically, we leverage a wide range of state-of-the-art open-source and closed-source LLMs and consider the following linguistic resources (components): dictionaries, parallel examples, grammar books, and Chain-of-Thought (CoT) prompting. For each component, we consider several variants that vary in the amount of information or the degree of relevance to the sentence to be translated. To quantify the influence of prior knowledge of Manchu in LLMs, we perform a character-level encipherment to disentangle the effect of LLMs' prior knowledge of Manchu from their in-context learning ability. In addition, we demonstrate a use case of our in-context MT system, using it as a data-augmentation tool to turn a monolingual Manchu corpus into a parallel corpus. With these synthetic parallel data incorporated into the training set, we fine-tune the mT5 model (Xue et al., 2021), achieving a substantial performance gain compared to the baseline that only uses actual parallel data.

The main contributions of this work are as follows: **(i)** We conduct a comprehensive investigation of in-context MT for Manchu, exploring the most important knowledge sources provided in the context, highlighting the positive role of high-quality dictionaries and closely related parallel examples. **(ii)** Using an enciphered version of Manchu, we isolate the limited prior knowledge of Manchu encoded in the LLMs considered in our work and show that most of their translation performance depends on their in-context learning abilities. **(iii)** We use in-context MT to generate synthetic parallel data from monolingual data of Manchu and measure how much this form of data augmentation actually benefits low-resource NMT.

2 Related Work

Low-resource NMT The challenges posed by limited parallel data has motivated extensive research on innovative strategies for low-resource NMT (Haddow et al., 2022; Yazar et al., 2023). Various approaches have been proposed to improve translation quality in such settings. Data augmentation techniques, such as back-translation (Sennrich et al., 2016; Edunov et al., 2018) and forward-translation (Bogoychev and Sennrich, 2020), have been widely used to generate synthetic parallel data and improve model performance. Data augmentation, coupled with unsupervised and semi-supervised methods for bilingual dictionary induction, has enabled translation with minimal parallel data, relying instead on monolingual resources (Lample et al., 2018; Artetxe et al., 2018). Transfer learning has also proven effective, where models pretrained on high-resource language pairs can be adapted to low-resource languages (Zoph et al., 2016; Tars et al., 2022; Her and Kruschwitz, 2024). Recent advancements in multilingual NMT also show that models trained on multiple language pairs can better deal with low-resource languages (Ko et al., 2021; Mohammadshahi et al., 2022; Costa-jussà et al., 2024). Despite these advancements, achieving high-quality translation in low-resource scenarios remains a significant challenge.

LLM-based In-context MT for Low-Resource Languages Although not explicitly trained for machine translation, LLMs can perform translation by following instructions and demonstrations in the prompt (Brown et al., 2020; Lin et al., 2022; Vilar et al., 2023). LLM-based MT, however, struggles with rare words that appear infrequently in the training data (Ghazvininejad et al., 2023). This issue is particularly pronounced for low-resource languages that are underrepresented in the LLM's training corpora (Le Scao et al., 2023; Touvron et al., 2023). To mitigate this, some studies incorporate linguistic resources into prompts, such as **dictionary entries** and **parallel sentence examples** (Ghazvininejad et al., 2023; Zhang et al., 2024a), as well as **grammars** (Tanzer et al., 2024; Hus and Anastasopoulos, 2024). Some works also include **morphological analyzers** to decompose input sentences into morphemes (Zhang et al., 2024b). Additionally, prompting strategies such as **CoT** reasoning have been explored in the context of MT (Elsner and Needle, 2023). However, little attention has been given to how the quality of each component affects the LLM-based in-context MT. Moreover, there is a lack of clear ablation studies disentangling the effects of an LLM's prior knowledge of the language and the linguistic information provided in context. Addressing these limitations of previous studies, our work systematically investigates the role of each of these components in

[3]Translation into from Manchu into Chinese is also considered in Appendix F.

LLM-based MT for low-resource languages, using Manchu as a case study.

3 Language, Data and General Setup

Manchu Language Manchu (ISO 639-3: mnc) is a critically endangered Tungusic language native to Northeast China. It is the traditional language of the Manchu people and was one of the official languages of the Qing dynasty (1644-1911) of China. Because of its significant historical importance, Manchu has been extensively studied, and there exist abundant linguistic resources, including dictionaries, grammar books, and some bilingual parallel sentences, which make Manchu well-suited for our case study. A more detailed description of the Manchu language is given in Appendix A.

Dictionary We use the comprehensive dictionary from Norman (2020),[4] which contains rich information such as the multiple senses for polysemous words as well as frequent collocations. It serves as our main Manchu-English lexicon. Additionally, we compile a dictionary for Manchu suffixes based on (Clark, 1980), which contains brief explanations for each suffix.[5]

Parallel Corpus The main source of parallel data is a Manchu-Chinese dictionary (Hu, 1994), which contains parallel example sentences for many dictionary entries.[6] We extract parallel sentences from the dictionary, followed by data-cleaning and filtering steps, to ensure that the Chinese sentences are in modern Standard Chinese. The result is a Manchu-Chinese parallel corpus consisting of 3,520 sentence pairs, encompassing diverse genres, including everyday conversations, historical records, and literary works. We then use the Google Cloud Translation API to translate the Chinese sentences into English, thereby creating a Manchu-English parallel corpus.[7]

Monolingual Corpus We also compile a monolingual Manchu corpus consisting of 42,240 sentences collected from websites, encompassing a diverse range of genres.[8] During our data augmentation experiment presented in §6, this monolingual Manchu corpus serves to build a synthetic Manchu-English parallel corpus.

Grammar We use two grammar books: a concise grammar book (Norman, 1965) and a more detailed grammar from (Gorelova, 2002).

Evaluation Set We compile a test set of 337 Manchu-English parallel sentences for evaluation. This test set consists of 70 sentences from *Nogeoldae*, a book containing dialogues in Manchu, paired with English translations (Zhang et al., 2024b), and 267 sentence pairs extracted from (Di Cosmo, 2007).[9] We have made sure that the parallel corpus and the evaluation set do not overlap.

Models We conduct our experiments with multiple LLMs: GPT-4o (Achiam et al., 2024), DeepSeek-V3 (Liu et al., 2024), and Llama3 models (Dubey et al., 2024). For the Llama3 family, we test models of varying sizes – 1B, 3B, 8B, and 70B – to evaluate how model size impacts performance.

Evaluation Metrics We use BLEU (Papineni et al., 2002) and chrF (Popović, 2015) to measure the translation quality, as implemented by SacreBLEU (Post, 2018).[10] Additionally, we use SBERT (Reimers and Gurevych, 2019), an encoding based-metric, which assesses the semantic relatedness between a hypothesis and a reference using the cosine similarity of their embeddings (scores are multiplied by 100 to ensure a uniform magnitude).

4 Assessing Each Component

Following the standard pipeline for in-context MT (Tanzer et al., 2024; Hus and Anastasopoulos, 2024; Zhang et al., 2024b,a), our goal is to conduct a rigorous investigation of the importance of each type of linguistic resource (component) and its quality to the translation performance. For each input Manchu sentence, a structured prompt is constructed by integrating various components. This prompt is then fed to the LLM to generate a response, from which the translation is extracted. The translation is finally evaluated against the ground truth reference using various metrics.

Formulation of Prompts We represent the prompt formulation as $\pi(\cdot)$ which takes several ar-

[4] https://buleku.org/home.

[5] Manchu exclusively uses suffixation, therefore neither prefixation nor circumfixation is involved.

[6] Data is available from https://gerel.net/.

[7] https://cloud.google.com/translate?hl=en

[8] https://manc.hu/ and https://gerel.net/

[9] https://github.com/ulingga/Manchu-English_babyMT.

[10] Signature: nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.4.3.

guments as input. Let $\mathbf{x}$ be the Manchu sentence to be translated. The simplest prompt is $\pi(\mathbf{x})$, which asks the LLM to directly translate $\mathbf{x}$ into the target language, without providing any additional information. The prompt template can be augmented by adding optional arguments as follows – each representing one component.[11]

- A morphological analyzer $\mu(\cdot)$, which transforms $\mathbf{x}$ into segmented and analyzed morphemes. The result is represented as $\mu(\mathbf{x})$.
- Dictionary entries D retrieved from a bilingual dictionary $\mathcal{D}$.
- Parallel examples P retrieved from a parallel corpus $\mathcal{P}$.
- Grammar excerpts G retrieved from a grammar book $\mathcal{G}$.
- CoT prompting instructions C selected from a set of prompting varieties $\mathcal{C}$.

Sequential Integration of Components Given that many components in our pipeline have multiple implementations of varying quality, exhaustively evaluating all possible combinations would be computationally infeasible. Therefore, we add components to $\pi(\cdot)$ sequentially and compare performance between implementations for that component. The best-performing one is used as a new baseline when we evaluate the next component. Specifically, starting with the simple baseline $\pi(\mathbf{x})$, we first add the morphological analyzer, a fundamental element for subsequent retrieval components, resulting in $\pi(\mu(\mathbf{x}))$.[12] We then consider components that have multiple variants. To begin with, we consider various ways to specify D and select the best one $\pi(\mu(\mathbf{x}), \mathrm{D}^*)$ which is the new baseline for subsequent add-ons. We then follow the order P, G, C, resulting in $\pi(\mu(\mathbf{x})^*, \mathrm{D}^*, \mathrm{P})$ (assessing multiple ways to select parallel examples), $\pi(\mu(\mathbf{x})^*, \mathrm{D}^*, \mathrm{P}^*, \mathrm{G})$ (assessing a variety of grammar excerpts), and $\pi(\mu(\mathbf{x})^*, \mathrm{D}^*, \mathrm{P}^*, \mathrm{G}^*, \mathrm{C})$ (assessing variants of CoT instruction). This order prioritizes components that are expected to be most beneficial, with less helpful components introduced at later stages, as suggested by previous works (Zhang et al., 2024b; Hus and Anastasopoulos, 2024). The pipeline is depicted in Figure 1. In the following sections, we study each component in detail and report experimental results obtained using the **GPT-4o model** on the evaluation set of 337 Manchu-English parallel sentences.

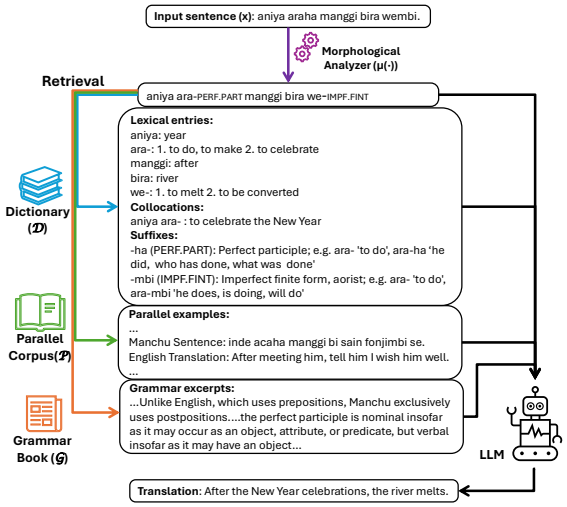

Figure 1: Illustration of the in-context MT pipeline with the components of π, i.e., $\mu(\mathbf{x})$, D, P and G.

4.1 Morphological Analysis

Morphological analysis is usually performed in a naive way in previous studies. For instance, Zhang et al. (2024b) simply perform a dictionary look-up – searching for **inflected word forms** in a dictionary, the coverage of which is limited.[13] As Manchu is an agglutinative language and exclusively uses suffixation, identifying word stems and suffixes is straightforward. Therefore, we implement a rule-based morphological analyzer that splits an input word into a stem and a sequence of suffixes. Both the **list of word stems** and the **set of allowed suffixes** are obtained from the dictionary. Our morphological analyzer then attempts to recursively detach a suffix from the end of a string until the remaining segment matches a known word stem. After the morphological analysis, a Manchu sentence is transformed into a list of morphemes (containing word stems and suffixes), which serves as the basis for retrieving dictionary entries, parallel examples, and grammar excerpts.

It is possible for a Manchu word to have multiple analyses. For example, *tere* could be a demonstrative pronoun meaning "that", or it could be analyzed as *te-re*, meaning "sitting" as the present participle of the verb *te* "to sit". In such cases, we include all possible analyses in the prompt and

[11] Prompt templates are illustrated in Appendix B.

[12] We only consider one version of $\mu(\mathbf{x})$, which constitutes an essential component for all other linguistic resources.

[13] The dictionary is from Norman (2020). It can be accessed at https://buleku.org/home.

let the LLM resolve the ambiguity by selecting the most contextually appropriate interpretation, as shown in Table 11 of Appendix I.

4.2 Dictionary

The dictionary $\mathcal{D}$ comprises lexical entries, suffixes, and their corresponding collocations. These three elements – lexical entries, suffixes, and collocations – form the foundation of our three variants:

- D^{l} includes only the **lexical** entries retrieved for each word in the input sentence, without explanation[14] for the suffixes.
- $\mathrm{D}^{\mathrm{l+s}}$ includes both the **lexical** entries and explanations for **suffixes** for all morphemes appearing in the input sentence.
- $\mathrm{D}^{\mathrm{l+s+c}}$ includes the **lexical** entries, explanations of **suffixes**, and the **collocations** for all morphemes in the input sentence.

Variant	BLEU	chrF	SBERT
$\pi(\mathbf{x})$	3.44	21.86	34.21
$\pi(\mu(\mathbf{x}))$	3.10	21.68	33.49
w/ D^{l}	7.40	31.84	58.91
w/ $\mathrm{D}^{\mathrm{l+s}}$	7.47	**32.93**	59.78
w/ $\mathrm{D}^{\mathrm{l+s+c}}$	**7.55**	32.71	**61.07**

Table 1: MT scores for direct prompting $\pi(\mathbf{x})$ and prompting with morphologically analyzed sentences $\pi(\mu(\mathbf{x}))$, and with **dictionary** entries of increasing complexities. **Bold**: best result for each column.

Comprehensive dictionary entries are important. As shown in Table 1, using a morphological analyzer alone is not helpful – $\pi(\mu(\mathbf{x}))$ performs worse than $\pi(\mathbf{x})$. This is expected as simply transforming the input sentence to a segmented list of morphemes does not provide the model much knowledge about Manchu. Once the explanations for the lexical entries are included, performance improves significantly. The translation quality can be further improved with the inclusion of suffixes ($\mathrm{D}^{\mathrm{l+s}}$) and then collocations ($\mathrm{D}^{\mathrm{l+s+c}}$). Although the chrF score of $\mathrm{D}^{\mathrm{l+s+c}}$ is slightly lower than $\mathrm{D}^{\mathrm{l+s}}$, both the BLEU and SBERT scores suggest that the $\mathrm{D}^{\mathrm{l+s+c}}$ delivers the best overall translations. Therefore, $\pi(\mu(\mathbf{x}), \mathrm{D}^{l+s+c})$ will be used as a new baseline when assessing the following component. We illustrate the benefits of dictionary information (lexical entries, suffixes, and collocations) for translation in Tables 11 and 12 in Appendix I.

[14]Explanations (in English) document the meaning and function of each suffix. See example in Appendix I.

4.3 Parallel Examples

Parallel examples are drawn from the corpus introduced in §3. Ideally, these parallel examples $\mathcal{P}$ should closely resemble the input sentence, as higher similarity with the source text is known to improve translation quality (Zhang et al., 2024a). To explore this, we construct three variants for $\mathcal{P}$, each exhibiting a different degree of similarity:

- P^{r} includes 10 parallel sentences **randomly** selected as few-shot examples.
- P^{d} includes up to 10 parallel sentences retrieved based on shared terms. As the parallel sentences are extracted from the **dictionary**, they are originally meant to illustrate the meaning of a specific dictionary entry. Therefore, for a lexeme in the input sentence, the parallel examples for its dictionary entry are retrieved.
- P^{bm} includes 10 parallel sentences retrieved using the **BM25** algorithm (Robertson et al., 1995) implemented by Rank-BM25.[15] The terms used by the retriever are morphemes segmented by the analyzer of §4.1.

Variant	BLEU	chrF	SBERT
$\pi(\mu(\mathbf{x}), \mathrm{D}^{*})$	7.55	32.71	61.07
w/ P^{r}	7.66	32.94	60.85
w/ P^{d}	8.10	32.95	61.04
w/ P^{bm}	**8.84**	**33.72**	**61.35**

Table 2: Performance comparison between the baseline (no parallel examples) and 3 ways to select **parallel examples**. **Bold**: best result for each column.

More similar parallel examples improve the translation. As shown in Table 2, randomly retrieved parallel examples provide only a slight improvement over the baseline, as they do not seem to introduce much useful information into the context. On the other hand, selecting parallel examples that are similar to the input sentence yields more noticeable improvements (see lines P^{d} and P^{bm} in Table 2). P^{bm} achieves the best performance across all 3 evaluation metrics, as BM25 aims to retrieve parallel examples that are globally similar to the input sentence; $\pi(\mu(\mathbf{x}), \mathrm{D}^{l+s+c}, \mathrm{P}^{\mathrm{bm}})$ will be used as a new baseline when assessing the following component. An example of how parallel examples help translation is in Table 13 in Appendix I.

[15]https://github.com/dorianbrown/rank_bm25

4.4 Grammar

As mentioned in §3, two grammar books – a short and a more detailed one – serve as source materials. For each book, we manually compile 26 tuples consisting of (*feature, excerpt*), in which each Manchu grammatical feature is paired with the corresponding excerpt from the short or the long grammar book. With our morphological analyzer, we extract a set of grammatical features from the source Manchu sentence and generate a tailored grammar combination accordingly, consisting of only excerpts that are relevant to that sentence. This approach is much more efficient than dumping the entire grammar book into the context. We consider 3 ways to retrieve excerpts G from grammar books:

- G^{s} is a combination of grammar excerpts, retrieved from the **short** book.
- G^{l} is a combination of grammar excerpts, retrieved from the **long** grammar book with more detailed explanations.
- $\mathrm{G}^{\mathrm{l+p}}$ additionally adds **parallel** examples that illustrate the grammar excerpts, which are originally included in the **long** grammar book.

In addition to the excerpts, we include a fixed paragraph shared by all variants, which contains basic information about the word order and typological features of Manchu (see Appendix B).

Variant	BLEU	chrF	SBERT
$\pi(\mu(\mathbf{x}), \mathrm{D}^*, \mathrm{P}^*)$	8.84	33.72	**61.35**
w/ G^{s}	8.26	33.12	60.70
w/ G^{l}	8.46	**33.79**	61.17
w/ $\mathrm{G}^{\mathrm{l+p}}$	**8.90**	33.77	60.40

Table 3: Performance comparison between the baseline without **grammar** and 3 different variants of retrieving grammar excerpts. **Bold**: best result for each column.

Grammars hardly help. As shown in Table 3, G^{s} yields scores worse than the baseline. With more detailed explanations, G^{l} leads to a slight improvement in chrF score, and when further accompanied by parallel examples, $\mathrm{G}^{\mathrm{l+p}}$ leads to a small improvement in BLEU score. Nevertheless, compared to the performance reported for the other components, i.e., dictionary and parallel examples (cf. Tables 1 and 2), the improvement seems marginal and is not reflected in SBERT scores. This suggests that grammars do not help much in in-context MT, which is consistent with the findings reported by Aycock et al. (2024). Nevertheless, we have found instances where grammar explanations could aid translation, such as the example of Table 14 in Appendix I. Moreover, since the next component – CoT – involves grammatical annotation and syntactic analysis, which are closely tied to the information provided in the grammar excerpts, we will still include the grammar component in the new baseline for assessing the CoT component. The variant $\mathrm{G}^{\mathrm{l+p}}$ is selected based on the BLEU score.

4.5 Chain-of-Thought

CoT prompting instructs LLMs to generate a series of intermediate results before solving the final task (Wei et al., 2022). We draw CoT prompt templates from `LingoLLM` (Zhang et al., 2024b), which are explicit instructions provided in the context. We consider 2 variants for CoT prompting C:

- C^{a} asks the LLM to **annotate** the grammatical and semantic features of each word in the sentence before computing the translation.
- $\mathrm{C}^{\mathrm{a+s}}$ asks the LLM to proceed step by step, first to annotate the grammatical and semantic features of each word, then analyze the sentence's **syntactic structure**, and finally produce the translation.

Variant	BLEU	chrF	SBERT
$\pi(\mu(\mathbf{x}), \mathrm{D}^*, \mathrm{P}^*, \mathrm{G}^*)$	**8.90**	**33.77**	**60.40**
w/ C^{a}	8.01	33.13	59.81
w/ $\mathrm{C}^{\mathrm{a+s}}$	8.49	33.43	59.01

Table 4: Performance comparison between the baseline without CoT prompting and 2 variants of CoT. **Bold**: best result for each column.

CoT does not help the model generate better translations. Explicitly prompting the model to perform intermediate generation steps results in a noticeable decline in both C^{a} and $\mathrm{C}^{\mathrm{a+s}}$. This aligns with the findings of Elsner and Needle (2023), where CoT does not improve performance. This discrepancy seems to arise from erroneous or incomplete deductions within the intermediate steps (cf. Table 15 in Appendix I). This further indicates that, even with the CoT prompting, the model is still unable to effectively utilize the grammar. Consequently, we exclude the CoT component and the Grammar component from our final pipeline.

Model	BLEU	chrF	SBERT
Llama3-1B	0.27	9.95	16.37
Llama3-3B	1.81	21.95	38.46
Llama3-8B	3.05	26.59	49.10
Llama3-70B	6.31	31.01	56.82
GPT-4o	8.84	33.72	61.35
DeepSeek-V3	**12.35**	**37.93**	**65.64**

Table 5: Performance of various LLMs using the best setting, i.e., $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$. **Bold**: best result for each column.

5 In-Depth Analysis of Performance

5.1 Performance Across Models

We have so far used the GPT-4o model to assess the importance of each component and its quality, finding that the best setting is $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$. We now study performance variation across models for this setting. Results are in Table 5.

Model size matters. The smallest model, i.e., Llama3-1B, yields an extremely low BLEU score of 0.27. When manually checking the translation, we found that the Llama3-1B model often does not follow the instructions, generating outputs where the translation is difficult to extract or missing entirely. With the size increase in the Llama3 family, we see a consistent improvement in translation scores. Through manual inspection of the translations from varying model sizes (see Table 17 in Appendix I), we observe that larger models not only exhibit better instruction-following abilities but are also better at leveraging the information included in the context. Therefore, we hypothesize that LLM-based MT relies on both good instruction-following and in-context learning abilities, which are closely related to the model size.

The performance could be underestimated The best performance is obtained with DeepSeek-V3, achieving BLEU scores of 12.35. The score is still low, especially when compared with LLM-based translation for high-resource languages (Alves et al., 2023; Sia et al., 2024). However, we often observe that the in-context translations are semantically close to the reference, yet exhibit significant differences in wordings, suggesting that BLEU and chrF scores actually underestimate the MT quality, as illustrated by the example in Table 6. When assessed with SBERT, the best-performing model (DeepSeek-V3) achieves a score of 65.64, indicating a strong semantic similarity between the translation and the reference.

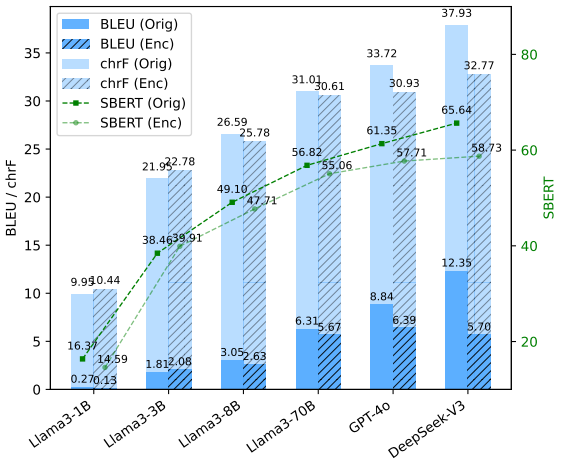

Figure 2: Performance comparison between enciphered $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$ and original $\pi(\mu(\mathbf{x})), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ across multiple LLMs.

Input:	*ereci julesi gurgu elgiyen*
Translation:	From this point forward, wild animals are abundant.
Reference:	From there onwards beasts were plentiful.
BLEU: 4.99	**chrF:** 24.02 **SBERT:** 62.42

Table 6: An example where BLEU and chrF scores (**sentence-level**) underestimate the translation quality, while SBERT better reflects the translation quality.

5.2 Exposing Prior Knowledge of Manchu with Character-Substitution Cipher

We have so far assumed that the MT performance of LLMs is mostly attributed to their **in-context abilities**, rather than to some **prior knowledge of Manchu** that can possibly be acquired during its training stage. To explore this question, we create a "fake Manchu" aimed at eliminating this possible confounding factor. As Yuan et al. (2024) and Marmonier et al. (2025) have demonstrated, LLMs' prior knowledge can be bypassed using substitution ciphers. In this work, we "encipher" all Manchu tokens by a simple character-level substitution cipher as follows: each vocalic character in Manchu (*a, e, i, o, u*), is substituted with the **next** character in this list, e.g., $a \rightarrow e$, $e \rightarrow i$, and $u \rightarrow a$. The same substitution rule applies to consonantal characters in (*b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z*). Using this scheme, a Manchu token *amban* is enciphered as *encep*.

The encipherment applies to all tokens in the input Manchu sentence as well as the linguistic resources involving Manchu, such as dictionary entries and parallel examples, while the English parts remain unchanged. This approach ensures that the LLM can only rely on the information provided in the prompt and its in-context learning ability. Enciphered prompts are denoted with the subscript $_e$ as in: $\pi(\mu(\mathbf{x})_e, \mathrm{D_e}, \mathrm{P_e})$. We experiment with the original template $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ and its enciphered version $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$.[16] The results achieved for multiple LLMs are in Figure 2.

LLMs already know some Manchu. The performance of the enciphered version tends to be slightly lower than the original version for all LLMs. This suggests that all models have seen some Manchu in their pretraining stage, possibly due to contamination – pretraining corpora often contain significant amounts of non-English texts, including many low-resourced ones (Blevins and Zettlemoyer, 2022). The performance drop is particularly noticeable for DeepSeek-V3. We hypothesize that DeepSeek-V3 has seen more Manchu data during its pretraining stage because it was trained on large Chinese corpora, which may contain more Manchu texts.

LLMs rely more on their in-context learning ability. Even though all LLMs have some prior knowledge of Manchu, as indicated by the drop in performance from the original version to the enciphered version, the enciphered versions can still achieve comparable results, and the performance gap remains relatively small (except for DeepSeek-V3). This confirms that LLMs are not fully relying on their prior knowledge, but are rather mainly depending on their in-context learning ability. This argument can be further supported by the consistent performance improvement for both the original and the enciphered Manchu texts when increasing the model size. Since the Llama3 family models are trained on the same data, the observed performance gain, such as from Llama3-1B to Llama3-70B, should be largely attributable to the enhanced in-context learning capabilities of the larger model.

5.3 Validating the Translation Quality Through Human Evaluation

In order to further validate the quality of our MT outputs beyond automatic evaluation metrics such as BLEU, chrF, and SBERT, we have conducted a human evaluation in the form of the Direct Assessment (DA) (Graham et al., 2013). Specifically, we have recruited 3 Manchu language experts who are fluent in both Manchu and English, and have asked them to rate how adequately the English translations express the meaning of their corresponding source sentences in Manchu, on a continuous scale from 0 to 100. The complete instruction given to the human raters is in Appendix H.

For the human evaluation, we randomly select 33 sentences from the evaluation set described in §3. For each sentence, we include the MT outputs of 3 variants using the GPT-4o model: $\pi(\mathbf{x})$ (direct prompting), $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ (the best setting), and $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$ (enciphered version of the best setting), resulting in a total of 99 evaluation items. The identities of the system variants are anonymized. In addition, the order of the items is randomized for the evaluation.

To account for potential differences in how individual raters use the scoring scale, the raw DA scores are normalized to z-scores before being aggregated across raters. The inter-rater agreement among the three raters is strong, with an average Pearson correlation coefficient $r = 0.864$.

We report the average z-scores as well as average raw DA scores in Table 7. The results show that the 2 variants enhanced with dictionary entries and parallel examples achieve substantially higher scores compared to the baseline $\pi(\mathbf{x})$. This further validates the effectiveness of our proposed pipeline. Moreover, $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ achieves higher average scores than the enciphered version $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$.

We run the Wilcoxon rank-sum test to test the statistical significance. The results indicate that both $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$ and $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ differ significantly from the baseline $\pi(\mathbf{x})$, both with $p < 0.001$. On the other hand, although the enciphered version has lower average scores, the difference between $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$ and

Variant	DA score	z-score
$\pi(\mathbf{x})$	29.04	-0.63
$\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$	56.12	0.19
$\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$	**64.47**	**0.44**

Table 7: Average DA scores and z-scores of the MT outputs across the 3 variants $\pi(\mathbf{x})$, $\pi(\mu(\mathbf{x})_e, \mathrm{D}_{\mathrm{e}}^{\mathrm{l+s+c}}, \mathrm{P}_{\mathrm{e}}^{\mathrm{bm}})$, and $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$. **Bold**: best result for each column.

[16] We use $\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ because the grammar excerpts contain a mixture of Manchu and English tokens, which makes it difficult to encipher only the Manchu tokens.

$\pi(\mu(\mathbf{x}), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ is not statistically significant, with $p = 0.27$. This also aligns with our previous finding using the automatic metrics, that the performance gap between the enciphered and the original version is relatively small: LLMs rely more on their in-context learning ability.

6 NMT Data Augmentation

We present a follow-up study where we use our in-context MT system to generate more parallel data for training an NMT model. This data augmentation approach follows the *forward-translation* method (Burlot and Yvon, 2018; Bogoychev and Sennrich, 2020).

Translating Monolingual Corpus. Specifically, we use our in-context MT system to translate 42,240 sentences from the monolingual Manchu corpus (cf. §3) into English, using our best-performing method $\pi(\mu(\mathbf{x})), \mathrm{D}^{\mathrm{l+s+c}}, \mathrm{P}^{\mathrm{bm}})$ with DeepSeek-V3. The resulting synthetic parallel corpus is combined with the real parallel corpus to train an NMT model of Manchu-to-English.

Fine-Tuning mT5. We fine-tune mT5-small (Xue et al., 2021), an encoder-decoder multilingual pre-trained model on the Manchu-to-English translation task. To systematically assess the impact of synthetic data, we use different data-mixing strategies, e.g., only real parallel data, or additionally with synthetic data that is several times larger than the real data. The performance is evaluated on the same evaluation set of 337 parallel sentences (cf. §3).

Figure 3 presents the results. The model trained exclusively on real data performs extremely badly across all metrics, suggesting that 3,520 parallel sentences are insufficient for training an effective NMT model. However, as more synthetic parallel data is introduced, performance improves consistently. The best-performing model – trained with real data and synthetic parallel data that are 12 times larger than the real data – achieves results comparable to or even surpassing Llama3-70B. The resulting fine-tuned mT5-small model only contains around 300M parameters and is significantly more efficient than a 70B-parameter in-context MT system. This study underscores the potential of leveraging in-context MT for data augmentation, enabling the development of more effective and efficient NMT models for low-resource languages.

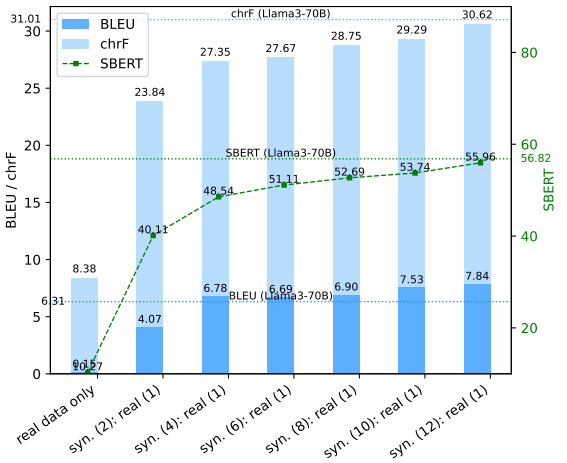

Figure 3: Performance comparison of the fine-tuned mT5 model using only real parallel data versus incorporating varying proportions of synthetic parallel data generated by our in-context MT system. We observe a steady improvement in performance as more synthetic parallel data is added, ultimately achieving scores that match the in-context MT results of Llama3-70B.

7 Conclusion

In this paper, we conduct a comprehensive investigation of in-context MT for low-resource languages, using Manchu as a case study. We examine the impact of different types of resources and the quality of each component on translation performance. Our findings highlight that high-quality dictionaries and properly retrieved parallel examples are the most influential factors, while grammar and CoT prompting appears to have no noticeable benefit. Furthermore, through the encipherment experiment, we disentangle the effects of LLMs' prior knowledge of Manchu from their in-context learning ability. Our results show that while LLMs possess some prior knowledge of the language, they primarily rely on in-context learning for translation. Finally, our follow-up study shows a practical application of in-context MT: generating synthetic parallel data. This approach has the potential to enhance NMT systems, offering a viable strategy for improving translation in low-resource languages.

Limitations

Our current work only includes a single language, Manchu, as a case study. Although our encipherment method can be considered a generalization effort applicable to any language unfamiliar to the LLMs, the encipherment did not alter the fundamental properties of Manchu, which is an aggluti-

native language characterized by a relatively clear separation between morphemes. It is not fully clear whether our findings extend to other typologically distinct languages.

We have only focused on the translation direction from Manchu to English and have not explored the reverse direction. However, if the goal is to produce synthetic parallel data of good quality, we believe it is advantageous to translate authentic low-resource language into a high-resource language that the LLM is proficient in. This ensures fluency and authenticity of the texts in both the source and target languages.

Lastly, we have only explored a limited range of CoT strategies. Our current results indicate that the extra CoT steps often introduce new errors, resulting in a poorer final translation. Future work could investigate ways to mitigate these undesired effects, such as through better prompt engineering or by providing guiding examples for the CoT process.

Acknowledgments

This research was supported by DFG (grant SCHU 2246/14-1). François Yvon has been partly funded by the French National Funding Agency (ANR) under the France 2030 program (ref. ANR-23-IACL-0007) and the Tralalam Project (ref. ANR-23-IAS1-0006). We are deeply thankful to Fresco Sam-Sin of the Manchu Foundation and Professor Hitoshi Kuribayashi from the Tohoku University, for generously granting us permission to use the digitized Manchu materials available on their websites. We also sincerely thank Manchu experts Chen Chen, Sulfa, and Zuoteng Li for their valuable contributions as raters in the human evaluation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2020. Domain, translationese and noise in synthetic data for neural machine translation. *Preprint*, arXiv:1911.03362.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Larry Clark. 1980. *Manchu suffix list*. Department of Asian Languages and Literatures. University of Washington.

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