Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding are Both the Problem

Anonymous ACL submission

Abstract

This work investigates the in-context learning abilities of pretrained large language models (LLMs) when instructed to translate text from a low-resource language into a high-resource language as part of an automated machine translation pipeline. We conduct a set of experiments translating Southern Quechua to Spanish and examine the informativity of various types of information retrieved from a constrained database of digitized pedagogical materials (dictionaries and grammar lessons) and parallel corpora. Using both automatic and human evaluation of model output, we conduct ablation studies that manipulate (1) context type (morpheme translations, grammar descriptions, and corpus examples), (2) retrieval methods (automated vs. manual), and (3) model type. Our results suggest that even relatively small LLMs are capable of utilizing prompt context for zero-shot low-resource translation when provided a minimally sufficient amount of relevant linguistic information. However, the variable effects of prompt type, retrieval method, model type, and language community-specific factors highlight the limitations of using even the best LLMs as translation systems for the majority of the world's 7,000+ languages and their speakers.

1 Introduction

The field has made great progress improving the quality of machine translation (MT) systems, but constraints on the amount and kinds of data available in the majority of the world's 7,000+ languages have led to yet another disparity in access and support for speakers of these languages: low-resource MT continues to be a major challenge (Hendy et al., 2023; Stap and Araabi, 2023; Robinson et al., 2023; Nicholas and Bhatia, 2023). While many of these languages lack the kinds of large, standardized corpora necessary for traditional methods, recent work shows it may be possible to leverage a smaller amount of existing resources, for example pedagogical materials used for language instruction, with Large Language Models (LLMs), albeit with varying results (Tanzer et al., 2024; Zhang et al., 2024; Elsner and Needle, 2023). These materials are often the result of community-driven or government-led initiatives to support language revitalization, reclamation, and mother-tongue education (Schreiner et al.; Riestenberg et al., 2024; Liu et al., 2022). Such discrepancies in the needs and priorities of academic, commercial, and community-led efforts to develop digital resources and language technologies is what Gessler (2022) terms the "NLP Gap".

In this study, we investigate one way to lessen the NLP Gap, comparing LLMs' in-context learning abilities when translating from a low-resource language (a Peruvian variety of Southern Quechua) to a high-resource language (Spanish) using information retrieved from a database of pedagogical materials. We replicate results of earlier studies on a new language pair by comparing the effects of morpheme translations, sentences from a parallel corpus, and passages from a grammar instruction document on translation quality. We then conduct a more focused analysis by annotating translation outputs by hand using a modified MQM error typology (Burchardt, 2013). Finally, we conduct an ablation study on the effects of automated retrieval by manually constructing prompts using the same set of materials.

Our results suggest that while, unsurprisingly, translation quality improves with model size, such improvements seem to primarily be the result of previous exposure to the low-resource language during model pretraining, rather an improved ability for the model to utilize prompt context, as evidenced by high scores in response to baseline (zero-shot) translation prompts. However, we also find evidence that in-context learning abilities may be inconsistent across different models of similar size. As found in previous studies, prompts containing morpheme and word-level translations reliably improve model outputs, but information from the grammar and corpus have a null or even negative effect on results. Human evaluation on a selection of outputs from two models – GPT-3.5 Turbo and GPT-4o – align with the quantitative measures we obtain using BLEURT (Sellam et al., 2020) as an automatic metric. Quantitative results also show an effect of automated retrieval on translation quality that is most evident in prompts containing morpheme translations and for models with lower baseline scores. Finally, we highlight a number of ethical concerns and limitations that arise from the proposed methods that are supported by our findings, and discuss the potential risks and challenges LLM-based methods for low-resource MT face moving forward.

2 LLMs for Machine Translation

Modern LLMs are now capable of translating many high-resource languages, but lack sufficient coverage of even modestly resourced languages to achieve comparable results without additional support (Kocmi et al., 2023). Retrieval-augmented generation (Rubin et al., 2022) may provide such support in the form of parallel sentences (Agrawal et al., 2022), dictionary definitions (Ghazvininejad et al., 2023; Lu et al., 2023) or other linguistic meta-knowledge such as a grammatical description. Retrieval-augmented methods offer exciting possibilities for low-resource translation, since the LLM might (in principle) be able to "teach itself" the language from learner-oriented resources produced by community members or language specialists.

Studies to date (Reid et al., 2024; Zhang et al., 2024; Elsner and Needle, 2023) experiment with four dimensions of variability: source language, LLM, type(s) of information retrieved, and retrieval method. Since the source languages in these studies have relatively little presence in public corpora or on the web, differing results across LLMs can tentatively be attributed to differences in their in-context learning and instruction following abilities.

All studies find that word-level translations are helpful additions to prompts. Zhang et al. (2024) and Tanzer et al. (2024) also add sentence pairs from a parallel corpus, while Elsner and Needle (2023) add usage examples from a dictionary. Each improve results, although to a lesser degree. Elsner and Needle (2023) and Zhang et al. (2024) experiment with small fixed "grammar lesson" passages to provide explicit syntactic instruction, but find these ineffective. Tanzer et al. (2024) uses passages retrieved from a grammar book, also with relatively disappointing results. Reid et al. (2024) use the entire grammar book and a very long-context model to obtain better translations, but without exploring the role explicit grammar instruction actually plays in doing so.

Zhang et al. (2024) find that sentences from the corpus retrieved using BM25 embeddings work better than random ones. Tanzer et al. (2024), however, report that retrieval with longest common substring (LCS) matching outperforms embedding-based retrieval. Overall, the question of how to best retrieve relevant passages containing grammar material or sentences in a low-resource language is still open. This also complicates the interpretation of the mostly-negative results found for grammar passages. It is not clear whether these stem from poor retrieval, from the LLMs' inability to process the retrieved content, or both. Moreover, although Reid et al. (2024) conducts human evaluation of the results for quality, to the best of our knowledge no study to date systematically investigates specific grammatical errors in the output.

3 Quechuan Languages

Quechua is a family of languages indigenous to the Andes in South America. This study focuses on varieties of Southern Quechua (S. Quechua, also known as *urin quechua* or *quechua sureño*) spoken in parts of Peru.[1] While previous studies investigated language/LLM pairs for which the baseline LLM lacked any pretrained knowledge, we find that newer LLMs can translate some S. Quechua sentences in a zero-shot setting. We expect this to be typical of many low-resource languages which, while often endangered, still may have some presence on the web.

Quechuan languages have by far the largest representation of all indigenous Latin American languages in NLP research (Tonja et al., 2024) and are often included in ACL-affiliated workshops, datasets, and shared tasks (Ebrahimi et al., 2022, 2023; Cotterell et al., 2020). S. Quechua has a robust language toolkit (Rios, 2015), including the morphological parser we use in our pipeline. It has also been the subject of numerous studies on MT for both text and speech, developed in conjunction with monolingual and parallel corpora (Rios, 2015;

[1]Unless noted otherwise, we use *Quechua* in this study to refer Southern Quechua and related varieties, following the practices of native speakers with whom we have relationships.

[TAREA] Traduce la siguiente frase del quechua al español.
Responde sólo con la traducción:
quechua: kay wasiqa turiypam
español:

Figure 1: Example BASELINE prompt. English: *[TASK] Translate the following sentence from Quechua to Spanish. Respond only with the translation: Quechua: kay wasiqa tirypam; Spanish:*

Cardenas et al., 2018; Ortega et al., 2020; Zevallos et al., 2022). Nonetheless, such tools continue to face challenges, and Quechuan languages continue to lack the resources necessary to develop most of today's state of the art models.

4 Methods

4.1 Data

We conduct experiments on a collection of 50 pairs of S. Quechua - Spanish sentences sourced from one of the author's personal notes. These were selected to highlight a range of specific grammatical phenomena at multiple levels of difficulty—they include simple clauses and tenses (e.g., *qam allinta tusunki* (tu bailas bien) 'you dance well') as well as more advanced constructions such as those involving past participles (e.g., *awasqay waliqa sumaqmi* (la falda que tejí es linda) 'the skirt I knit is lovely' and simultaneous events (e.g., *qamqa takita uyarispa wasiykita pichachkanki* (tú estás limpiando tu casa escuchando música) 'you're cleaning your house listening to music'. The first author, a foreign-language student of S. Quechua, received permission from her instructor to use notes from their lessons for the study. All sentence pairs were inspected by the instructor, a native bilingual speaker of both S. Quechua and Peruvian Spanish, to eliminate any errors and confirm the accuracy of all reference translations.

4.2 Prompt Construction

As a baseline, each sentence is inserted into a prompt template that instructs the model in Spanish to translate the S. Quechua sentence into Spanish and respond only with the translation (Figure 1). We automate a process for building on this template and compare the effects of adding information from three different sources to the prompt context.

4.2.1 Morpheme Translations (MORPH)

We use a morphological parser (Rios, 2015) to segment each word of the source segment into canonical morphemes, each with gloss symbols and a Spanish translation.[2] Some morphemes have multiple candidate meanings, all of which are retrieved. As an example, the word *rantikuq* is segmented as *ranti-ku-q* and glossed as "comprar.DB.VRoot-DB.VDeriv.+RflxInt-+Ag.NS." While numerous orthographic standards have been developed and promoted across Quechuan-speaking communities in South America, considerable variation in orthographic conventions may be found even within a particular community or variety (Rios and Castro Mamani, 2014). We discuss the implications of this for our results in Section 4.2.5.

We supplement the output from the parser using a Quechua-Spanish bilingual dictionary (Qheswa Simi Hamut'ana Kurak Suntur, 2005). We retrieve any dictionary entry whose headword exactly matches a canonical morpheme in our segmentation. By default, we include all senses and any usage examples or contextual information in the dictionary entry as part of the prompt. We then concatenate the output of the parser with the retrieved dictionary entries and include this MORPH information as prompt context preceding the source sentence and baseline translation prompt.

4.2.2 Grammar Descriptions (GRAMMAR)

We also experiment with the inclusion of grammar lessons found in student-facing pedagogical materials, retrieving grammatical explanations relevant to each source sentence from a PDF document developed for students and teachers of S. Quechua. (Pinto Tapia et al., 2005). The document is organized into short sections (1-3 sentences, plus paradigm tables or usage examples) that describe the particular grammatical concept associated with an affix in Quechua. For each source sentence, we retrieve sections associated with any affix listed in the document that is an exact match of a canonical morpheme and include this in prompts using contextual information from the grammar. This improves on the methods described in Tanzer et al. (2024), who use LCS-based retrieval over an entire textbook, and Elsner and Needle (2023); Zhang et al. (2024), whose grammatical description remains consistent across prompts regardless of the source text being translated.

[2] We set aside valid concerns regarding the theoretical status of the *morpheme* for this study and define a morph(eme) loosely as a recognizable form-meaning pair that recurs in a language.

4.2.3 Parallel Usage Examples (CORPUS)

Finally, we experiment with sentence-level examples from a S. Quechua-Spanish parallel corpus designed for traditional NLP tasks. We combine data made available via the AmericasNLP 2021 Shared Task on Open Machine Translation and the 2023 IWSLT shared task on low-resource SLT (Mager et al., 2021; Agarwal et al., 2023; Agić and Vulić, 2019; Ortega et al., 2020; Tiedemann, 2012). For each source sentence, we retrieve the three best matches from the corpus using a LCS search against the full source sentence.

4.2.4 Combined prompt types

Combinations of information from all three sources yields 8 total conditions, including the baseline. An example prompt from each information source is given in Appendix E.

4.2.5 Manually Revised Prompts

To compute a soft upper bound on the improvements possible with better retrieval, we conduct an additional set of experiments using manually revised prompts. We first examine the content retrieved from the morphological parser, dictionary, and grammar document and remove all instances of ambiguity and irrelevant or misleading information from the prompt context.[3]

For example, many S. Quechua speakers use the term *runasimi* (lit: 'people mouth', 'the people's language'), as an endonym for the language. The parser, however, returns only the literal decomposition (*runa* 'ser humanos'/'people' and *simi* 'boca'/'mouth'), and the dictionary does not list *runasimi* as a headword but rather as one of eight different senses of *simi*. We thus remove all such irrelevant examples and translations from the prompt and retain only the content indicating a translation of *runasimi* in the linguistic sense.

We also manually retrieve content from the dictionary and grammar documents that were overlooked by the automated retriever. For example, the verb *yanuy* 'to cook' does not appear as a headword in the dictionary, but rather as a regional variant of *wayk'uy* 'to cook'. We also eliminate content from the grammar that was retrieved because of syncretism, or mistakes that cascaded from the morphological parser to result in irrelevant retrievals.

[3]We do not experiment with retrieval methods for corpus examples, which were retrieved using LCS match in both conditions. Improving on LCS-based retrieval remains an open question in low-resource LLM-MT, and we leave this for future work.

We manually parse each source sentence to only retrieve and include relevant information in the prompt context. All content in the revised prompts is sourced from the same material available to the automated retriever systems, and we do not add any additional information or use supplemental materials of any sort to create the revised prompts.

4.3 Models

We experiment with three proprietary models, GPT-3.5 Turbo (`gpt-3.5-turbo-0125`, Brown et al., 2020), GPT-4o (`gpt-4o`, Achiam et al., 2023), and Gemini 1.5 Pro (`gemini-1.5-pro`, Reid et al., 2024), and one open-source model, Llama 3 (`llama-3-8b-instruct`, AI@Meta, 2024). We use the pretrained models with their default settings, and do not adjust hyperparameters or conduct any finetuning as part of our experiments.

4.4 Evaluation

We conduct both automatic and human evaluations to identify trends in model errors and outputs in the various experimental conditions. We use BLEURT as an automatic metric, and report mean BLEURT scores across items as the primary quantitative measure of translation quality for each of the conditions and models. We also use an adapted MQM schema to conduct qualitative human evaluation of the outputs of GPT-3.5 and GPT-4o for all prompts with automatic retrieval.

Each item selected for human evaluation is annotated by at least one of the authors by comparing the model's output to the source text and reference translation.[4] We refer to the complete MQM typology to design our own four-dimensional framework of commonly attested errors in LLM-MT, each with a defined set of specific subtypes. Precise definitions and examples for all error categories and subtypes may be found in Appendix D.

Many of the categories in our schema are defined as in the core MQM framework. However, to capture some of the key behaviors reported in previous studies on LLM-MT and to evaluate the effects of prompt type on model outputs, we make the following adjustments. First, we utilize the Addition and Omission errors defined as Accuracy subtypes in the original MQM typology, but distinguish these from three additional subtypes: Substitution - Incorrect Subject, Substitution - Incorrect

[4]We discuss limitations on this process given the authors' respective proficiency levels in S. Quechua and Spanish, as well as the steps we take to address them, in Section 8.

Tense/Aspect/Modality (TAM), and Substitution - Other. This is intended to capture LLM translations that differ from the source in terms of discrete lexical material or case, person, number, and/or TAM markings while otherwise maintaining the lexical and structural content needed to appropriately translate the source text.

Rather than including Mistranslation and MT Hallucination as Accuracy Error subtypes as in the original MQM typology, we define a separate Non-Translation category with three possible subtypes: Complete Mistranslation, Mistranslation with Lexical Correspondences, and Refusal. The third dimension of our typology, Model Error, was ultimately not used to classify any output in this study, but characterizes more generic model "misbehavior" such as failing to follow instructions, producing garbled text, or inappropriately generating content in the source language. Finally, Target Errors identify outputs that are ungrammatical, stylistically inappropriate, or semantically incoherent in the target language, regardless of their accuracy.

Detailed annotation guidelines were drafted and agreed upon to encourage consistency across annotators and experimental items. Annotators are instructed to identify and tag up to three specific errors for each translation output, with the exception of Target Errors, which do not count towards the three-error maximum. Each model output is also tagged for quality along a four-point scale as defined in Table 7.

Before proceeding with annotation over the larger dataset, both annotators also completed a test evaluation of the same 12 experimental items (96 sentences total) to assess inter-annotator agreement. Statistical measures (κ = 0.72 for quality judgments, α = 0.55 for error categories) indicated some discrepancies in annotator judgments, especially for categories, since determining the three most important errors is especially subjective. These were identified and discussed, and agreement was ultimately deemed sufficient to proceed.

5 Results

5.1 Quality metrics

We present BLEURT scores for prompts generated using automated retrieval in Table 1 and summarize human quality judgments for GPT-3.5 and GPT-4o with automated retrieval in Table 2. The complete distribution of quality ratings across all prompt types for these two models is provided in Appendix

	GPT3.5	GPT4o	Gem.	Lla3
BASE	0.19	0.66	0.56	0.15
CORPUS	0.27	0.59	0.49	0.19
GRAM	0.23	0.56	0.55	0.17
MORPH	0.44	0.54	0.61	0.39
C+G	0.26	0.59	0.54	0.21
C+M	0.44	0.59	0.59	0.36
G+M	0.41	0.53	0.61	0.39
C+G+M	0.43	0.57	0.61	0.15

Table 1: Mean BLEURT scores by LLM and prompt type. Shaded rows include morpheme contexts.

LLM	GPT-3.5	GPT-4o
BASE	21	108
CORPUS	43	101
GRAMMAR	33	99
MORPH	79	102
CORPUS-GRAMMAR	41	101
C+M	75	110
G+M	68	100
C+G+M	77	109

Table 2: Human-annotated quality ratings summarized as $3 \times high + 2 \times med + low$. Shaded rows include morpheme contexts.

	GPT3.5	GPT4	Gem.	Lla3
G-AUTO	0.23	0.56	0.55	0.17
G-MAN	0.24	0.58	0.54	0.15
M-AUTO	0.44	0.54	0.61	0.39
M-MAN	0.56	0.63	0.66	0.49
CGM-AUTO	0.43	0.57	0.61	0.15
CGM-MAN	0.54	0.63	0.63	0.26

Table 3: Comparison of mean BLEURT scores for automatic versus manual retrieval of material in GRAMMAR, MORPH, and CORPUS-GRAMMAR-MORPH prompts.

F. We find clear effects of LLM, prompt type, and retrieval method, as well as interactions among all three factors.

Gemini and GPT-4o outperform Llama 3 and GPT-3.5 for every prompt type. This gap is highest for the least informative prompts, indicating that the Llama 3 and GPT-3.5 base models have relatively poor coverage of S. Quechua, while GPT-4o and Gemini have much better coverage. The effect is evident in both BLEURT scores and human quality evaluations.

Effects of prompt type are mediated by the quality of the pretrained model. Llama 3 and GPT-3.5 show a clear improvement in quality when MORPH information is included in the prompt. Gemini also improves when this information is added, but to a lesser extent. GPT-4o, on the other hand, performs best in response to the BASELINE (zero-shot) prompts, which attain the highest BLEURT scores across all models, prompt types, and retrieval methods evaluated in this study. In other words, providing additional information in the prompt's context actually *degrades* GPT-4o's ability to translate from S. Quechua to Spanish in all experimental conditions.

5.2 Effects of Automated Retrieval

To highlight the effects of automated retrieval on model output, we present BLEURT scores for a selection of prompt types and all four models in Table 3 (full scores may be found in Appendix F). The effect of manual retrieval for MORPH information is positive for all models, although this gap is smallest for Gemini (probably because its performance for these prompts is already highest). The effect for GRAMMAR prompts is either minor or negative.

5.3 Human Analysis of Translation Errors

The most common error type identified by the annotators is Substitution - Other, which includes a diverse assortment of lexical and phrasal incongruencies of varying degrees of severity. These are largely item-specific and therefore hard to characterize as a group. Using the error categories described in Section 4.4, we instead identify three more clearly interpretable phenomena and provide a detailed discussion of each in the following sections. We present counts for selected prompt types in Table 4, with examples in Appendix A and counts for all errors in Appendix G.

5.4 Mistranslations

Outright mistranslations are most common for GPT-3.5, making up 30 of the 50 responses in the BASELINE condition. We also consider outputs that retain only minimal traces of the source content, which we label as Mistranslations with Lexical Correspondence. Approximately 1/3 of the 637 total errors tagged across all prompt types for GPT-3.5 are mistranslations of either type, roughly split between complete mistranslations and those with lexical correspondence (15.07% and 18.37%, respectively, of all errors tagged for GPT-3.5).

As reported in previous work, adding morpheme- and word-level translations to the prompt greatly reduces the rate of this kind of response. GPT4o also produces drastically fewer mistranslations compared to its predecessor. However, it is notable that both models produce at least one mistranslation for each prompt type. In general, complete mistranslations are in fluent Spanish and contain no overt indications that something has been misrepresented. We return to the ethical implications of these errors in the Discussion.

We also note that many of the items tagged as Mistranslation with Lexical Correspondence show correspondence only for words that were already in Spanish in the source text. For example, some sentences contain Spanish loan names for the days of the week. While some of these errors are produced using deceptively fluent Spanish as described above, we find many to be accompanied by semantic incoherence or ungrammaticality in the output. We discuss such target language fluency errors in the following section.

5.5 Target Fluency

Target Fluency errors occur when the output is not grammatical, coherent, or stylistically appropriate – for instance, if an output contains a nonsensical repetition or a verb with missing arguments. Outputs of this type bear a strong similarity to human "translationese" in that structural features of the source language may surface in the translation at the expense of naturalness (Freitag et al., 2019; Koppel and Ordan, 2011). Both GPT-3.5 and GPT-4o tend to produce more such outputs when the prompt is more informative – 10 to 20% of the time (5-10 instances per 50) in prompts with morpheme translations.

		BASE	MORPH	C+G+M
Mistranslation: complete + lexical correspondence	GPT-3.5	45	11	12
	GPT-4o	4	6	4
Target Fluency: grammar + coherence + style	GPT-3.5	0	14	10
	GPT-4o	3	13	9
Grammatical Divergence: subject + TAM	GPT-3.5	0	24	31
	GPT-4o	17	13	11

Table 4: Counts of human-annotated error types (per 50 sentences) by LLM and prompt type.

5.6 Grammatical Divergence

We group misrendered verbal subjects and tense/aspect/morphology (TAM) markers together as Grammatical Divergence errors. Such errors are distinct from the Target Fluency errors described in the previous section— the Spanish output is grammatical, but fails to accurately reflect the syntax of the source. Although they are not the only grammatical phenomena that may be similarly misrendered, we select subject and TAM markers for analysis as they are straightforward to identify and give a good indication of how well the LLMs cope with more abstract information about the meanings of functional morphemes. TAM divergences are much more prevalent than divergences in subject; for instance, only one of GPT-4o's 13 Grammatical Divergence errors in the MORPH condition misrender the subject marker.

Grammatical Divergence errors are annotated only for sentences that are not mistranslated outright, so GPT-3.5 produces none of these in the BASELINE condition. For more informative prompts, it is clear that GPT-4o is better than GPT-3.5 at translating both functional and lexical meanings. However, a relatively large number of sentences (over 20%) still contain such an error even with the highest performing model and prompt type. The relatively small drop in error between different prompt types for GPT-4o suggests that neither the corpus-based usage examples nor example paradigms and descriptions from the grammar document can fully prevent this type of error.

6 Discussion

We observe large differences between LLMs, both in terms of the overall quality of their generated translations as well as the effects of prompt type on their outputs. GPT-4o and Gemini, which have the highest baseline scores, benefit least from additional information— their performance with CORPUS and GRAMMAR information actually decreases. This occurs even with manually curated prompts, suggesting it is not an effect of including irrelevant material. On the other hand, it does not represent a ceiling on quality, since both models continue to make errors (GPT-4o produces 10 LOW-quality translations in our set of 50). These results suggest that even relevant grammar explanations, when written in prose with examples, do little to help the newest generation of LLMs to translate a low-resource language such as Southern Quechua.

Although GPT-4o and Gemini results are similar in many ways, we do find evidence for differences in their in-context learning abilities. Baseline prompts and the GPT-4o model produce the highest BLEURT scores across the dataset, but these outputs still show a number of errors characteristic of LLMs, particularly lexical substitution errors that are not necessarily corrected with the inclusion of more context. In contrast, Gemini, which has near-comparable performance across prompt types, shows an increase in scores when prompts include MORPH information, regardless of retrieval type, suggesting a greater ability to identify and utilize relevant word- and morph-level translations in the prompt's context. Previous work suggests that newer builds of GPT-4 are less capable of following instructions (Chen et al., 2023); such differences may be masked by the effects of pretraining when automatically evaluating translations. This suggests that researchers should continue to carefully select and compare among different LLMs when experimenting with retrieval-based translation.

Finally, we identify a number of translation errors of varying types that appear to be due to language-specific characteristics, for example ambiguity from syncretism in grammatical markers, polysemous lexical items, or the orthographic and lexical variation discussed in Section 4.2.5. It may be possible to moderate such effects with additional refinement of the retrieval database and methods, which we leave for future work.

6.1 Ethical concerns

Both our work and much of the previous work in this paradigm is motivated by the desire to close the "NLP Gap" among researchers, community members, and software developers interested in low-resource language technologies. Machine translation is listed as a welcome topic of research by some (though not all) members of American indigenous communities (Mager et al., 2023), and is potentially an important tool for language learners (Jolley and Maimone, 2022). Even an imperfect translation system might be a useful tool for users with a clear understanding of its limitations. However, the systems evaluated in this work have two problematic tendencies that limit their potential for deployment in real community settings.

First, unfaithful translations often tend to be highly fluent (Section 5.4). While fluency ratings for older MT systems correlate well with accuracy scores, and have even been used as a proxy for overall translation quality (Gamon et al., 2005; Estrella et al., 2007), this correlation is reversed for our systems. LLMs are well-known for making false statements that seem plausible and authoritative (Bickmore et al., 2018; Dinan et al., 2021); this could be particularly problematic when they project illusions of expertise at the expense of an already marginalized group.

Second, some mistranslations identified in our study appear to draw on stereotypes of indigenous groups (Appendix 6). These are most apparent for the BASELINE system and GPT-3.5, but also (less frequently) occur with more informative prompts and better LLMs. Stereotypical sentences can involve flowery language with an emphasis on tradition or connectedness to nature (Erhart and Hall, 2019), as well as the unprompted addition of indigenous Andean cultural customs and products (traditional medicine, chicha) to translations that are otherwise faithful to the source text. The overall effect is to exoticize Southern Quechua speakers and writers in ways that the original sentences do not. Similar stereotypes have also been noted in LLM-generated responses to open-ended prompts (Cheng et al., 2023; Delgado Solorzano and Toxtli, 2023; Shieh et al., 2024).

While we prompt models to output only the translation for evaluation purposes, models may have some capacity to explain or qualify their translations and give reminders for responsible use of the technology. Should a retrieval-based translation system ever be deployed in a real-world setting for language learning, its developers should maximize transparency by presenting the content of any retrieved information and its source to the user along with the translation, reminding users directly of potential inaccuracies, and offering vetted resources for additional fact-checking when available.

7 Conclusion

Our results suggest a number of key limitations and concerns regarding the use of LLMs in a low-resource MT context, and have greater implications for our understanding of the seemingly "humanlike" conceptual, analytical, and in-context learning abilities of LLMs.

For the majority of the world's language communities and their speakers, powering and supplying LLMs with enough pretraining data to overcome their limitations is not feasible. We therefore offer the following suggestions to those looking to develop low-resource LLM-MT: (1), improve data structures and methods for interacting with a language-specific database for retrieval-aided generation. (2), continue analysis of the mechanisms driving in-context learning in LLMs, for example by comparing ICL to the effects of finetuning (Dai et al., 2023), (3) experiment with prompt structures and techniques, for example by altering the order information (Liu et al., 2024) or by iteratively or prompting the model to guide its reasoning towards a suitable translation (Wang et al., 2022).

Finally, we wish to emphasize the continued risks of prematurely deploying this or similar methods in any low-resource language community, particularly given the vulnerability and disproportionate lack of resources many such communities face in domains where these technologies would likely be used. As AI research continues to rapidly develop, we urge those conducting it to increase community engagement, amplify the voices of those traditionally at a disadvantage, and collaboratively develop research infrastructures that may lessen the NLP Gap. While there's still much to be done before low-resource LLM-MT may be safely implemented, we believe such a tool has the potential to empower speakers of any variety, including nonstandard varieties of traditional "high-resource" languages such as English, to develop technologies that reflect their preferences and serve their unique needs.

8 Limitations

Limitations on the scope and replicability of this work may be attributed to one or more characteristics of the data and models used in this study, in addition to limitations inherent to the respective identities of its authors. First, the BLEURT scores we report are limited in their statistical validity. We have conducted some constrained tests to explore potential variance in scores, but expenses associated with text generation using proprietary models such as those developed by OpenAI and Google on a larger dataset may be prohibitive. This is compounded by the widely-acknowledged "black box" nature of the models powering both LLMs and BLEURT, as well as an increasing opacity with respect to the exact content and methods used to pretrain modern state of the art LLMs. For this reason, we focus our discussion on those results that show clear trends in both the quantitative and human evaluations we conduct.

There are also some constraints on our study and its methodology that are largely tied to linguistic factors, such as variation in orthography (and the need for digitized text-based resources as a prerequisite) as well as the lexical and grammatical variation that may be found in all languages, particularly the low-resource varieties we wish to support. Our results suggest it may be possible to guide the outputs of LLMs towards the specific usage conventions of a given community, but this is itself limited by the content of the materials used to develop the database from which prompt contexts are retrieved. Neither of the authors is a native speaker of any Quechua or Spanish varieties, and only one is a student of these languages with relationships to Quechua speakers and communities. While we have strived to be consistent in the Quechua and Spanish varieties used in our study (both the dictionary and grammar materials were provided by the same instructor who shared and proofread the 50 sentence pairs we use, and we select a morphological parser and corpora intended for use with Southern Quechua), variation is widespread among and within Quechua-speaking communities, and we do not have access to a dictionary, grammar, morphological parser, and corpus developed by a unified and consistent set of authors. Such variation is language- and community-dependent and bound to constrain potential applications of our methods. Future work should continue to explore ways to faithfully represent the diversity of linguistic conventions employed by communities interested in developing such technologies.

We acknowledge, as well, limitations that arise from the size of our dataset and database and the methods used to curate them. The 50 sentence pairs we use were selected to highlight a range of specific grammatical phenomena, not all of which were well represented in our database, and differ in their structural complexity. We are grateful for the guidance provided by the Quechua instructor whose lessons were a source for such examples and proofread the sentences before their inclusion in our experiments, but limited as well by our status as non-native speakers. Human evaluation of model outputs was for this reason primarily constructed using machine-translated English texts as references, but was inspected by the Spanish- and Quechua-speaking author to remove a small number of evaluations that reflect linguistic discrepancies between Quechua, Spanish, and English or inaccuracies in the machine translated English.

9 Ethics Statement

We consulted the first author's Quechua instructor, who gave us permission to use the sentences from the notes in this project and verified their accuracy. The instructor will be acknowledged by name if the paper is accepted. We cite the Quechua dictionary and grammar materials used to provide prompt information, and believe that our use of these materials is consonant with their original purpose. However, we do not distribute machine-readable versions of them as a contribution of this project, since this would violate the rights of the publisher.

The authors annotated the translation output themselves, so no human subjects approval/consent/compensation was required.

There are numerous ethical issues related to the training and use of LLMs, such as labor issues and energy costs. While these issues are inextricable from the methods used in this project, we believe the potential impact of making low-resource translation viable and accessible to minority language communities who want them (our primary goal in this line of research) outweighs the problems inherent in using LLMs at all.

We discuss the potential risks of deploying systems like the ones described here futher in Section 6.1 of the main text.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. See also: https://openai.com/index/hello-gpt-4o/.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, et al. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *Preprint*, arXiv:2212.02437.

AI@Meta. 2024. Llama 3 model card.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of siri, alexa, and google assistant. *J Med Internet Res*, 20(9):e11510.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165. See also: https://openai.com/index/new-embedding-models-and-api-updates/.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP*, 2:21.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Analyzing chatgpt's behavior shifts over time. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2020. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. *Preprint*, arXiv:1810.07125.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *Preprint*, arXiv:2212.10559.

Cecilia Delgado Solorzano and Carlos Toxtli. 2023. Evaluating machine perception of indigeneity: An analysis of chatgpt's perceptions of indigenous roles in diverse scenarios.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *Preprint*, arXiv:2107.03451.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, et al. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. 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.

Ryan S Erhart and Deborah L Hall. 2019. A descriptive and comparative analysis of the content of stereotypes about native americans. *Race and Social Problems*, 11:225–242.

Paula Estrella, Andrei Popescu-Belis, and Maghi King. 2007. A new method for the study of correlations between MT evaluation metrics. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages: Papers*, Skövde, Sweden.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

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.

Jason R Jolley and Luciane Maimone. 2022. Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal*, 14(1):26–44.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, et al. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *Preprint*, arXiv:2305.06575.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *Preprint*, arXiv:2306.07377.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Miguel Ángel Pinto Tapia, Luis Quispe Zúñiga, et al. 2005. Didáctica quechua i. Documento de trabajo, Dirección Regional de Educación Apurímac | Dirección Gestión Pedagógica. Elaborado por Mgt. Miguel Ángel Pinto Tapia, Esp.DREA EBI-RURAL-HUASCARÁN. Contact: pitman_38@hotmail.com, mapt_38@yahoo.es.

Qheswa Simi Hamut'ana Kurak Suntur. 2005. *Diccionario Quechua - Español - Quechua Qheswa - Español - Qheswa Simi Taqe*, 2 edition. Multiservicios e Imprenta Edmundo Pantigozo EIRL, Cusco, Peru.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Katherine J. Riestenberg, Ally Freemond, Brook Danielle Lillehaugen, and Jonathan N. Washington. 2024. Prioritizing Community Partners' Goals in Projects to Support Indigenous Language Revitalization. In *Decolonizing Linguistics*. Oxford University Press.

Annette Rios. 2015. *A basic language technology toolkit for quechua*. Ph.D. thesis, University of Zurich.

Annette Rios and Richard Castro Mamani. 2014. Morphological disambiguation and text normalization for southern quechua varieties.

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.

Sylvia L.R. Schreiner, Lane Schwartz, Benjamin Hunt, and Emily Chen. Multidirectional leveraging for computational morphology and language documentation and revitalization. *Language documentation and conservation*, 14.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Preprint*, arXiv:2004.04696.

Evan Shieh, Faye-Marie Vassel, Cassidy Sugimoto, and Thema Monroe-White. 2024. Laissez-faire harms: Algorithmic biases in generative language models. *Preprint*, arXiv:2404.07475.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. Nlp progress in indigenous latin american languages. *Preprint*, arXiv:2404.05365.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. *Preprint*, arXiv:2203.08383.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. *Preprint*, arXiv:2402.19167.

A Example errors

Original	Reference	Translation	Model - Prompt (Error)
		Mistranslation	
ñuqaqa sapa punchaw llamkani aylluyta uywanaypaq	yo trabajo todos los dias para mantener a mi familia; *I work every day to support my family*	yo sé que en algún momento volveré a mi comunidad; *I know that at some point I will return to my community*	GPT-3.5 - BASELINE (Complete Mistranslation)
ñuqaqa illani runakunawan rimayta munaptiy	yo viajo porque me gustar hablar con las personas; *I travel because I like to talk to people*	yo quiero hablar con las personas cuando amanece; *I want to talk to people when the sun rises.*	GPT-4o C+G+M (Mistranslation - Lexical Correspondence)
		Fluency	
qamqa yachayta tukuptiyki, puñunki	cuando termines de estudiar, tu duermes; *when you finish studying, you sleep*	acabaré de aprender lo que tú duermes; *I will finish learning what you sleep*	GPT-3.5 - C+G+M (Coherence: unresolvable pronoun)
payqa lunesmanta jueveskama kallpan tullu niraq kachkananpaq	él corre de lunes a jueves para estar delgado; *he runs from monday to thursday to be slim*	él sigue siendo delgado desde el lunes hasta el jueves para estar; *He remains thin from Monday to Thursday to be.*	GPT-4o - GRAMMAR (Grammar: missing predicate with ‘estar’)
		Grammar	
qam allinta tusunki	tu bailas bien; *You dance well*	tú bailarás bien; *You will dance well*	GPT-3.5 - C+G+M (TAM: tense)
qamqa mamaykipa wasinta yanuq rinki	tú irás a cocinar a la casa de tu mamá; you will go to cook at your mother’s house	voy a cocinar en la casa de tu madre; *I’m going to cook at your mother’s house*	GPT-3.5 - MORPH (Subject: ‘you’ to ‘I’)
qamqa misita qawaspa qari warata awachkarqanki	tú estabas tejiendo un pantalón de varón mirando al gato; *you were knitting a pair of men’s pants looking at the cat*	tú mirabas al gato mientras el hombre tejía el pantalón; *You looked at the cat while the man knitted the pants.*	GPT-4o - C+G+M (Subject: ‘you’ to ‘man’)
qam runasimita yachapayaptiyki, allintam rimanki	si practicas el quechua lo hablarás bien; *If you practice Quechua you will speak it well*	tú ya sabías el runasimi, hablas muy bien; *you already knew runasimi, you speak very well*	GPT-4o - MORPH (TAM: missing conditional)

Table 5: Examples of errors analyzed in Section 5.3, one error per type.

B Unprompted stereotypes

Source	Reference	Translation	Model - PROMPT TYPE
payqa qatupi ñañanpa rantisqan apita rurarqan	ella preparó una mazamorra que compró su hermana en el mercado; *she prepared a mazamorra [pudding] that her sister bought at the market*	el sol brilla en nuestra tierra, calienta y nos da vida; *The sun shines on our land, warms us and gives us life.*	GPT-3.5 - BASE
paykunaqa yachachiqninkupa rakta medicina liwru apasqanta likanku	ellos leerán el libro grueso de medicina que trajo su profesor; *They will read the thick medical book that their teacher brought*	los médicos de la comunidad nos enseñaron a utilizar plantas medicinales; *The doctors from the community taught us how to use medicinal plants.*	GPT-3.5 - GRAMMAR
qamkunapa taytamamaykichikta watunaykichik, paykunata purikuq apanaykichik	ustedes tienen que visitar a sus padres y llevarlos a pasear; *you have to visit your parents and take them for a walk*	ustedes deben honrar a sus padres, guiarlos en su camino; *You must honor your parents, guide them on their path.*	GPT-4o - BASE
ñuqaqa upyayta munani	yo quiero beber; *I want to drink*	yo quiero beber chicha; *I want to drink chicha*	Gemini - BASE

Table 6: Examples of errors involving cultural stereotypes (Section 6.1).

C Quality Descriptions

Quality	Description
High	Output is an accurate and/or acceptable translation of the source content.
Med	Output contains errors that prevent it from being an acceptable translation, but is generally high in quality otherwise.
Low	Output contains errors that prevent it from being an acceptable translation, with minor correspondences that vaguely identify it as relevant to the source.
None	Output does not appear to be relevant to the source.

Table 7: Quality Descriptions

D Annotation error typology

Dimension	Error	Description
Accuracy	Addition	Translation includes information not present in the source, but does not result in the displacement of source content.
Accuracy	Omission	Translation is missing content from the source.
Accuracy	Substitution - Subject	The translated segment contains content identified as relevant to the source in other spans, but substitutes novel subject markers for those present in the source in the highlighted span; Classify an error as a “substitution” when the error appears to result in both Addition and Omission errors that cannot be distinguished into two distinct spans.
Accuracy	Substitution - TAM	The translated segment contains content identified as relevant to the source in other spans, but substitutes novel TAM for those present in the source in the highlighted span; Classify an error as a “substitution” when the error appears to result in both Addition and Omission errors that cannot be distinguished into two distinct spans.
Accuracy	Substitution - Other	Substitution errors that do not involve mistranslated subject markers or TAM. See above.
Accuracy	Overtranslation	Error occurring in the target content that is inappropriately more specific than the source content.
Accuracy	Undertranslation	Error occurring in the target content that is inappropriately less specific than the source content.
Target Error	Grammar	Other spans in the translated segment may be identified as relevant to the source, but the highlighted span is not grammatical in the target language.
Target Error	Coherence	Other spans in the translated segment may be identified as relevant to the source, but the highlighted span is unnatural or incoherent in the target language.
Target Error	Style/Register	Other spans in the translated segment may be identified as relevant to the source, but the highlighted span is produced in a style or register that is inappropriate given the content.
Non-Translation	Complete Mistranslation	The entire segment is coherent in the target language but the core predicate shows no immediate connection to the reference translation.
Non-Translation	Mistranslation - Lexical Correspondence	The entire segment is coherent in the target language but only minor correspondences to the reference translation may be identified.
Non-Translation	Refusal	Model does not attempt to translate into the target language, e.g., because it "does not understand".
Model error	Garbled	Output does not contain coherent text in the target language.
Model error	ChattyGPT	Output contains translated content, but is wordy, over-explanatory, and/or abruptly truncated.

Table 8: Adapted MQM typology for human error annotation

E Example Prompts

The following are examples of prompts generated used automated retrieval from the database.

BASELINE

[TAREA] Traduce la siguiente frase del quechua al español. Responde sólo con la traducción:
quechua: qam allinta tusunki
español:

MORPHS-ONLY

[CONTEXTO]
qam: [PrnPers+2.Sg]
allin: bueno [D̂B][NRoot]
ta: [+Acc][Cas]
tusu: bailar [VRoot][D̂B]
nki: [+2.Sg.Subj][VPers]
allin. adj. Bueno (término de aprobación). SINÓN: kusa. EJEM: allin p'unchay, buenos días: allin tuta, buenas noches; allin tutamanta, buena mañana, buenos días; allin inti chinkay, buenas tardes; allin iñiyniyoq, de buena fe, fiel, justo, íntegro: allin nunayoq, de espíritu bueno; allin puriq, de comportamiento bueno; allin puriy, comportamiento bueno; allin rikuy, tratamiento bueno; allin rikuq, el que trata bien; allin ruway, obrar bien, beneficiar; lo que se hace bien, beneficioso; allin ruwaq, el que hace bien; allin yuyay, pensar bien; pensamiento bueno; allin qolqeyoq, poseedor de plata fina; adinerado.
ta. s. Gram. Sufijo que desempeña los papeles de artículo y preposición. EJEM: llamata qatiy, arrea la llama; Urkusmanta hamuni, vengo de Urcos.

[TAREA] Traduce la siguiente frase . . .

GRAMMAR-ONLY

[CONTEXTO]
ta: CASO ACUSATIVO. Su marca es –ta, esta es una marca de objeto directo con los verbos que no son de movimiento (quietud). Ejemplo:
Quyllur–ta qhawani Veo una estrella
T'anta–ta apay Lleva pan
Ñuqa quylluyta qhawani
Pedrucha t'antata rantin
En cambio con los verbos de movimiento –ta indica (hacia) que es igual a meta. Ejemplos:
Punu–ta rini Voy a Puno
Llaqta-ta risaq Iré al pueblo
Hamawt'anchis Punuta rinqa
Llanta umalliq llaqtata richkan
nki: FLEXIÓN DE TIEMPO. TIEMPO FUTURO. TIEMPO FUTURO. Los sufijos para cada una de las personas gramaticales son: saq, nki, nqa, sun, saqku, nkichis, nqaku; en singular y plural respectivamente.
Ejemplos:
Puklla-saq jugaré
Puklla-nki jugarás
Puklla-nqa jugará
Puklla-sun jugaremos
Puklla-saqku jugaremos
Puklla-nkichis Uds. jugarán
Puklla-nqaku ellos jugarán

[TAREA] Traduce la siguiente frase . . .

CORPUS-ONLY

[CONTEXTO]
quechua: rimanakunapaq wawakunapa rimasqan simi aswan allinta takyachinaraq piwanpas maywanpas mana manchakuspa rimananpaq chaymi qillqanapaqpas ñawichanapaqpas aswan allin kanqa
español: para este diálogo saber la lengua que dominan los niños sería importante para que ellos se expresen sin miedo de ahí será que la escritura y la lectura salga de manera óptima
quechua: kay tiqsipi sumaq rimanakunapaqa kawsayninchikmi allinta kallpachawanchik runakunahina allinta tiyanapaq chaymi ñuqanchikkqa allinta ñawichayta qillqayta yachananchik ñawpa ayllunchikkuna rurasqankuta maytukunapi tukuy puyñukunapi tiqsi muyu qhawarisqankuta
español: para vivir en armonía tenemos que conocer bien nuestra forma de vivir y luego escribir leer tambien a valorar lo que nos dejaron nuestros antecesores en cada visión sobre el mundo
quechua: winsislawcha chayarqamuptinsi tuparquspanku allinta qatunakusqanku suwakuypi purinankupaq
español: cuando había llegado wenseslau y a su encuentro se habían reforzarón para andar a robar

[TAREA] Traduce la siguiente frase . . .

F Full quality scores

This section contains the full tables of BLEURT and human-annotated quality scores. Table 9 contains the full results summarized in Tables 1 and 3 of the main text. Table 10 and Table 11 contain the full scores summarized in Table 3.

	GPT-3.5		**GPT-4o**		**Gemini-1.5**		**Llama 3**	
Prompt	auto	manual	auto	manual	auto	manual	auto	manual
BASELINE	0.19	0.22	0.66	0.66	0.56	0.57	0.15	0.16
CORPUS-ONLY	0.27	0.29	0.59	0.61	0.49	0.47	0.19	0.18
GRAMMAR-ONLY	0.23	0.24	0.56	0.58	0.55	0.54	0.17	0.15
MORPH-ONLY	0.44	0.56	0.54	0.63	0.61	0.66	0.39	0.49
CORPUS-GRAMMAR	0.26	0.28	0.59	0.59	0.54	0.53	0.21	0.21
CORPUS-MORPH	0.44	0.52	0.59	0.64	0.59	0.64	0.36	0.38
GRAMMAR-MORPH	0.41	0.54	0.53	0.61	0.61	0.64	0.39	0.37
CORPUS-GRAMMAR-MORPH	0.43	0.54	0.57	0.63	0.61	0.63	0.15	0.26

Table 9: BLEURT scores for all LLMs and prompt types.

GPT-3.5 Turbo

	None	**Low**	**Med**	**High**
BASELINE	31	17	2	0
CORPUS-ONLY	18	23	8	1
GRAMMAR-ONLY	20	27	2	1
MORPHS-ONLY	3	22	16	9
CORPUS-GRAMMAR	18	23	9	0
CORPUS-MORPHS	2	28	12	8
GRAMMAR-MORPHS	3	29	13	5
CORPUS-GRAMMAR-MORPHS	2	27	12	9

Table 10: Human quality annotation of GPT-3.5 outputs with automated retrieval (raw counts out of 50) by prompt type.

GPT-4o

	None	**Low**	**Med**	**High**
BASELINE	0	10	20	20
CORPUS-ONLY	1	16	13	20
GRAMMAR-ONLY	0	17	16	17
MORPHS-ONLY	0	13	18	19
CORPUS-GRAMMAR	0	14	17	19
CORPUS-MORPHS	0	10	17	23
GRAMMAR-MORPHS	0	19	14	17
CORPUS-GRAMMAR-MORPHS	0	9	20	21

Table 11: Human quality annotation of GPT-4o outputs with automated retrieval (raw counts out of 50) by prompt type.

G Full error counts

This section contains the full counts of annotated errors by category and prompt type.

GPT-3.5 Turbo

	BASE	C	G	M	C+G	C+M	G+M	C+G+M	TOTAL
None	0	1	1	6	0	8	3	5	24
Addition	0	5	3	14	1	9	10	11	53
Omission	3	9	2	13	2	5	9	9	52
Substitution - Subject	0	3	0	7	0	9	9	12	40
Substitution - TAM	0	11	3	17	6	19	19	19	94
Substitution - Other	4	9	4	13	6	16	14	13	79
Overtranslation	1	1	1	4	0	2	3	2	14
Undertranslation	0	0	0	2	1	2	2	2	9
Target Error - Grammar	0	1	1	4	2	3	3	1	15
Target Error - Coherence	0	0	3	5	2	3	7	7	27
Target Error - Style/Register	0	3	0	5	2	3	1	2	16
Complete Mistranslation	30	19	21	2	18	2	2	2	96
Mistranslation - Lexical Correspondence	15	13	23	9	21	11	15	10	117
Refusal	1	0	0	0	0	0	0	0	1
Total	54	75	62	101	61	92	97	95	637

Table 12: Human error type annotation of GPT-3.5 outputs with automated retrieval (raw counts, up to 3 errors per sentence) by prompt type.

GPT-4o

	BASE	C	G	M	C+G	C+M	G+M	C+G+M	TOTAL
None	15	16	10	16	13	19	14	18	121
Addition	2	5	7	5	4	1	6	4	34
Omission	8	7	6	7	6	3	5	5	47
Substitution - Subject	1	2	0	1	2	1	2	2	11
Substitution - Other	22	24	22	18	19	18	17	20	160
Substitution - TAM	16	17	19	12	13	10	11	9	107
Overtranslation	2	1	0	2	2	2	1	2	12
Undertranslation	6	1	3	1	3	0	1	2	17
Target Error - Grammar	1	3	4	4	1	2	6	1	22
Target Error - Coherence	1	3	4	5	4	5	9	5	36
Target Error - Style/Register	1	2	3	4	4	2	4	3	23
Complete Mistranslation	0	1	0	0	0	0	0	0	1
Mistranslation - Lexical Correspondence	4	3	5	6	6	6	9	4	43
Total	79	85	83	81	77	69	85	75	634

Table 13: Human error type annotation of GPT-4o outputs with automated retrieval (raw counts, up to 3 errors per sentence) by prompt type.