On the Robustness of Morphosyntactic Transformation with Large Language Models: The Case of Quechua Collao

Pool Pocco
Chana Research Group
Pontificia Universidad Católica del Perú
Lima, Peru
pool.pocco@pucp.edu.pe

Arturo Oncevay
Chana Research Group
Pontificia Universidad Católica del Perú
Lima, Peru
arturo.oncevay@pucp.edu.pe

Abstract

Morphosyntactic transformation poses significant challenges for large language models (LLMs) in low-resource, morphologically rich languages, where multiple grammatical categories are often encoded within a single word. Applying controlled grammatical changes while preserving meaning requires both linguistic precision and robust generalization. We introduce a morphosyntactically controlled transformation dataset for Quechua Collao, built from a normalized Spanish–Quechua parallel corpus with explicit annotations and transformation labels. The dataset defines a controlled sentence-level transformation task, where models generate target sentences given a source sentence and a structured specification of grammatical changes. We evaluate multiple LLMs under varying numbers of in-context examples, selection strategies, and the use of lightweight morphological hints. Results show that performance depends strongly on prompt design and task formulation rather than model size alone. Increasing the number of examples yields model-dependent gains, benefiting smaller models more, while larger models remain relatively stable. Morphological hints provide selective improvements depending on the transformation type. These findings show that robustness arises from the interaction between model behavior, context size, and linguistic structure, highlighting the importance of controlled experimental design in low-resource settings.

1 Introduction

Morphosyntactic transformation is a structured generation task in which a system modifies specific grammatical properties of a source sentence while preserving its propositional content. Unlike word-level inflection, sentence-level transformation may require coordinated changes across multiple tokens or the insertion and deletion of functional elements. This challenge becomes particularly pronounced in morphologically rich languages, where several grammatical features are encoded within a single word through the concatenation of affixes, and even minimal modifications can entail non-trivial structural adjustments.

Southern Quechua (ISO 639-3: quz), particularly its Collao variety, provides a suitable testbed for this task. As part of the Quechuan language family, it exhibits a predominantly agglutinative morphology in which grammatical categories such as person, tense, aspect, number, and evidentiality are expressed through suffixation (Cerrón-Palomino, 2003; Chuquimamani Valer et al., 2021). This results in dense morphological structure and productive inflectional paradigms, where accurate transformation requires identifying and modifying specific morphemes while preserving the remaining structure.

In addition, Quechua Collao is situated in a low-resource setting, characterized by limited availability of digitized corpora, annotated data, and computational tools. Joshi et al. (2020) formalize this condition through a taxonomy of language resource levels, showing that most of the world's languages fall into categories with minimal computational support. Under these constraints, approaches based on large pretrained language models and in-context learning have become particularly attractive.

Recent work has explored sentence-level morphosyntactic transformation in low-resource languages, notably through shared tasks in the AmericasNLP workshop (De Gibert et al., 2025; Chiruzzo et al., 2024). In these settings, systems are required to apply explicit grammatical changes—such as person, polarity, or tense—to a source sentence and generate the corresponding modified output. Within these tasks, LLMs prompted with few-shot examples have emerged as competitive approaches, with prior studies showing that both prompt enrichment with morphosyntactic information (Vasselli et al., 2024) and example selection strategies

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 128–146

(Lupicki et al., 2025) can influence performance.

However, these efforts do not specifically address Quechua Collao, nor do they provide datasets with explicit morphosyntactic control tailored to its linguistic properties. While prior work on Quechua has primarily focused on tasks such as machine translation and related applications (Cueva Medina et al., 2024), the availability of standardized, annotated resources for controlled morphosyntactic transformation remains limited. This gap is particularly relevant for morphologically rich and low-resource languages, where the absence of linguistically structured datasets constrains both model evaluation and the systematic study of generation behavior (Camacho Caballero and Zevallos Salazar, 2020; Joshi et al., 2020).

Despite these advances, the relative contribution of different design choices remains insufficiently understood. Existing work typically reports improvements from individual techniques, but does not systematically analyze how factors such as the number of in-context examples, example selection strategies, and the inclusion of lightweight morphological hints interact under controlled conditions, particularly in relation to differences across models and transformation complexity.

In this work, we construct a morphosyntactically controlled transformation dataset for Quechua Collao, derived from a normalized Spanish–Quechua parallel corpus and enriched with explicit morphosyntactic annotations and transformation labels. This resource enables a controlled formulation of sentence-level morphosyntactic transformation, where models generate target sentences given a source sentence and a structured specification of grammatical changes.

Using this framework, we conduct a systematic study of in-context learning with multiple LLMs, varying the number of examples, example selection strategies, and the use of lightweight morphological hints. In particular, we design a linguistically motivated k-shot selection strategy that prioritizes morphosyntactic relevance while constraining spurious variation in the prompt. This setup allows us to analyze how design choices in prompt construction affect the robustness and consistency of morphosyntactic generation in a low-resource setting.

Source	Sipasqa chukuta achalan *The young woman adorns the hat*
Change	{NUMBER:PL}
Target	Sipasqa chukukunata achalan *The young woman adorns the hats*

Table 1: Example of a morphosyntactic transformation instance.

Metric	**Spanish**	**Quechua**
Sentence pairs	22,581	22,581
Total tokens	449,580	285,605
Unique tokens	52,632	73,637
Avg. tokens per sentence	19.91	12.65

Table 2: Statistics of the Spanish–Quechua parallel corpus.

2 Task and Dataset

We consider the task of sentence-level morphosyntactic transformation in Quechua Collao. Given a source sentence and a set of morphosyntactic changes, the task consists of generating a target sentence that reflects the specified transformation while preserving its propositional meaning. Table 1 illustrates the task format.

2.1 Corpus Construction

The dataset is derived from a Spanish–Quechua parallel corpus compiled from bilingual educational materials and publicly available texts. The Spanish side provides a semantic anchor and supports the initial annotation stage, while the transformation task itself is defined over Quechua sentences. The corpus was processed through text extraction, normalization[1], sentence segmentation, and heuristic alignment to obtain parallel sentence pairs. This parallel corpus serves as the basis for subsequent morphosyntactic annotation and transformation generation.

Table 2 summarizes the resulting corpus. Although the Quechua side contains fewer tokens overall, it exhibits a higher number of unique word forms and shorter average sentence length. This pattern reflects the agglutinative nature of Quechua morphology: grammatical information that often requires multiple words in Spanish can be encoded through productive suffixation attached to a single lexical root in Quechua.

[1]Orthographic normalization follows the official conventions for Southern Quechua described in Chuquimamani Valer et al. (2021).

Category	Example markers
PERSON	-ni, -nki, -n, -nchik
TENSE	-rqa, -sqa, -saq
ASPECT	-chka
NUMBER	-kuna
POSSESSION	-y, -n, -yki
EVIDENTIALITY	-mi, -si
POLARITY	mana ... -chu

Table 3: Main morphosyntactic categories and typical surface realizations in Quechua Collao.

2.2 Morphosyntactic Annotation

To enable controlled transformations, Quechua sentences were annotated with morphosyntactic features such as person, tense, aspect, number, possession, evidentiality, and polarity. Annotation is performed using a semi-automatic process based on surface morphological patterns, followed by manual verification by native speakers to ensure linguistic consistency.

Table 3 summarizes the main morphosyntactic categories represented in the dataset and their typical morphological realizations in Quechua Collao. These categories correspond to core grammatical distinctions in the language and follow the descriptive framework proposed in previous linguistic work (Chuquimamani Valer et al., 2021). A full inventory of labels and abbreviations is provided in Appendix C.

2.3 Transformation Dataset

From the annotated corpus, we construct a dataset of transformation instances, each defined as a tuple (*source sentence*, *change*, *target sentence*) (De Gibert et al., 2025). This dataset enables the systematic study of controlled morphosyntactic transformations in Quechua Collao, a setting for which no prior structured resources are available. Target sentences are obtained through controlled modification of morphosyntactic features and manually verified to preserve grammaticality and the original propositional content. Further details on the transformation pipeline are provided in Appendix B.

Table 4 provides an overview of this task-ready resource.

In addition to the minimal representation, some experimental configurations incorporate lightweight morphological hints that specify the expected surface realization of target categories. These hints are derived from the same morphosyntactic markers and are used in enriched prompting setups to provide explicit linguistic guidance during inference.

Metric	Value
Transformation instances	6328
Unique source sentences	471
Unique target sentences	474
Unique change specifications	402
Avg. transformations per source	13.44

Table 4: Statistics of the morphosyntactic transformation dataset used for in-context learning experiments.

Label	Expected realization
NUMBER:PL	noun + -kuna
POSS:3_SI	noun + -n
TENSE:PST_NEXP	verb + -sqa
TYPE:NEG	mana ... -chu
MODE:POT	verb + -man

Table 5: Examples of lightweight morphological hints used in prompting.

Table 6 presents the distribution of morphosyntactic categories in the transformation field. Person, possession, and tense transformations account for the majority of cases, reflecting their central role in Quechua verbal and nominal morphology and their high productivity in natural language usage.

The resulting dataset serves as the basis for the in-context learning experiments described in the following section.

3 Experimental Setup

3.1 In-Context Learning Setup

All experiments are conducted in an in-context learning setting using instruction-tuned large language models. Each prompt is formulated as a structured table of resolved examples followed by a final query with an empty target field, so that the model must infer the required transformation by analogy.[2] We evaluate three instruction-tuned LLMs selected to contrast parameter scale and generation behavior: Mistral 24B Instruct (Jiang et al., 2023), Qwen 14B Instruct (Bai et al., 2023), and the reasoning-oriented OpenAI GPT-OSS-20B (OpenAI et al., 2025).

To ensure comparability across runs, decoding is deterministic in all configurations, using temperature $= 0$ and top-$k = 1$. The maximum output length is fixed to 64 tokens for Mistral and Qwen, while GPT-OSS-20B is generated without

[2]A full example of the prompting format is provided in Appendix A.

Category	Frequency	Percentage
PERSON	5676	41.9%
POSS	2685	19.8%
TENSE	2385	17.6%
NUMBER	1135	8.4%
EVID	519	3.8%
MODE	397	2.9%
ASPECT	387	2.9%
TYPE	208	1.5%
PERSON_OBJ	132	1.0%
SUBTYPE	30	0.2%

Table 6: Distribution of morphosyntactic categories in the transformation field (*Change*).

Factor	Values
Dataset features	Source, Change, Target, Hints
Models	Qwen2.5-14B-Instruct Mistral-24B-Instruct GPT-OSS-20B
Shots (K)	{5, 10, 15}
Transformation size	{1, 2}
Prompting modes	w/o hints; w/ hints
Decoding	temperature = 0; top-k = 1 max tokens = 64 (Mistral/Qwen) no explicit cap (GPT-OSS)
Inference setup	LM Studio (local API)

Table 7: Experimental setup.

an explicit cap.[3] All models are executed locally through LM Studio, which also allows us to record inference latency under the same execution environment. Additional inference details are provided in Appendix E.

3.2 k-shot Selection Strategy

The quality of the in-context examples is a central factor in this task. Rather than sampling examples at random, we use a hierarchical retrieval strategy designed to preserve morphosyntactic relevance while minimizing distracting variation.

Selection proceeds in three stages:

- **Exact match.** The system first retrieves examples whose `Change` specification exactly matches the target instance, requiring identical morphosyntactic categories and values (e.g., `PERSON:1_PL_INC`, `TENSE:PST_EXP`).

- **Primary-category match.** If the required number of examples is not reached, the system backs off to instances sharing the same primary morphosyntactic category, defined as the dominant category within the `Change` specification (e.g., PERSON, TENSE, POSS), regardless of the specific feature values.

- **Similarity-based fallback.** Remaining slots are filled using TF-IDF similarity over the source sentence in order to preserve lexical and structural proximity.

[3]The 64-token limit exceeds the expected length of a single transformed sentence and prevents unnecessarily long generations for Mistral and Qwen. All models are prompted to output a `FINAL:` marker for retrieving the generated target; GPT-OSS-20B is left uncapped because it may produce intermediate reasoning before reaching that marker. This may affect latency comparability, but target retrieval remains uniform across models.

This retrieval process is further constrained in three ways. First, we apply a *subset constraint*, which prevents examples from introducing morphosyntactic categories that are absent from the target transformation. Second, we follow a *cluster-first* policy, which prioritizes examples derived from the same transformation cluster whenever possible. Third, evaluation follows a *leave-one-out* scheme, so that the target instance never appears among its own demonstrations. The full selection workflow is illustrated in Appendix D.

3.3 Experimental Factors

The experiments vary four controlled factors: model, number of in-context examples, use of morphological hints, and transformation complexity. Table 7 summarizes these settings.

We use the full transformation pool for retrieval and evaluation, rather than fixed train/dev/test partitions, since the goal is not to fine-tune model parameters but to measure controlled generalization by analogy. The number of demonstrations $K \in \{5, 10, 15\}$ allows us to test whether additional context improves performance. We also compare prompts with and without lightweight morphological hints, which provide surface-oriented cues derived from the target transformation. Finally, we distinguish between single-category transformations (*size* = 1) and compositional transformations involving two categories (*size* = 2), allowing us to evaluate how model behavior changes with structural complexity.

3.4 Evaluation Metrics

We evaluate model outputs using automatic metrics that capture both exact correctness and surface-level similarity.

Exact match accuracy measures the proportion of predictions that exactly match the reference target sentence after normalization.

chrF computes character n-gram F-scores between predicted and reference sentences, providing a more fine-grained measure of similarity that is robust to minor variations in morphologically rich languages.

In addition to automatic metrics, we perform a human-in-the-loop evaluation to assess grammatical correctness, naturalness, and semantic consistency of the generated outputs. Details on the evaluation protocol and results are provided in Appendix F.

4 Results

Model performance varies substantially across context sizes, revealing distinct sensitivities to the number of in-context examples. Table 8 summarizes corpus-level results across models and values of K.

Mistral 24B achieves strong accuracy with minimal context, indicating that it can perform robustly even with a small number of demonstrations. In contrast, Qwen 14B exhibits consistent gains as K increases, reaching the highest overall performance at larger context sizes. GPT-OSS 20B shows more gradual improvements with additional examples, but remains less competitive in exact-match accuracy.

These patterns indicate that the effect of increasing K is strongly model-dependent. Mistral remains comparatively stable across configurations, suggesting limited reliance on additional contextual support. Qwen, by contrast, benefits directly from larger context, indicating a stronger dependence on in-context signals for generalizing morphosyntactic transformations. GPT-OSS follows an intermediate trend, with moderate but less consistent gains.

Overall, additional in-context examples do not provide uniform benefits across models. Their effectiveness depends on how each model leverages contextual information during inference, motivating the more detailed analyses that follow.

4.1 Model Comparison and Stability

Robustness vs. sensitivity. Model performance reveals a clear trade-off between robustness and sensitivity to experimental conditions. While all systems achieve comparable average performance, their behavior across configurations differs substantially.

	Mistral 24B		GPT-OSS 20B		Qwen 14B	
K	**Acc.**	**chrF**	**Acc.**	**chrF**	**Acc.**	**chrF**
5	52.54	90.69	44.07	91.21	44.92	92.98
10	**55.08**	**90.28**	48.73	92.63	54.66	94.40
15	53.39	89.83	**53.81**	**93.95**	**58.90**	**94.14**

Table 8: Corpus-level results by model and number of in-context examples (K). Bold indicates the best-performing configuration for each model across values of K.

Table 9 summarizes performance variability across context sizes using mean and standard deviation as indicators of stability. Lower variance reflects more consistent behavior across configurations, while higher variance indicates sensitivity to changes in prompting conditions.

Mistral exhibits consistently low variance, indicating stable performance that is largely invariant to both the number of in-context examples and the use of hints. This suggests that the model internalizes morphosyntactic transformations without relying heavily on prompt-specific cues.

In contrast, Qwen displays markedly higher variability, reflecting strong sensitivity to contextual factors. Its performance improves as additional examples are provided, indicating a more direct reliance on in-context signals for generalizing transformation patterns, particularly in settings involving higher structural complexity.

GPT-OSS occupies an intermediate position. Although it benefits from increased context, its improvements are less consistent and do not follow a stable trend, suggesting a limited ability to systematically exploit additional examples compared to Qwen, while lacking the robustness observed in Mistral.

Accuracy vs. surface similarity. A complementary distinction emerges between exact-match correctness and surface similarity. Despite achieving high chrF scores, Qwen and GPT-OSS do not consistently match this performance in exact accuracy, indicating that morphologically similar outputs may still diverge from the correct transformation. This highlights the importance of jointly evaluating both metrics in morphologically rich languages.

Overall, these results indicate that model behavior in this task is systematic rather than random. Mistral emerges as the most robust model, Qwen

Model	Acc.	chrF
Mistral 24B	**53.67**± 1.06	90.27± 0.35
Qwen 14B	52.82± 5.85	**93.84**± 0.61
GPT-OSS 20B	48.87± 3.98	92.60± 1.12

Table 9: Performance stability across models (averaged over $K = 5, 10, 15$).
Values are reported as mean $\pm$ standard deviation over $n = 708$ instances per model.

as the most context-sensitive, and GPT-OSS as an intermediate system with moderate adaptability. These differences motivate the more fine-grained analyses of contextual effects and prompting strategies presented in the following subsections.

4.2 Effect of Context Size

The effect of the number of in-context examples (K) is not uniform across models and is better captured through statistical analysis than raw metric differences. To assess whether changes in K produce systematic variation, we apply Kruskal–Wallis tests for global comparisons across context sizes and Mann–Whitney U tests for pairwise contrasts.[4]

For Qwen, the effect of K is statistically significant across configurations, confirming a strong dependence on contextual information. Performance improves consistently as additional examples are provided, indicating that the model relies on in-context signals to infer transformation rules, particularly in settings with lower frequency or higher structural complexity. In this regime, additional examples act as explicit guidance that supports generalization by analogy.

In contrast, Mistral does not exhibit statistically significant differences across values of K, indicating that its performance remains largely invariant to the number of demonstrations. This reinforces the robustness observed in the previous subsection: the model appears to internalize morphosyntactic regularities in a way that does not depend strongly on the quantity of contextual support at inference time.

GPT-OSS shows a weaker and less consistent effect. Although performance tends to improve with larger K, the differences are comparatively mild and not systematically significant, suggesting that the model can partially benefit from additional context without exploiting it as consistently as Qwen.

[4]All statistical tests are conducted using a significance level of $\alpha = 0.05$.

Model	Metric	Test	p-value	Sig.
Mistral 24B	Accuracy	Kruskal (K)	0.853	n.s.
	chrF	Kruskal (K)	0.915	n.s.
GPT-OSS 20B	Accuracy	MW (5 vs 15)	0.034	*
	chrF	MW (5 vs 15)	0.033	*
Qwen 14B	Accuracy	Kruskal (K)	0.0077	**
	chrF	Kruskal (K)	0.0042	**
	Accuracy	MW (5 vs 15)	0.0024	**
	chrF	MW (5 vs 15)	0.0012	**

Table 10: Effect of the number of in-context examples (K) on model performance.
Significance levels: n.s. ($p \geq 0.05$, no consistent effect), * ($p < 0.05$, mild effect), ** ($p < 0.01$, strong effect).

Overall, these findings indicate that increasing K is not universally beneficial. Its effectiveness depends on the model's inference strategy: context-sensitive models derive substantial gains from additional examples, while more robust models exhibit diminishing returns.

4.3 Effect of Morphological Hints

We evaluate the impact of morphological hints through paired comparisons between configurations with and without hints, controlling for model, context size, and input instance. To determine whether observed differences are systematic, we apply the Wilcoxon signed-rank test. Table 11 summarizes the results.

For Mistral, the effect is statistically significant across metrics, indicating that hints consistently improve performance. Rather than compensating for missing knowledge, hints appear to reinforce correct morphological decisions by providing surface-level constraints aligned with the model's internal representations. This results in more stable and accurate outputs, particularly in suffix selection.

Qwen exhibits a more selective effect. While hints improve accuracy, their impact on chrF is not consistent, suggesting that their primary contribution lies in guiding the correct application of the requested transformation rather than improving surface similarity. This indicates that Qwen benefits from explicit cues, but does not fully integrate them across all aspects of generation.

In contrast, GPT-OSS does not show statistically significant differences between configurations. This suggests that the model does not consistently leverage hints, and that its predictions are influenced by factors not directly modulated by the additional information provided.

Overall, these findings indicate that morpholog-

Model	Metric	p-value	Sig.
Mistral 24B	Accuracy	$< 1 \times 10^{-6}$	***
	chrF	$< 1 \times 10^{-6}$	***
Qwen 14B	Accuracy	5.55×10^{-3}	**
	chrF	5.21×10^{-1}	n.s.
GPT-OSS 20B	Accuracy	4.54×10^{-1}	n.s.
	chrF	1.66×10^{-1}	n.s.

Table 11: Effect of morphological hints evaluated with the Wilcoxon signed-rank test (paired comparisons between configurations with and without hints).
Significance levels: n.s. ($p \geq 0.05$, no consistent effect), * ($p < 0.05$, mild effect), ** ($p < 0.01$, moderate improvement), *** ($p < 0.001$, strong improvement).

ical hints act as auxiliary signals whose effectiveness depends on the model's inference strategy. They function as a stabilizing mechanism for models with stronger internal representations, and as partial guidance for more context-sensitive systems, but do not provide universal improvements.

Qualitative examples illustrating these effects are provided in Appendix G, where hints guide correct suffix selection in some cases while having no effect in others.

4.4 Efficiency: Performance vs. Latency

Increasing the number of in-context examples (K) improves performance for some models, but also increases the amount of information processed at inference time, leading to longer end-to-end response times. This trade-off is particularly relevant in low-resource settings, where larger prompts may be required to achieve competitive performance. Latency trends across models and values of K are reported in Appendix H.

Differences across models reflect varying degrees of dependence on contextual scaling. Qwen benefits the most from increasing K, but requires longer prompts to reach strong performance. In contrast, Mistral remains comparatively stable with fewer examples, reducing its reliance on extensive context. GPT-OSS exhibits less consistent behavior, with weaker and less predictable gains from additional examples. Overall, these results suggest that improvements obtained through larger context windows should be considered jointly with their computational cost, particularly when resources are constrained.

4.5 Performance by Morphological Category

To better understand the sources of variability observed in previous sections, we analyze performance across morphosyntactic categories. While aggregate metrics suggest relatively stable behavior, Table 12 reveals that performance is not homogeneous across transformation types, but instead reflects systematic differences in linguistic complexity and data distribution. Category-wise trends across models are visualized in Appendix I.

Category	n	Model	Acc.	chrF
ASPECT	90	Mistral	52.22	88.33
	90	OSS	46.67	93.33
	90	Qwen	58.89	95.24
NUMBER	51	Mistral	32.35	84.09
	51	OSS	43.14	90.86
	51	Qwen	23.53	87.49
PERSON	120	Mistral	56.67	90.37
	120	OSS	40.83	91.87
	120	Qwen	54.17	93.77
POSS	120	Mistral	65.83	88.67
	120	OSS	65.83	94.39
	120	Qwen	75.83	95.74
TYPE	84	Mistral	47.62	92.96
	84	OSS	47.62	91.01
	84	Qwen	41.67	92.73

Table 12: Performance by morphosyntactic category and model.

Categories involving regular and localized morphological alternations—such as *PERSON* and *POSS*—are handled consistently well across models. In particular, *POSS* achieves the highest overall accuracy, while *PERSON* maintains strong performance across systems. These categories benefit from predictable suffixal patterns and higher frequency in the dataset, which generally correlates with improved performance by increasing the likelihood of retrieving relevant k-shot examples during inference.

In contrast, *NUMBER* emerges as the most challenging category across all models. Performance drops substantially, especially for Qwen, highlighting a notable exception to this trend: despite its relatively high frequency, NUMBER remains difficult due to structural ambiguity. Plural marking in Quechua can apply to multiple candidate nouns within a sentence, requiring the model to correctly identify the scope of the transformation. This introduces ambiguity that is absent in more localized morphological changes and increases sensitivity to example selection.

A similar, though less pronounced, pattern is observed for *TYPE*. These transformations often involve coordinating multiple elements (e.g., *mana* +

verb + *-chu*), making them partially syntactic rather than strictly morphological. As a result, models may preserve surface similarity while failing to fully reconstruct the intended grammatical structure, leading to moderate accuracy despite high chrF scores.

Differences across models further highlight the interaction between model-specific behavior and category-specific difficulty. Qwen achieves the highest scores in categories such as *ASPECT* and *POSS*, suggesting effective pattern extraction when sufficient contextual evidence is available, but degrades sharply in *NUMBER*, indicating higher sensitivity to ambiguity. Mistral, in contrast, exhibits more balanced performance across categories, maintaining consistently high chrF scores and moderate accuracy even in more challenging cases. GPT-OSS shows intermediate behavior, with relatively strong performance in *NUMBER* but less consistent results overall.

Taken together, these findings indicate that robustness in this task is not determined solely by model size or prompting strategy, but also by the intrinsic properties of each morphosyntactic transformation. Categories with regular structure and higher frequency yield more stable predictions, whereas structurally complex or underrepresented transformations introduce variability that cannot be fully mitigated by increasing k-shot examples alone.

5 Discussion

Across models, the most reliable behavior is not associated with the highest scores, but with consistency across configurations. Mistral 24B exhibits low variance and stable performance regardless of prompt conditions, suggesting that it internalizes morphosyntactic transformations in a way that reduces dependence on external cues. In contrast, Qwen 14B achieves competitive results but shows greater sensitivity to changes in context size and prompting configuration. This indicates that robustness is better characterized by reproducibility than by isolated metric gains.

The interaction between model size and the number of in-context examples reveals two distinct strategies. Larger models, such as Mistral, achieve strong performance with minimal context, whereas smaller models, such as Qwen, rely on increasing K to reach comparable results. However, increasing the number of examples shifts part of the modeling burden from the model itself to data availability. In low-resource settings, where curated examples are scarce and costly to obtain, this introduces an additional constraint: improvements obtained through larger context windows may not scale in practice due to limitations in data collection, annotation, and curation. As a result, robustness depends not only on performance, but on the ability to maintain it under constrained data and inference conditions.

Variation across morphosyntactic categories shows that performance is not determined solely by model architecture or prompting strategy, but also by the intrinsic properties of each transformation. Categories with regular and localized morphology, such as *PERSON* and *POSS*, are consistently easier to model, while structurally complex or ambiguous transformations, such as *NUMBER* or multi-element constructions in *TYPE*, remain challenging even under favorable prompting conditions. This suggests that certain forms of linguistic complexity introduce variability that cannot be fully mitigated through additional context or prompt engineering alone.

6 Conclusion

Our results show that performance in morphosyntactic transformation is shaped not only by model size, but by the interaction between model behavior, context size, and linguistic complexity. Larger models such as Mistral achieve stable performance with minimal contextual support, while smaller models such as Qwen benefit more from increasing the number of in-context examples. Analysis across morphosyntactic categories further reveals that structurally complex transformations remain challenging regardless of prompting strategy.

These findings highlight that robustness in low-resource settings depends on the alignment between linguistic structure and prompt design, rather than model size alone. Future work could extend morphosyntactic coverage, further analyze category-level imbalance, and explore alternative modeling approaches, including hybrid systems that integrate explicit linguistic knowledge with in-context learning.

7 Limitations

Dataset scope and coverage. The proposed dataset, while enabling controlled morphosyntactic transformations, is limited in size and coverage. Al-

though it contains 6,328 transformation instances derived from 471 source sentences, it does not exhaustively represent the full range of morphosyntactic phenomena in Quechua Collao. In particular, the distribution of transformation types is inherently imbalanced, reflecting both the availability of source data and the selective focus on specific grammatical categories.

Linguistic coverage. The dataset focuses on a subset of morphosyntactic categories and does not fully capture the richness of Quechua Collao morphology. While it includes core features such as person, tense, aspect, number, and polarity, other relevant phenomena—such as case marking, subordination, and more complex derivational processes—are not systematically represented. As a result, the task formulation reflects a simplified view of the language, which may limit the generalization of the findings to more complex or less structured linguistic contexts.

Modeling and experimental setup. Our experiments are restricted to a small set of instruction-tuned language models evaluated under a specific in-context learning setup. We do not explore alternative approaches such as fine-tuning, hybrid systems combining symbolic rules with neural models, or retrieval-augmented methods. As a result, the findings are specific to the considered models and prompting strategies, and future work could examine whether similar patterns hold under different modeling paradigms or training regimes.

Evaluation and human validation. The evaluation primarily relies on automatic metrics such as exact match accuracy and chrF, which may not fully capture linguistic acceptability or naturalness in morphologically rich languages. Although we complement this analysis with a human-in-the-loop validation involving native speakers and expert evaluators, this component is limited in scale and based on a relatively small sample of generated instances. While the results show consistent and high ratings across evaluators, the limited number of participants and examples restricts the generalizability of these findings.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Luis Camacho Caballero and Rodolfo Zevallos Salazar. 2020. Lingüística computacional para la revitalización y el poliglotismo. *Revista Letras UNMSM*, 91(134):184–198.

Rodolfo Cerrón-Palomino. 2003. *Lingüística quechua*, 2 edition. Centro de Estudios Regionales Andinos Bartolomé de Las Casas, Cusco.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Nonato Rufino Chuquimamani Valer, Oscar Chávez Gonzales, Felix Alain Riveros Paravicino, César Jara Luna, Moisés Cárdenas Guzmán, and Melquíades Quintasi Mamani. 2021. *Urin qichwa qillqay yachana mayt'u = Manual de escritura quechua sureño*. Ministerio de Educación, Lima.

Beatrice Cueva Medina, Gabriel Fabrizio Tuco Casquino, and José Alfredo Sulla-Torres. 2024. Development of a neural machine translation model optimized with bert for translation from quechua to spanish. In *Proceedings of the 22nd LACCEI International Multi-Conference for Engineering, Education, and Technology*, pages 1–7, San Jose, Costa Rica. LACCEI.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and

fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, and David Yarowsky. 2025. JHU's submission to the AmericasNLP 2025 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 105–111, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

A Prompt Example

We provide an example of the prompt format used for in-context learning. The prompt consists of a table of resolved examples followed by a final instance with an empty target field, which the model must complete.

Each row specifies a source sentence, a set of morphosyntactic changes, and the corresponding transformed target. The final instance follows the same structure but omits the target, requiring the model to infer the correct transformation based on the preceding examples.

```
Source                     | Change            | Target
---------------------------------------------------------------------------
Sipasqa chukuta achalan    | {NUMBER:PL}       | Sipasqa chukukunata achalan
Qayna wayk'urqani wasiypi  | {PERSON:1_PL_INC} | Qayna wayk'urqanchik wasiypi
...
---------------------------------------------------------------------------
<FINAL INSTANCE>
Source                      | Change            | Target
Sipasqa chukuta achalan     | {NUMBER:PL}       |
```

The model is instructed to produce a single output line in the format:

```
FINAL: <target_sentence>
```

This structured format constrains generation by explicitly aligning input transformations with output examples, enabling controlled application of morphosyntactic changes.

B Dataset Construction

B.1 Overview of the Construction Pipeline

The dataset used in this work is constructed through a multi-stage pipeline that transforms a raw parallel corpus into a structured resource for controlled morphosyntactic transformation. The process consists of three main stages: (i) parallel corpus construction, where bilingual sentence pairs are extracted, cleaned, and aligned; (ii) morphosyntactic annotation, where linguistic features are automatically identified and manually validated; and (iii) transformation dataset construction, where annotated sentence pairs are converted into controlled transformation instances. This pipeline enables the generation of high-quality training and evaluation examples in which morphosyntactic changes are explicitly specified and systematically applied, while preserving the original meaning of the source sentence. Figure 1 provides an overview of this process.

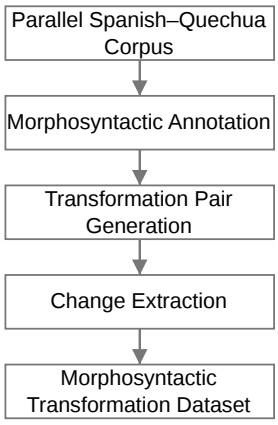

Figure 1: Pipeline for generating the morphosyntactic transformation dataset.

B.2 Parallel Corpus Construction

The parallel corpus is compiled from a combination of heterogeneous bilingual sources, including religious publications and official educational materials. These sources provide complementary properties: while religious texts contribute large-scale and diverse sentence pairs, educational materials offer more consistent orthographic and grammatical usage.

Corpus construction involves multiple preprocessing steps. First, textual data are extracted either through optical character recognition (OCR) for scanned documents or through structured parsing for digital sources. The extracted text is then cleaned to remove formatting artifacts such as line breaks, special characters, and segmentation inconsistencies.

Sentence segmentation is applied to obtain aligned units, followed by heuristic sentence alignment based on length and structural similarity. To ensure alignment quality, a subset of sentence pairs is manually reviewed and corrected.

Given the high degree of orthographic variation in the raw data, especially in non-standardized sources, an additional normalization step is applied using rule-based transformations informed by the official orthographic conventions for Southern Quechua (Chuquimamani Valer et al., 2021), which we adopt here as the reference standard for Quechua Collao (quz). Existing Quechua resources often exhibit inconsistent spelling and may mix surface variants, which introduces noise and affects downstream modeling. In our corpus, this includes forms such as “noqa” or “nuqa”, normalized to “ñuqa”, “qollqe”, normalized to “qullqi”, and “huasi”, normalized to “wasi”. Applying a consistent normalization scheme ensures uniform surface forms across the corpus, improves the reliability of morphosyntactic analysis, and reduces variability unrelated to the linguistic transformations of interest.

B.3 Morphosyntactic Annotation

Morphosyntactic annotation is performed through a semi-automatic pipeline that combines rule-based analysis with syntactic parsing and manual validation. The goal of this stage is to assign structured linguistic features to each sentence, enabling the controlled generation of morphosyntactic transformations.

In the automatic stage, candidate annotations are generated using complementary sources of linguistic evidence. For Quechua, rule-based pattern matching is applied to identify morphological markers such as person, number, tense, and evidentiality. For Spanish, syntactic parsing is used to extract grammatical information that can support disambiguation. These sources are then combined to produce a set of candidate tags along with associated evidence.

Given the ambiguity of certain morphological markers and the potential noise introduced by automatic processing, a manual validation step is performed by native speakers. This step resolves ambiguous cases, adds missing labels, corrects an-

notation errors, and ensures that the assigned tags accurately reflect the intended grammatical structure of each sentence.

For example, the suffix *-n* may mark third-person verbal agreement, as in *rima-n* 'he/she speaks', or third-person nominal possession, as in *wasi-n* 'his/her house'. Such cases are resolved during manual validation by considering the lexical category and syntactic context of the marked word. Similarly, possessive forms may require an epenthetic *-ni* after consonant-final roots, as in *yawar-ni-n* 'his/her blood', rather than the simpler vowel-final pattern *wasi-n*. These checks ensure that the final tags reflect the intended morphosyntactic function rather than only surface string matching.

The final annotated corpus includes, for each sentence, a set of validated morphosyntactic tags that serve as the basis for constructing transformation pairs. Figure 2 illustrates an example of this process, showing both automatically extracted candidates and their manually validated counterparts.

B.4 Transformation Dataset Construction

The transformation dataset is constructed from the annotated corpus by generating pairs of base and transformed sentences under controlled morphosyntactic modifications. Each transformation corresponds to a change in a subset of morphosyntactic features while preserving the underlying propositional meaning of the original sentence.

Controlled transformation generation. Starting from an annotated base sentence, one or more transformed variants are generated by modifying specific grammatical categories such as person, number, tense, or polarity. This process goes beyond surface-level rule application: transformations may require coordinated changes across multiple elements in the sentence to maintain agreement and grammaticality.

This design enables the creation of multiple valid transformation instances from a single base sentence, increasing dataset diversity without introducing uncontrolled noise. **Unlike standard data augmentation approaches**, which rely on perturbations or random variation, transformations in this dataset are linguistically grounded and explicitly defined through morphosyntactic features.

Transformation representation. Each instance is represented as a tuple of the form *(Source, Change, Target)*, where *Change* encodes the difference between the morphosyntactic features of the source and target sentences. This structured representation allows models to infer transformations from explicit grammatical specifications.

Change computation. The set of changes is computed by comparing the annotated features of the source and target sentences, retaining only those categories that differ. This ensures that each transformation is minimally specified and avoids introducing irrelevant information.

Table 13 illustrates how multiple controlled transformations can be generated from a single base sentence.

B.5 Dataset Representation and Change Computation

Each transformation instance is represented as a tuple of the form (Source, Change, Target), where the Change component encodes the morphosyntactic differences between the source and target sentences. The *Source* corresponds to the base sentence, the *Target* to the transformed variant, and the *Change* component encodes the specific morphosyntactic differences between them.

Change computation. The *Change* representation is obtained by comparing the validated morphosyntactic annotations of the source and target sentences. Each set of tags is first normalized into a dictionary of the form *CATEGORY:VALUE*. The change is then defined as the subset of categories whose values differ between the source and target, ensuring that only the relevant grammatical modifications are retained.

For example, given the following pair:

- **Source:** Sipasqa chukuta achalan {PERSON:3_SI, TENSE:PRE_SIM}
- **Target:** Sipaskunaqa chukunkuta achalanku {PERSON:3_PL, NUMBER:PL, POSS:3_PL, TENSE:PRE_SIM}

the resulting change is:

{PERSON:3_PL, NUMBER:PL, POSS:3_PL}

This procedure ensures that transformations are minimally specified and avoids including features that remain unchanged. To further improve consistency and reproducibility, the elements in the *Change* field follow a fixed ordering of categories (e.g., PERSON, NUMBER, POSS, TENSE, ASPECT, MODE, TYPE), which standardizes the representation across all instances.

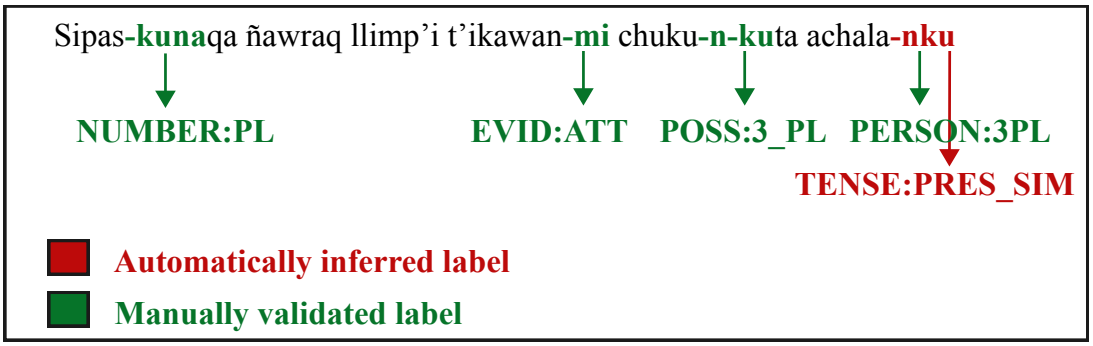

Figure 2: Example of morphosyntactic annotation showing automatically extracted candidates and manually validated tags.

Source	Change	Target
Purikuq runaqa chay wasiman chayarqan	{PERSON:3_PL, NUMBER:PL}	Purikuq runakunaqa chay wasiman chayarqanku
Purikuq runaqa chay wasiman chayarqan	{TENSE:PST_NEXP}	Purikuq runaqa chay wasiman chayasqa
Purikuq runaqa chay wasiman chayarqan	{TENSE:PRE_SIM, ASPECT:PRG}	Purikuq runaqa chay wasiman chayachkan
Purikuq runaqa chay wasiman chayarqan	{MODE:POT}	Purikuq runaqa chay wasiman chayanman

Table 13: Examples of controlled morphosyntactic transformations derived from a single base sentence.

This structured representation provides a clear interface for in-context learning, allowing models to infer transformations from explicit linguistic specifications rather than implicit patterns.

C Morphosyntactic Label Inventory

The morphosyntactic labels used in the *Change* field were selected to cover productive grammatical categories in Quechua Collao that can be modeled as controlled sentence-level transformations. The inventory includes subject person and number, verbal tense, aspect and mood, polarity, interrogation, nominal possession, nominal pluralization, object-person marking, and evidentiality. Although this inventory does not exhaust the full morphology of the language, it provides a structured set of categories for the controlled transformations studied in this work.

Labels follow a `CATEGORY:VALUE` format. For example, `PERSON:1_PL_INC` denotes a subject-person transformation whose value is first-person plural inclusive, while `TENSE:PST_NEXP` denotes a tense transformation to non-experienced past. When only the category is discussed, we use the category name alone, such as PERSON, TENSE, or NUMBER.

Table 14 summarizes the main labels, their meanings, and typical surface realizations.

Table 15 summarizes the main value abbreviations used in label values.

D k-shot Selection Workflow

Figure 3 details the retrieval workflow used to select in-context examples for each test instance. The procedure first prioritizes exact matches in the `Change` specification, then backs off to examples sharing the same primary morphosyntactic category, and finally fills remaining slots using TF-IDF similarity over the source sentence. This ordering is designed to preserve morphosyntactic relevance before relying on surface similarity.

Two additional constraints make the retrieved context auditable and comparable across runs. First, a subset constraint prevents examples from introducing morphosyntactic categories that are absent from the target transformation. Second, a leave-one-out constraint ensures that the evaluated instance never appears among its own demonstrations. Whenever possible, examples from the same transformation cluster are prioritized to preserve lexical and structural proximity.

E Experimental Conditions

This appendix provides additional implementation details for reproducibility, complementing the experimental setup summarized in the main paper.

Label	Meaning	Typical realization	Example
PERSON	Subject person/number	*-ni, -nki, -n, -nchik, -yku, -nku*	*rima-n* ‘he/she speaks’
TENSE	Verbal tense	*-rqa, -sqa, -saq, -nqa*	*wayk’u-rqa-ni* ‘I cooked’
ASPECT	Verbal aspect	*-chka*	*riku-chka-n* ‘he/she is seeing’
MODE	Verbal mood/modality	*-man, -chus*	*taki-n-man* ‘he/she could sing’
TYPE	Clause type or polarity	*mana ... -chu, ama ... -chu*	*mana yacha-ni-chu* ‘I do not know’
SUBTYPE	Clause subtype	*-chu*	*riku-n-chu?* ‘does he/she see?’
NUMBER	Nominal number	*-kuna*	*waka-kuna* ‘cows’
POSS	Nominal possession	*-y, -yki, -n, -nchik, -yku, -nku*	*wasi-y* ‘my house’
PERSON_OBJ	Subject–object person relation	*-yki, -wa-nki, -su-nki*	*qu-yki* ‘I give you’
EVID	Evidentiality	*-mi/-m, -si/-s*	*amawta-m* ‘the teacher’ (attested)

Table 14: Morphosyntactic labels used in the transformation specifications.

Abbreviation	Meaning
1, 2, 3	First, second, third person
SI	Singular
PL	Plural
INC	Inclusive
EXC	Exclusive
PRE_SIM	Present/simple or non-future simple
PST_EXP	Experienced past
PST_NEXP	Non-experienced past
FUT_SIM	Future/simple
PRG	Progressive aspect
POT	Potential mood
DUB	Dubitative mood
IMP	Imperative
NEG	Negation
PROH	Prohibitive
INT	Interrogative
ATT	Attestative evidential
REP	Reportative evidential

Table 15: Abbreviations used in morphosyntactic label values.

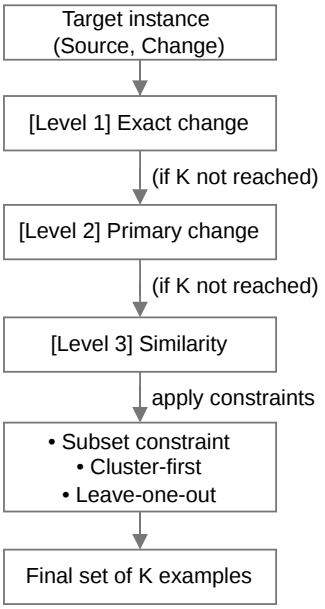

Figure 3: Hierarchical k-shot selection strategy combining exact matching, primary-category retrieval, and similarity-based fallback under structural constraints.

E.1 Local Inference Environment

All experiments were run locally through LM Studio on a Windows machine with an AMD Ryzen 5 3400G CPU, 32 GB of RAM, and AMD Radeon RX Vega 11 integrated graphics reported with 2 GB of adapter memory. No NVIDIA/CUDA device was available in this setup.

E.2 Model Files and Quantization

The evaluated models were loaded in GGUF format through LM Studio. Table 16 reports the exact local model variants used in the final experiments, including quantization format and model size as reported by LM Studio.

E.3 Model Selection Rationale

The model set was selected to provide a compact comparison among instruction-tuned LLMs with different parameter scales and generation profiles. Mistral 24B represents the larger instruction-tuned model in our comparison, Qwen 14B provides a smaller multilingual instruction-tuned alternative, and GPT-OSS 20B introduces a reasoning-oriented generation behavior. The goal was not to exhaustively benchmark all available models, but to analyze how different model profiles respond to the same controlled morphosyntactic transformation setting.

E.4 Decoding and Paired Prompting

All runs used deterministic decoding with temperature $= 0$ and top-$k = 1$, following the reproducibility setting used throughout the pipeline. Mistral and Qwen used a 64-token output cap, which was

Model	Local LM Studio identifier	Quantization	Size
Mistral 24B	`mistral-small-3.2-24b-instruct-2506`	Q4_K_M	15.21 GB
Qwen 14B	`qwen-qwen2.5-14b-instruct-1m`	Q4_K_M	8.99 GB
GPT-OSS 20B	`openai/gpt-oss-20b`	MXFP4	12.11 GB

Table 16: Local model files used in the experiments, as reported by LM Studio.

sufficient for the required single-sentence target format. GPT-OSS 20B was left uncapped only to ensure that the model could reach the required `FINAL:` output line when producing intermediate reasoning.

The experimental grid combines model, number of in-context examples ($K \in \{5, 10, 15\}$), transformation size (1 or 2 morphosyntactic categories), and prompting mode (with or without morphological hints). Paired runs with and without hints used the same selected k-shot examples, so that differences between prompting modes reflect the presence of hints rather than changes in the demonstrations.

F Human Evaluation

F.1 Evaluation Setup

To complement automatic metrics, we conduct a human-in-the-loop evaluation of model outputs. A subset of 44 transformation instances is selected to cover a diverse range of morphosyntactic categories and transformation types, including both single-category and compositional cases.

The evaluated outputs are generated using the best-performing configuration identified in the main experiments (Mistral 24B Instruct). The evaluation is carried out by two native speakers of Quechua Collao and one expert annotator, ensuring both linguistic competence and consistency in judgments.

F.2 Evaluation Protocol

Each generated instance is evaluated along three dimensions:

Grammatical correctness measures whether the output follows the morphosyntactic rules of Quechua Collao and correctly applies the specified transformation.

Naturalness assesses the fluency and acceptability of the generated sentence from a native speaker perspective.

Semantic consistency evaluates whether the output preserves the meaning of the source sentence while accurately reflecting the intended transformation.

Criterion	Average Score
Grammatical correctness	4.87 ($\pm$ 0.60)
Naturalness	4.76 ($\pm$ 0.60)
Semantic consistency	4.82 ($\pm$ 0.60)

Table 17: Human evaluation results (Likert scale from 1 to 5; $\pm$ indicates standard deviation).

Evaluators assign scores using a Likert scale from 1 (very poor) to 5 (excellent) for each dimension. This setup captures both exact-match correctness and perceived fluency, which are particularly important for morphologically rich and low-resource languages.

F.3 Results

Table 17 reports the average scores across evaluators for each evaluation dimension. The results show consistently high ratings, indicating that the generated outputs are generally well-formed, natural, and semantically coherent.

The relatively low and uniform standard deviation indicates stable judgments among evaluators, suggesting that the outputs align well with native speaker intuitions. These findings complement the automatic evaluation results reported in the main text, providing additional evidence of the practical validity of the generated transformations.

G Qualitative Comparison: Hints vs No Hints

Table 18 presents representative examples from Mistral 24B comparing predictions with and without morphological hints. These examples illustrate how hints influence the application of morphosyntactic transformations at the token level.

In particular, hints guide the model toward the correct insertion or modification of morphological markers (e.g., plural suffixes or evidential markers), reducing errors in morpheme placement. However, their effect is not uniform across categories, as some transformations (e.g., negation) are consistently handled correctly even without additional

guidance.

H Latency Analysis

Figure 4 shows response time as a function of the number of in-context examples (K) across models, with and without morphological hints, for both single-change ($size = 1$) and multi-change ($size = 2$) settings.

Latency is measured as the end-to-end response time (in milliseconds) under local inference, computed as the elapsed time between sending a prompt to the model and receiving the generated output. This measurement reflects the combined cost of processing longer prompts and generating responses, rather than an isolated estimate of model-internal inference speed.

Response time increases consistently with K for Mistral and Qwen, reflecting the expected cost of processing longer inputs. In contrast, GPT-OSS exhibits less predictable behavior, with non-monotonic trends across configurations, suggesting that response time is influenced not only by input length but also by model-specific implementation and decoding dynamics.

The effect of morphological hints on response time is comparatively small, indicating that adding structured guidance does not substantially increase computational cost. Overall, these results highlight a trade-off between context size and computational cost, as well as variability in how different models scale with longer inputs.

I Category-wise Performance

This appendix provides additional visualizations of model performance across morphosyntactic categories, disaggregated by the number of in-context examples (K).

Figure 5 shows accuracy by category and model. Performance varies substantially across categories, with transformations such as NUMBER consistently yielding lower accuracy, while categories like POSS and ASPECT exhibit more stable and higher performance. These differences reflect variation in structural complexity and the difficulty of correctly applying specific morphological operations.

The effect of increasing K is also category-dependent. For example, Qwen shows notable improvements in categories such as POSS and ASPECT as more examples are provided, while gains are less consistent for more challenging transformations like NUMBER. This supports the observation that the benefits of additional context depend on both the model and the type of transformation.

Figure 6 presents chrF scores by category. While chrF remains high across most categories, discrepancies between chrF and accuracy are evident, particularly in more complex transformations. This indicates that outputs can be lexically similar to the target while still failing to capture the correct morphological structure.

Together, these plots complement the main results by highlighting how performance varies not only across models and configurations, but also across linguistic categories with differing structural properties.

Change	Source	Target	No hints	With hints
NUMBER:PL	Yana yuraq llimp'iyuq wakataqa allqa waka nini	Yana yuraq llimp'iyuq **wakakunataqa** allqa waka nini	Yana yuraq **llimp'ikuniyuq** wakataqa allqa waka nini	Yana yuraq llimp'iyuq **wakakunataqa** allqa waka nini
TYPE:NEG	Qhipa wiñay warmiqa aqsuta riqsin	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**
EVID:ATT	Inkataqa amawta yana-pasqa	Inkataqa **amawtam** yanapasqa	Inkataqa **amawtami** yanapasqa	Inkataqa **amawtam** yanapasqa

Table 18: Qualitative comparison of predictions from Mistral 24B with and without morphological hints. Bold indicates the modified segment relative to the source.

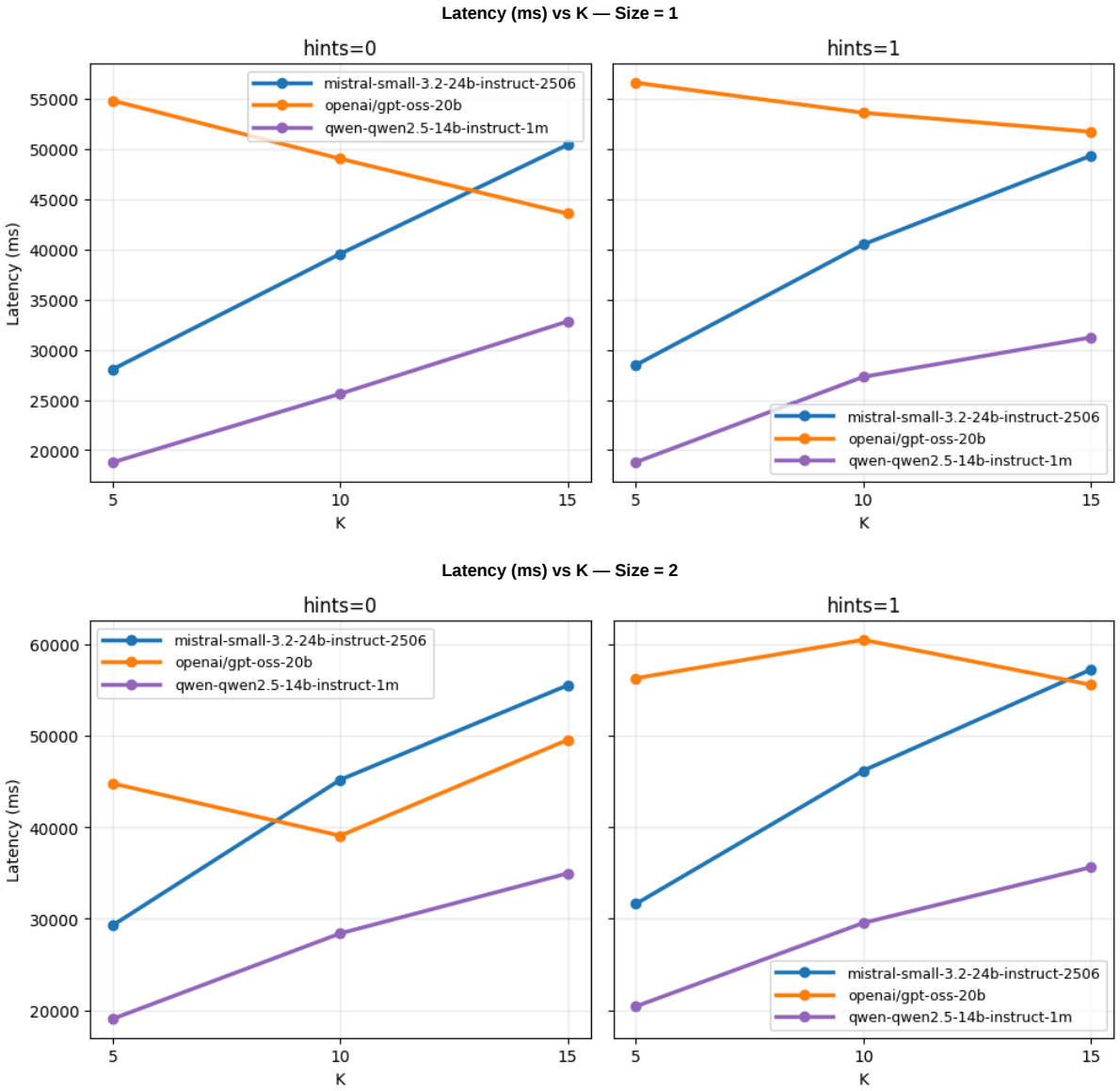

Figure 4: Inference latency as a function of the number of in-context examples (K) across models, with and without morphological hints, for $size = 1$ and $size = 2$.

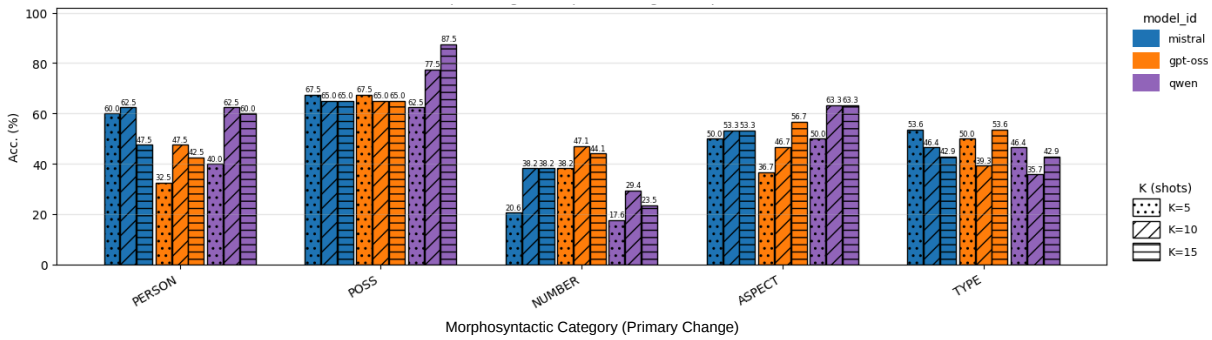

Figure 5: Accuracy by morphosyntactic category and model, disaggregated by the number of in-context examples (K).

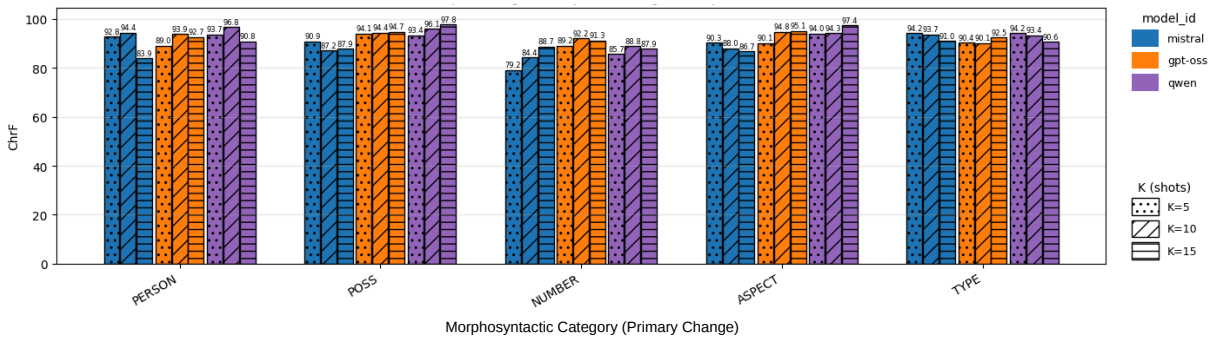

Figure 6: chrF scores by morphosyntactic category and model, disaggregated by the number of in-context examples (K).