Supervised Segmentation for Southern Quechua: Baselines, Priors, and LLM Augmentation

Anonymous ACL submission

Abstract

We propose *supervised* segmentation for Quechua (focus: Cusco/Southern) under low-resource conditions and analyze how *LLM-generated* synthetic data can augment scarce gold labels. We implement and compare: (i) BiLSTM boundary taggers (character vs. grapheme), (ii) BiLSTM-CRF, (iii) a Transformer seq2seq transducer, and (iv) hybrid models that inject unsupervised or rule-based priors (Morfessor, a Decision-Tree prior with a K-teacher, and an HMM prior) plus morphotactic rejection filters that eliminate spurious segmentations and prevent error propagation. Using a unified formulation and metrics, we summarize trade-offs across data regimes. We also document a practical, governance-compliant pipeline to produce few-shot LLM segmentations from a large crawl. We release code and trained models; a 920-word gold set is released, while a private 6,896-word set is used for training only.[1]

1 Introduction

Quechua forms long, morphologically complex words via agglutinative and derivational processes, so segmentation is a prerequisite for tagging, parsing, and translation (Adelaar and Muysken, 2004; Cusihuamán Gutiérrez, 2001). Gold-segmented data are limited. In this paper, we study label-efficient supervised segmenters and how carefully constrained large language model (*LLM*) synthetic examples can help. Our goals are as follows: (1) Implement a compact but representative set of supervised segmenters for Quechua; (2) provide a unified evaluation protocol; (3) document a reproducible and compliant synthetic-data pipeline. In doing so, we contribute:

- End-to-end implementations: BiLSTM (char/grapheme), BiLSTM-CRF, Transformer seq2seq, and hybrid priors (Morfessor, DT+K-teacher, HMM) with simple validators.
- A unified setup (normalization, graphemes, label projection, metrics) for boundary tagging and transduction.
- A crawl-based, few-shot LLM generation pipeline that honors data-governance constraints, with code and models publicly released.

The remainder of this paper is organized as follows. Section 2 provides essential linguistic context for understanding Quechua morphology, followed by our problem formulation in Section 3. We then detail our datasets (Section 4), review related work (Section 5), and present our supervised methods (Section 6) before describing our synthetic data pipeline (Section 7).

2 Linguistic Background

2.1 The Quechua Language Family

Quechua is a *family* spread across the central Andes with substantial internal diversity and uneven mutual intelligibility (Adelaar and Muysken, 2004; Mannheim, 1991). A canonical split distinguishes Quechua I (Central) and Quechua II (Peripheral), the latter including Southern Quechua (II-C), to which Cusco belongs (Torero, 2007; Adelaar and Muysken, 2004; Cusihuamán Gutiérrez, 2001). Grammars for Cusco/Cusco-Collao detail agglutinative morphotactics and rich derivational/inflectional inventories relevant to segmentation (Cusihuamán Gutiérrez, 2001). Glottolog provides up-to-date identifiers for varieties (Hammarström et al., 2024).

2.2 Orthography

We normalize to a practical grapheme inventory for tokenization: {`ch`, `ll`, `rr`, `tr`, `kw`, `ph`, `a`,

[1] Private data are not redistributed or sent to external LLMs.

b, d, e, f, g, h, i, k, l, m, n, ñ, o, p, q, r, s, t, u, v, w, x, y}. Southern Quechua contrasts plain/aspirated/ejective stops; orthographies mark aspiration with *h* and often treat *ch/ll/rr* as digraphs (Cusihuamán Gutiérrez, 2001; Adelaar and Muysken, 2004). Peruvian guidelines list official graphemes and variety-specific clusters (e.g., *tr*, *sh*) (Ministerio de Cultura del Perú, 2024). Cusco is typically analyzed with three vowels /i, a, u/ though five-vowel spellings persist regionally (Adelaar and Muysken, 2004; Cerrón-Palomino, 1992; Academia Mayor de la Lengua Quechua, 2005). We keep *k* vs. *q* distinct and retain Spanish letters for loans.

3 Problem Formulation

Quechua's agglutinative/derivational morphology yields long morpheme chains; segmentation is required for downstream NLP (Adelaar and Muysken, 2004; Cusihuamán Gutiérrez, 2001). Our gold supervision consists of human-produced morpheme boundaries (e.g., *rikuchkani* → *riku-chka-ni*). Let Σ be the grapheme set (Section 2.2). We study:

- **Boundary tagging:** predict a boundary vector $\mathbf{b} \in \{0,1\}^{T-1}$ over grapheme positions (BIO/BMES).
- **Transduction:** generate a segmented string (e.g., `seg1+seg2+...`) and convert to spans.

We report boundary F1, segment F1, and exact match. To evaluate these approaches, we utilize two complementary datasets that balance data availability with governance constraints.

4 Datasets

Private native-annotated (6,896 words): collected by a private researcher (who wishes to remain unnamed) from speakers of Southern Quechua. These stories were gathered for linguistic documentation but without the necessary permissions for public release. They may therefore only be used for appropriate research benefiting the community and cannot be redistributed or ingested by third-party language models. We deduplicate across sources and ensure that no private materials are used in prompts, demonstrations, or released artifacts. **Public** (920 words): consists of short narratives and personal recounts by community members describing their experiences during the COVID-19 pandemic. The 920-word subset will be released publicly with the code and models to support reproducible evaluation. **Data structure:** Each dataset is formatted as a two-column CSV file.

- **Column 1 – Word:** the raw surface form in standardized orthography.
- **Column 2 – Segmentation:** the corresponding morphological segmentation, represented as the same word with morpheme boundaries separated by whitespace.

For example, the entry `rikuchkani` in Column 1 would appear as `riku chka ni` in Column 2.

Before presenting our methods, we situate our work within existing approaches to Quechua morphological analysis and segmentation.

5 Related Work

Normalization/analysis. A cascaded FST+CRF normalizer for Southern Quechua gains ∼22% coverage when integrating a Spanish FST for loanword handling (Gonzales and Mamani, 2014); Rios (2015) describe the broader rule-based toolkit including orthographic standardization and morphological disambiguation. **FST analyzers.** Cusco-focused finite-state analyzers model morphotactics and morphophonology with high internal consistency but show reduced coverage for Bolivian, Ayacucho, and Collao varieties (Rios, 2015; Himoro and Pareja-Lora, 2022). Their limited generalization underscores the need for segmentation approaches that can adapt across dialects without extensive hand-crafting. **Subwording for MT.** Morphology-aware tokenization continues to yield gains in low-resource translation: BPE-Guided segmentation improves BLEU and chrF for Q→Es systems (Ortega et al., 2021); PRPE outperforms BPE and Unigram for Q–Es LSTM models and achieves better morpheme alignment (Chen and Fazio, 2021); and PRPE enhances NER precision for morphologically rich classes in QuBERT experiments (Zevallos et al., 2022). Cross-language results remain mixed, suggesting that segmentation granularity interacts with both morphological complexity and corpus balance (Sälevä and Lignos, 2021). **Resources.** Available Quechua resources include QuBERT (450k normalized Southern Quechua sentences with tokenization variants and a BERT model) (Zevallos et al., 2022) and the AmericasNLP shared benchmarks for multilingual evaluation (Vázquez et al., 2021; Ebrahimi et al., 2023; Moreno, 2021). Together, these works frame

the need for controlled morphological segmentation datasets and interpretable baselines for agglutinative, low-resource languages such as Quechua.

Building on their theoretical foundations while addressing their inherent unsupervised limitations, we now present our supervised segmentation framework.

6 Supervised Methods for Quechua

6.1 Extracting Linguistic Features

To strengthen the Decision Tree (DT) prior and provide additional linguistic signal in this low-resource setting, we enriched the training data with a set of surface-level phonological and morphological features derived from the gold segmentations. We selected a DT-based prior because Quechua is a highly derivational and suffixing agglutinative language, where morphemes attach in largely deterministic and hierarchical order. This structured morphology allows rule-like dependencies between suffix positions, which a DT can capture through interpretable feature splits. We hypothesized that encoding such regularities explicitly would provide meaningful inductive bias and complement the downstream neural models trained on sparse data.

Each word in the annotated dataset was represented as a consonant-vowel (CV) sequence, along with a trimmed version corresponding to its final segment. Words without internal segmentation were excluded to ensure consistent structure. Using these representations, we computed several interpretable linguistic indicators: overall word length, counts of vowels and consonants across the entire word, counts of vowels and consonants in the final segment, the number of morpheme splits, and the frequency of the glides *y* and *w* both across the word and within non-initial morphemes.

These features capture regularities such as syllable composition, vowel-consonant balance, and morphophonemic patterns common in Quechua word formation. The enriched training data were then used to train the DT prior jointly with the gold segmentation labels, allowing the model to exploit explicit linguistic structure alongside distributed representations from the neural encoder.

6.2 Model Architectures

We implement a progression of learners that all predict morpheme boundaries over grapheme sequences but differ in tokenization, architecture, priors, and decoding:

- **BiLSTM (char)**: Embedding → BiLSTM → per-position logits; masked BCEWithLogitsLoss; grid over size/dropout/LR; greedy decoding.
- **BiLSTM (grapheme)**: NFC+lowercase; apostrophe harmonization; longest-match grapheme tokenizer; same model/training as above.
- **BiLSTM-CRF**: BiLSTM emissions with a CRF and Viterbi decoding; the notebook also includes dataset diagnostics (length vs. morpheme count, basic regressors, outlier removal).
- **Transformer seq2seq**: encoder-decoder with positional encodings; input characters, output uses + to join segments; autoregressive decoding.
- **Morfessor-informed BiLSTM**: ensemble Morfessor over raw text yields per-position boundary probabilities; concatenated as an extra feature channel to embeddings.
- **BiLSTM + DT prior + K teacher**: a Decision Tree predicts local boundary priors (fused into the BiLSTM); a regression-based K-teacher (counts-based privileged features) regularizes the expected number of cuts; decoding via a lexicon-constrained beam with cut/OOV penalties.
- **BiLSTM + DT prior + Markov filter**: Again, Decision Tree predicts local boundary priors (fused into the BiLSTM) but we replace beam search with a simple suffix-set validator; report both "reject invalid as no-cuts" and "valid-only" modes. A suffix-HMM (forward-backward) provides globally coherent boundary posteriors concatenated to embeddings; same suffix validator; hyperparameters tuned with Optuna.[2]

6.3 Novelty and Design Motivation

Unlike prior low-resource segmentation systems that rely solely on supervised sequence tagging

[2]The validator maintains a suffix inventory extracted from training data and rejects any segmentation containing morphemes outside this set, preventing the model from hallucinating non-existent suffixes.

or unsupervised morphological priors, our framework integrates **structured global priors** and **governed post-hoc validation** into a unified supervised training pipeline. Specifically, we introduce a **suffix-informed HMM prior** that constrains the sequence model toward morphologically coherent boundaries, and a **hard rejection validator** that enforces linguistic consistency at inference time without access to ground truth. This combination forms a lightweight, reproducible control layer that improves segmentation quality under limited supervision, and paired with governed LLM-based data augmentation constitutes a new, generalizable strategy for low-resource morphological analysis.

Given the severe data constraints identified above (only 6,896 training words), we explore whether modern LLMs can provide useful training signal through synthetic segmentations.

6.4 Rejection Filter Mechanism

The rejection filter serves as a critical post-processing component that validates predicted segmentations against known morphotactic constraints. Specifically, the filter rejects any spurious segmentations predicted by the BiLSTM that violate Quechua's suffix inventory or morpheme ordering rules, preventing cascade errors from propagating to downstream NLP tasks. This hard constraint enforcement operates by: (1) checking each predicted segment against a curated suffix lexicon derived from training data, (2) validating morpheme sequences against observed morphotactic patterns, and (3) rejecting entire word segmentations that contain invalid morphemes, defaulting to unsegmented forms rather than propagating errors. This conservative approach trades overall recall for precision, ensuring that only linguistically valid segmentations pass through the pipeline.

7 Synthetic Data via LLMs

We augment training data with LLM segmentations in a way that avoids exposing private examples.

Candidate selection. From a large crawl, we keep only tokens fully tokenizable by our grapheme inventory (Section 2.2). For each token we compute a 4-grapheme "root". We load two labeled resources (our training and validation datasets), extract their first segment as a root, and select *unlabeled* corpus tokens whose root lies in the intersection, biasing toward in-domain morphotactics while avoiding label leak.

Few-shot pool. We construct word analysis data by intersecting the two labeled sources and the crawl at surface+root level, keeping items with agreeing first segments. This file (publicly permissible items only) provides {word, segmentation} pairs for few-shot prompts.

Prompting & API. For each target word, we sample 37 few-shot pairs and use a single instruction prompt ("segment into morphemes; output only segments, space-separated"). The number 37 was chosen to remain within the context-length and rate-limit constraints of the API while preserving representative coverage of morphological patterns. We query `gpt-4o` with `temperature=0` and robust rate-limit handling (honors `Retry-After`; marks failures as `[API-FAILED]`). All outputs are written to disk.

Label projection and mixing. Accepted LLM outputs are converted to boundary labels (and BIO/BMES when needed) and concatenated with gold data for training. In our codebase, mixing is implemented as simple joint training (shuffle each epoch).

8 Experimental Setup

Preprocessing. All text is NFC-normalized, lowercased, and stripped of control characters. We require inputs to be tokenizable by the grapheme inventory (Section 2.2); Spanish letters are kept for loans. Grapheme longest-match tokenization feeds boundary taggers; seq2seq uses character inputs with +-joined targets.

Systems compared. We report the systems in Section 6 plus trivial references (NoCuts, AllBetweenChars). Decoding is greedy (taggers), Viterbi (CRF), autoregressive (seq2seq), or validated (beam or Markov filter).

Hyperparameters and compute. Neural models use Adam/AdamW with gradient clipping (1.0), dropout, and early stopping on dev boundary F1. We sweep small grids: embedding {64–256}, BiLSTM hidden {128–384}, layers {1–2}, dropout {0.1–0.5}, LR in $[10^{-4}, 10^{-3}]$; for seq2seq $d_{\text{model}} \in [128, 384]$, heads {4,8}, layers {2–4}. Prior weights ($\lambda_{\text{prior}}, \lambda_K$) are tuned in short ranges; HMM variants use Optuna. Runs fit on a single commodity GPU (or CPU with smaller batches).

Evaluation protocol. We report boundary F1, segment F1, and exact match; results are averaged

over multiple seeds (mean±stdev). The public 920-word set serves as test data; we carve a small dev split from training for early stopping. LLM outputs are used only for training; private items are never sent to external APIs.

Having established our experimental framework, we now evaluate the performance of our proposed methods across three key dimensions: baseline supervised performance, the impact of synthetic data augmentation, and ablation studies on data selection strategies.

9 Results

We report Exact Match Accuracy, defined as the percentage of words in the test set for which the predicted segmentation exactly matches one of the acceptable gold-standard variants. We analyze the progression from baseline architectures to our proposed hybrid models, the impact of augmenting training data with synthetically segmented words from LLMs, and ablation studies regarding the quantity and selection of synthetic data.

As no prior supervised segmentation systems exist for Quechua, there are no directly comparable baselines. To the best of our knowledge, this work constitutes the first linguistically informed supervised segmentation approach developed for the language.

9.1 Supervised Baselines

We first evaluated several purely supervised architectures to establish a baseline and determine the most effective core segmenter. The results are summarized in Tables 1 and 2.

Our initial character-level BiLSTM achieved an accuracy of 50.4%. Moving to a grapheme-based tokenization aimed at capturing Quechua digraphs did not immediately improve performance, yielding 49.3%. We observed moderate improvements by incorporating structural constraints and morphological features. Adding a Conditional Random Field (CRF) layer improved accuracy to 55.0%, while incorporating unsupervised features from Morfessor as an input channel reached 54.4%.

Notably, a sequence-to-sequence Transformer approach performed poorly with 38.5% accuracy. This confirms that for low-resource scenarios with limited training pairs, inductive biases present in BiLSTM-based sequence taggers are more effective than data-hungry Transformer architectures.

We then evaluated our proposed hybrid architectures, which incorporate linguistic priors during training and a rule-based rejection filter during post-processing. The complex model, utilizing a beam search and multiple feature sets, achieved 53.5%. However, our streamlined architectures proved superior.

Model Architecture	**Exact Match Accuracy**
Transformer Seq2Seq	38.5%
BiLSTM (Character)	49.3%
BiLSTM (Grapheme)	52.8%
BiLSTM + Morfessor Features	54.4%
BiLSTM-CRF	55.0%
BiLSTM + DT Prior	55.5%
BiLSTM + HMM Prior	**58.0%**
With Rejection Filter:	
BiLSTM + DT Prior (Filtered)	69.0%
BiLSTM + HMM Prior (Filtered)	**70.0%**

Table 1: Exact Match Accuracy of Supervised Segmentation Models.

The DT-LSTM-Markov model, utilizing a local Decision Tree prior, achieved a raw accuracy of 55.5%. Applying the post-processing suffix rejection filter significantly boosted this to 69.0%. This substantial improvement (13.5% absolute) demonstrates the filter's effectiveness in eliminating spurious segmentations: the BiLSTM may hallucinate morpheme boundaries that create invalid suffixes or violate Quechua's morphotactic constraints, but the rejection filter catches these errors and prevents them from corrupting downstream analyses. Rather than propagating uncertain or malformed segmentations, the filter conservatively maintains the original unsegmented form when validation fails.

The boundary-level F1 scores of the segmentation models follow a similar overall pattern, with one notable deviation. The DT-LSTM-Markov model achieves higher boundary F1 scores than the Markov-LSTM-Markov model with HMM priors. This difference arises from the DT model's tendency toward over-segmentation: it predicts a larger number of suffix boundaries, which inflates boundary-level recall and consequently the F1 score. However, this same behavior leads to lower overall segmentation accuracy, as excessive boundary predictions result in incorrect morpheme splits.

9.2 Impact of Synthetic Data

To address data scarcity, we explored augmenting our training set with words segmented by LLMs

Model Architecture	Precision	Recall	F1
Transformer Seq2Seq	0.725	0.718	0.722
BiLSTM (Character)	0.803	0.835	0.800
BiLSTM (Grapheme)	0.820	0.847	0.818
BiLSTM + Morfessor Features	0.8213	0.8546	0.8186
BiLSTM-CRF	0.782	0.832	0.806
BiLSTM + DT Prior	0.815	0.881	0.847
BiLSTM + HMM Prior	**0.782**	**0.852**	**0.816**
With Rejection Filter:			
BiLSTM + DT Prior (Filtered)	**0.945**	**0.964**	**0.955**
BiLSTM + HMM Prior (Filtered)	0.916	0.934	0.925

Table 2: Boundary-level Precision, Recall, and F1 scores for supervised segmentation models.

(GPT-4o and GPT-3.5) via few-shot prompting. We utilized our best-performing supervised model (Markov-LSTM-Markov) as the base for these experiments.

Augmenting the training data with 2,869 words segmented by GPT-4o yielded substantial improvements. The raw model accuracy increased from the baseline of 58.0% to **63.7%**. When combined with the rejection filter, the accuracy reached **74.8%**, demonstrating that synthetic data can effectively teach the model valid morphological patterns that pass the suffix validator.

We observed a distinct difference in the utility of data generated by different LLMs. In a controlled comparison using 469 common words, augmentation with GPT-3.5 segmentations resulted in a performance of 55.2% (raw) and 66.3% (filtered), which was detrimental compared to the non-augmented baseline. Conversely, GPT-4o generated data that improved performance.

Qualitative analysis of the LLM outputs revealed that while both models provided confident linguistic justifications for their segmentations (citing roots, evidentials, and agglutinative rules), GPT-4o's outputs aligned better with our gold standard. GPT-3.5's explanations, while linguistically detailed, often accompanied segmentations that did not generalize well to the test set. GPT-4o demonstrated some awareness of its own potential for errors, explicitly noting possibilities of "over-segmentation" or "artifacts of overly literal parsing" in its justifications, yet its generated training data proved empirically more robust.

9.2.1 GPT-4o vs. GPT-3.5 for Synthetic Data

We observe that the reduced performance of data generated by GPT-3.5 arises from an interaction between tokenization, model capacity, and noisy pretraining priors. Let T denote the subword tokenizer mapping a surface string w to tokens $(t_1, \ldots, t_k)$. When T fragments Quechua roots or suffix clusters (e.g., $T(\texttt{kasu}) = (\texttt{ka}, \texttt{su})$, $T(\texttt{-naku-}) = (\texttt{na}, \texttt{ku})$), the objective "insert spaces at morpheme boundaries" becomes ill-posed: the model's natural breakpoints align with subword cuts rather than morphs.

In few-shot conditions, the conditional distribution $p_\theta(y \mid x, \mathcal{D}_{\text{shots}})$ reflects tension between a strong pretraining prior, favoring high-frequency subword continuations, dialectal normalizations, and gloss-like formatting, and a weak in-context likelihood derived from the small rule set in the prompt. As capacity increases, this prior dominates, yielding outputs shaped by internal regularities (e.g., hyphenated glossing, over-segmentation) instead of the intended surface segmentation policy. This produces characteristic "subword artifact exposure" (`k u`, `qku`) and drift toward analyses the model "knows," even when they contradict the input.

Beyond tokenization, larger models amplify spurious pattern induction. Given heterogeneous Quechua data, a high-capacity model readily infers latent rules that overgeneralize frequent shapes (e.g., `-na` $\leftrightarrow$ `-ni`, `-qti` $\mapsto$ `qta`) and "helpfully" normalizes or reanalyzes the surface, drifting from segmentation toward interlinear glossing. This explains the emergence of hyphenation, invented morphemes, and violations of root-shape expectations (e.g., shredded bisyllabic roots) in GPT-3.5. In contrast, GPT-4o or smaller models, with flatter priors or better tokenizer alignment, adhere more closely to local constraints and behave as conservative segmenters rather than analysts. Thus, scaling gains in general reasoning can *reduce* robustness on narrow, morphologically rigid tasks: larger models find patterns, including incorrect ones, when supervision is limited and the tokenization-morphology interface is misaligned.

9.3 Ablation Studies

We conducted ablation studies to analyze the effects of the quantity and selection strategy of the synthetic training data. These experiments used a slightly different baseline split than that used in

Table 3: Ablation on Synthetic Data Quantity and Selection Strategy

Synthetic Data Added	Raw Accuracy	With Rejection Filter
Baseline (0 words)	56.3%	67.0%
+ First 100	58.4%	68.5%
+ First 200	**62.1%**	72.7%
+ First 300	60.7%	69.1%
+ Random 100	61.3%	**72.8%**

Section 9.1. We report the results of the ablation studies in Table 3.

Starting with a baseline of 0 synthetic words (56.3% raw / 67.0% filtered), we incrementally added synthetic data selected from the beginning of the corpus ("First N") and compared it to random sampling.

Adding the "First 100" words improved filtered accuracy to 68.5%. Performance peaked with the "First 200" words at **72.7%** filtered accuracy (62.1% raw). However, increasing the set to the "First 300" words caused a degradation in performance to 69.1% filtered accuracy. This suggests that beyond a certain point, the noise inherent in synthetic segmentation outweighs the benefit of additional training examples.

Furthermore, selection strategy matters. Adding 100 *randomly sampled* synthetic words achieved a filtered accuracy of **72.8%** (61.3% raw). This outperforms the "First 100" strategy and matches the peak performance of the "First 200", indicating that a diverse selection of synthetic examples is more valuable than a larger, potentially homogenous block of data.

These quantitative results reveal important patterns that warrant deeper investigation. To understand the systematic behaviors underlying model performance, we conducted a detailed error analysis, described in the next section.

10 Error Analysis

A qualitative and corpus-linguistic analysis was conducted in collaboration with trained linguists specializing in Quechua morphology. Their evaluation, combined with quantitative inspection of model outputs, revealed several systematic tendencies in segmentation behavior.

We find that Quechua word segmentation generally follows a **linear morphemic trajectory**, reflecting the language's strongly agglutinative structure. Words that deviate from this broad linearity tend to trigger characteristic model errors: **longer words** are frequently *over-segmented*, while **shorter or morphologically opaque words** are often *under-segmented*.

Another consistent source of difficulty is the persistence of **Spanish loanwords** in Quechua texts. These borrowings often violate Quechua's canonical bisyllabic root pattern, thereby confusing both sequence models and morphologically informed priors. For instance, the word *anchayllapunitaq* has been consistently mis-segmented by all models in our study, illustrating the model's sensitivity to length and morpheme stacking. Similarly, *explicamuwanku* contains the Spanish-derived root *explica*, which falls outside Quechua's traditional root formation constraints and leads to unreliable boundary predictions. Conversely, the word *umankus* is typically under-segmented as *umanku-s* instead of the correct *uma-nku-s*, showing how short and phonologically simple roots may be incorrectly treated as monomorphemic.

Overall, these findings highlight both the morphological regularity and the contact-induced variability of Quechua, underscoring the importance of linguistically informed priors and curated data in achieving robust segmentation performance.

These systematic patterns and failure modes inform a set of practical recommendations for future work on morphological segmentation in similar low-resource settings.

11 Best Practices and Recommendations

Based on our extensive experimentation with supervised segmentation for Southern Quechua, we recommend the following practices for future work on morphological analysis of low-resource agglutinative languages:

Architectural Choices: For extremely low-resource scenarios (<10K labeled words), BiLSTM-based sequence taggers with linguistic priors significantly outperform Transformer architectures. The inductive biases of recurrent models prove more valuable than the representational capacity of Transformers when data is scarce.

Prior Integration: Combining soft priors during training (via Decision Trees or HMMs) with hard rejection filters during inference yields substantial improvements (58%→70% in our experiments). The rejection filter specifically prevents spurious segmentations by validating each predicted morpheme against known suffix inventories and mor-

photactic patterns, blocking errors from cascading through the NLP pipeline. This two-stage approach—soft guidance during training followed by hard constraints during inference—proves more effective than either strategy alone. We recommend implementing suffix validators based on observed morphotactic patterns as a lightweight post-processing step that can catch and eliminate implausible morpheme sequences that neural models might confidently but incorrectly predict.

Grapheme Tokenization: Despite minimal immediate gains, grapheme-aware tokenization that respects language-specific digraphs (*ch*, *ll*, *rr*) provides a more principled foundation for morphological modeling and should be preferred over character-level processing.

LLM Augmentation Strategy: When using LLM-generated synthetic data: (1) Use stronger models (GPT-4 class) over weaker ones despite cost differences; quality matters more than quantity; (2) Implement root-based selection to bias toward in-domain morphology; (3) Limit augmentation to 100–200 carefully selected examples to avoid noise accumulation; (4) After capping the augmentation set to 100–200 examples, select them randomly rather than sequentially to minimize overgeneralization and pattern memorization by the LLM.

Data Governance: Maintain strict separation between private linguistic resources and external API calls. Our pipeline demonstrates that meaningful augmentation is possible without exposing sensitive indigenous data to commercial systems.

12 Conclusion

We presented a comprehensive study of supervised morphological segmentation for Southern Quechua under low-resource conditions, comparing multiple neural architectures and demonstrating how carefully constrained LLM-generated synthetic data can augment scarce gold annotations. Our key contributions include: (1) systematic implementation and evaluation of supervised segmenters ranging from BiLSTM taggers to Transformer transducers; (2) novel hybrid architectures combining neural sequence models with linguistic priors and morphotactic validators; (3) a governance-compliant pipeline for LLM augmentation that respects indigenous data sovereignty; and (4) empirical demonstration that quality matters more than quantity in synthetic data selection.

Our best-performing system, a BiLSTM with Suffix HMM priors and rejection filtering, augmented with 200 GPT-4 synthetic examples, achieves 74.8% exact match accuracy, improving 26.3% absolute over naive baselines. The success of linguistically-informed priors and hard rejection filters (improving accuracy by 12%) underscores the value of incorporating explicit morphological knowledge even in neural approaches. Our analysis reveals that model failures concentrate on Spanish loanwords and morphologically complex long words, pointing toward necessary future improvements.

This work establishes the first supervised segmentation baselines for Quechua and demonstrates that careful integration of linguistic knowledge, neural learning, and synthetic augmentation can partially overcome data scarcity. Future directions include: expanding to other Quechua varieties, investigating cross-dialectal transfer, developing better handling of code-mixed text, and creating interactive tools that allow native speakers to correct and improve model outputs. Most importantly, we emphasize that technological progress must align with community goals for language maintenance and revitalization; our methods are means toward supporting indigenous language communities, not ends in themselves.

13 Limitations

Our study has several important limitations that should guide interpretation and future work:

Data Scale: Our experiments use only 6,896 training words and 920 test words, which limits the generalizability of our findings. While this reflects real-world constraints in indigenous language documentation, larger-scale evaluation would strengthen our conclusions.

Dialectal Coverage: We focus exclusively on Cusco/Southern Quechua (Quechua II-C). Given the substantial internal diversity of the Quechua family and uneven mutual intelligibility across varieties, our models may not transfer to Central Quechua or other Peripheral varieties without retraining.

Annotation Consistency: The private dataset comes from multiple annotators with potentially varying segmentation philosophies. While we attempt to harmonize these through preprocessing, inconsistencies in boundary decisions (particularly for derivational morphology) may limit model performance.

LLM Dependency: Our augmentation strategy relies on commercial LLM APIs, introducing concerns about reproducibility, cost, and long-term availability. The quality of synthetic data is also sensitive to prompt engineering and model updates beyond our control.

Evaluation Metrics: Our exact match metric is strict and may undervalue partially correct segmentations. Alternative metrics considering partial credit for boundary placement could provide more nuanced assessment.

Spanish Loanword Handling: Our models struggle with Spanish borrowings that violate Quechua morphotactics. Given the prevalence of code-mixing in contemporary Quechua texts, this represents a significant practical limitation.

Computational Requirements: While our models are lightweight, the LLM augmentation step requires API access and incurs costs that may be prohibitive for some research contexts.

14 Ethics Statement

This work involves Southern Quechua, an indigenous language spoken primarily in Peru and Bolivia. We acknowledge several ethical considerations:

Indigenous Data Sovereignty: We respect the principle that indigenous communities maintain authority over their linguistic and cultural data. The private dataset (6,896 words) was created by native speakers under specific governance constraints that forbid redistribution without conset and external LLM ingestion. We strictly honor these constraints throughout our pipeline.

Community Benefit: While our technical contributions advance NLP capabilities, we recognize that technology development must serve community needs. We commit to making our models freely available and prioritize applications that support language maintenance and education.

Cultural Sensitivity: Morphological segmentation is not culturally neutral; different communities may have varying perspectives on word analysis. Our gold standard represents one analytical tradition and should not be imposed as the sole correct interpretation.

Synthetic Data Risks: Using LLMs to generate Quechua segmentations raises concerns about perpetuating errors or biases present in limited training data. We mitigate this through careful validation but acknowledge that synthetic data can never fully replace native speaker knowledge.

Digital Divide: Our reliance on computational resources and API access may exclude community members without technical infrastructure. Future work should prioritize accessibility and develop tools usable in low-connectivity contexts.

Consent and Attribution: All human annotators provided informed consent for their linguistic contributions. However, we note that the crawled web data used for candidate selection may include text from authors who did not explicitly consent to computational analysis.

No Harm Principle: We have designed our pipeline to prevent exposure of private indigenous data to commercial systems while still benefiting from modern NLP techniques. This balanced approach aims to advance scientific knowledge while protecting community resources.

This work addresses a fundamental challenge in low-resource NLP: developing effective morphological segmentation with minimal supervised data while respecting indigenous data governance.

References

Academia Mayor de la Lengua Quechua. 2005. *Diccionario Qheswa–Español / Español–Qheswa*. Gobierno Regional Cusco, Cusco. Ortografía de cinco vocales en uso regional.

Willem F. H. Adelaar and Pieter C. Muysken. 2004. *The Languages of the Andes*. Cambridge University Press, Cambridge.

Rodolfo Cerrón-Palomino. 1992. Standardization in andean languages. *Working Papers in Educational Linguistics (WPEL)*, 8(1):1–13.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT) at MT Summit XVIII*, pages 20–31, Virtual.

Antonio Cusihuamán Gutiérrez. 2001. *Gramática quechua : Cuzco-Collao : Antonio Cusihuaman Gutiérrez*, 2. ed edition. Centro Bartoloméde las Casas, Lima.

Amin Ebrahimi, John Ortega, Katharina Kann, Manuel Mager, and et al. 2023. Findings of the americasnlp 2023 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP 2023)*, pages 215–235.

Annette Rios Gonzales and Richard Alexander Castro Mamani. 2014. Morphological disambiguation and text normalization for southern quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 39–47, Dublin, Ireland.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog database 5.0. Zenodo.

Marcelo Yuji Himoro and Antonio Pareja-Lora. 2022. Preliminary results on the evaluation of computational tools for the analysis of quechua and aymara. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, pages 5450–5459, Marseille, France.

Bruce Mannheim. 1991. *The Language of the Inka since the European Invasion*. University of Texas Press, Austin.

Ministerio de Cultura del Perú. 2024. Quechua (alfabeto oficial y normativas vigentes). Base de Datos de Pueblos Indígenas del Perú (BDPI). Cita las resoluciones: R.M. Nº 1218-85-ED; R.D. Nº 0282-2013-ED (3 vocales); R.D. Nº 0293-2013-ED.

Oswaldo Moreno. 2021. The repucs spanish–quechua submission to the americasnlp 2021 shared task on open machine translation. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP 2021)*, pages 265–271.

John E. Ortega, Richard A. Castro Mamani, and Kyunghyun Cho. 2021. Neural machine translation with a polysynthetic low-resource language. *Machine Translation*, 34(3):325–346.

Annette Rios. 2015. *A Basic Language Technology Toolkit for Quechua*. Ph.D. thesis, University of Zurich, Zurich, Switzerland.

Jenna Sälevä and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the EACL 2021 Student Research Workshop*, pages 164–174.

Alfredo Torero. 2007. *El quechua y la historia social andina*. Fondo Editorial del Pedagógico San Marcos, Lima. 1st ed. 1974.

Raúl Vázquez, Sami Virpioja, Jörg Tiedemann, and et al. 2021. The helsinki submission to the americasnlp 2021 shared task. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP 2021)*, pages 199–209.

Rodolfo J. Zevallos, John E. Ortega, William Tzu-Liang Chen, Richard A. Castro Mamani, and Brett Fazio. 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 NLP (DeepLo)*, pages 1–13, Seattle, USA.

A Prompts Used

This appendix lists the exact or representative prompts used for generating synthetic segmentation data.

A.1 Zero-shot Prompt

The following system message was used for zero-shot segmentation queries:

```
system_message = (
"You are an expert in Quechua
linguistics. Your task is to segment a
given Quechua word into its constituent
morphemes. "
"The morphemes should be separated
by spaces. Please provide only the
segmented output, "
"with no additional explanation or
commentary."
)
```

A.2 Few-shot Prompt (Pseudocode Description)

For few-shot prompting, the model was given the same task as above but preceded by a small number of {word, segmentation} examples drawn from the gold data. The construction of the prompt follows this logic:

```
1. Initialize a string variable prompt
with task instructions identical to
the zero-shot system message.

2. Append the section header
"Examples:".

3. For each example (word, segmentation)
pair in the training data:
a. Retrieve the gold segmentation from
either Morph_split_str
or Morph_split fields.
b. If Morph_split is a Python list, join
its elements with spaces.
c. Add a line "<word>: <segmentation>"
to the prompt.

4. Append a new section header "Words to
Segment:".

5. For each word in the target set
(words_to_segment):
Add a line "<word>" to the prompt.

6. Send this complete prompt string to
the LLM with temperature = 0.
```

This few-shot prompt format provides explicit segmentation examples before asking the model to segment unseen Quechua words.