CART: Stability-Aware Evidence Selection for Cascade-Augmented Retrieval in Low-Resource Machine Translation

Anonymous ACL submission

Abstract

Low-resource languages (LRLs) with complex morphosyntax and limited usable parallel data remain challenging for modern translation systems. Although large language models (LLMs) enable cross-lingual transfer, their translations for LRLs often exhibit unstable decoding under weak lexical grounding, leading to semantic drift and hallucination. We study this failure mode in Lakota, a morphologically rich Indigenous language lacking any standard MT system, and introduce CART, an inference-time Cascade-Augmented Retrieval Translation framework that mitigates instability without finetuning or language-specific analyzers. CART progressively constrains model behavior through input canonicalization, multi-channel retrieval, and stability-aware evidence selection, which filters retrieved examples based on whether they induce consistent short-form translations under minimal prompt perturbations rather than similarity alone. On a heterogeneous English–Lakota corpus of approximately 70k mixed-source pairs, CART improves EN→LKT BLEU by +5.8 over direct prompting and by +2.9 over similarity-based retrieval, with corresponding BERTScore gains of +0.07 and +0.03, respectively. These results suggest that generator-aware evidence selection can reduce hallucination and improve robustness in morphologically rich LRL settings, with Lakota serving as an illustrative case study.

1 Introduction

Recent advances in large language models have shifted machine translation (MT) away from purely parameter-centric learning and toward inference-time reasoning. Modern systems increasingly blend generation with external evidence, reflecting a broader recognition that many translation errors arise not simply from limited model capacity but from insufficient alignment between the model's internal representations and the linguistic patterns present in the input (Haddow et al., 2022). Retrieval provides a way to inject context explicitly, but its reliability varies depending on the linguistic and resource conditions of the target language (Ranathunga et al., 2023). In low-resource settings, insufficient grounding often manifests behaviorally as unstable decoding.

For LRLs, these conditions differ sharply from those of high-resource settings. Many LRLs exhibit rich morphology, flexible constituent order, and orthographic variation, all of which complicate surface-level matching and reduce the reliability of generation (Ataman and Federico, 2018). Lakota, a Plains Siouan language spoken in the northern United States, exemplifies these challenges (Mager et al., 2023). It employs primarily SOV order, encodes argument structure through intricate verbal morphology, and contains highly productive affixation and phonological alternation, features that cause the same lexical item to appear in multiple shapes (Nzeyimana, 2024). Under these conditions, semantically similar but structurally incompatible retrieved examples may encode incompatible morphological or argument-structure realizations, causing similarity-based retrieval to destabilize generation instead of supporting it (Hangya et al., 2023).

Languages like Lakota face acute scarcity of digitized text and parallel corpora, and existing materials often consist of dictionaries, learning resources, and scattered community efforts rather than large-scale parallel data (Chen and Abdul-Mageed, 2023). Under such constraints, designing systems that maximize the utility of available evidence becomes more important than developing increasingly complex architectures tailored to high-resource conditions (Ranjith et al., 2023).

These challenges also complicate evaluation. Languages with rich morphology and flexible word order permit multiple valid surface realizations, which can limit the interpretability of surface-oriented metrics such as BLEU (Lakew et al., 2020;

Adelani et al., 2022). We therefore report multiple metrics and complement them with qualitative analysis.

Finally, many Indigenous and endangered languages are central to collective identity and cultural continuity (Mager et al., 2023). In communities working to maintain or revitalize such languages, access to reliable technological support can influence whether future learners engage with the language at all. As fluent speakers age and archival resources remain limited, the window for creating usable tools is narrowing. Developing translation methods that are accurate, adaptable, and respectful of linguistic norms therefore has implications beyond technical performance.

These considerations illustrate one viable design approach for translation systems in low-resource settings. Rather than modifying core architectures, we focus on inference-time mechanisms that progressively constrain model behavior. In particular, we study how normalization, retrieval, and evidence filtering can be composed to remove examples that destabilize decoding, yielding more faithful translations in morphologically rich, low-resource settings.

2 Related Works

Research on low-resource MT spans data augmentation, multilingual transfer, structural modeling, and, more recently, retrieval-augmented generation (RAG) and retrieval-augmented prompting with large language models. We group related work into two strands most relevant to our approach: (1) data-centric and multilingual strategies, and (2) structure-aligned and retrieval-based methods.

2.1 Data-Centric Approaches

A large body of work addresses low-resource MT by increasing effective data scale. Back-translation (Sennrich et al., 2016) augments parallel corpora using synthetic data, while multilingual parameter sharing (Johnson et al., 2017) and massively multilingual models (Aharoni et al., 2019; Arivazhagan et al., 2019) leverage cross-lingual transfer from high-resource languages. Pretrained sequence-to-sequence models such as mBART (Liu et al., 2020) and large multilingual MT systems such as M2M-100 (Fan et al., 2021) further extend this paradigm through joint training and cross-lingual denoising.

However, these approaches rely on assumptions that frequently fail for endangered or structurally complex languages, including the availability of substantial monolingual data and the presence of closely related, well-represented languages (Ranjith et al., 2023). When these assumptions do not hold, multilingual systems may exhibit severe degradation under domain shift or typological divergence, often producing ungrammatical or semantically incoherent output when the target language is absent from training data (Kocmi and Federmann, 2023). These limitations motivate inference-time methods that incorporate external evidence rather than relying solely on parameter scaling.

2.2 Structure-Aware and Retrieval-Based Methods

A complementary line of work incorporates linguistic structure to better handle morphologically rich languages. Morphology-aware encodings (Ataman and Federico, 2018) and systems incorporating explicit morphological features (Nzeyimana, 2024) reduce sparsity, while language-specific pretrained encoders such as Kinyabert (Nzeyimana and Rubungo, 2022) capture finer-grained regularities. In practice, however, such approaches often require morphological analyzers, annotated corpora, or standardized orthography, resources that are unavailable for many low-resource languages.

Retrieval-based methods instead introduce external evidence at inference time. Parallel corpus mining with multilingual embeddings (Artetxe and Schwenk, 2019) demonstrates that targeted retrieval can outperform larger but noisier datasets. More recent work on retrieval-augmented prompting and fragment-based in-context learning (Merx et al., 2024; Frontull and Ströhle, 2025) shows that providing aligned exemplars can improve consistency in low-resource translation. In-context learning studies (Agrawal et al., 2023; Cahyawijaya et al., 2024) further demonstrate that structure-aligned examples can guide generation in the absence of task-specific training.

Related work has also explored self-consistency decoding (Wang et al., 2022) and uncertainty-aware generation using confidence or log-probability statistics, primarily by resampling or aggregating multiple outputs at decoding time.

However, existing retrieval-augmented approaches typically select evidence based on semantic or structural similarity alone and treat retrieval as independent of the generator's decoding behavior. This assumption can break down in low-resource settings, where semantically sim-

ilar but structurally incompatible examples may induce divergent or unstable outputs. Unlike self-consistency or uncertainty-based decoding methods, which resample or aggregate outputs at decoding time, prior retrieval-based MT work has not systematically treated decoding instability under minimal prompt perturbations as a first-class failure mode. In contrast, our work explicitly treats instability under minimal prompt perturbations as a negative signal and introduces generator-aware evidence filtering to discard retrieved examples that consistently destabilize generation prior to decoding.

Together, these findings suggest that while retrieval and structural alignment are valuable, effective low-resource MT also requires mechanisms that account for how retrieved evidence interacts with the generator's decoding behavior, particularly in languages characterized by rich morphology, surface variation, and sparse supervision.

3 Methodology

3.1 Task Setting and Design Rationale

We consider the bidirectional translation task between English and Lakota using an LLM augmented with external linguistic evidence, where the input is x and the output is y. In this setting, weak lexical grounding often leads the model to entertain multiple competing interpretations, which manifests behaviorally as unstable decoding. Small prompt perturbations can yield divergent translations, and CART is designed around this failure mode. Supporting steps such as normalization and multi-channel retrieval improve evidence coverage and alignment, while the primary control signal comes from discarding examples that destabilize generation, yielding a compact and reliable context for deterministic decoding. CART follows a staged, inference-time cascade in which each component progressively constrains the model's hypothesis space, from input normalization and retrieval to stability-aware filtering and deterministic decoding.

3.2 Candidate Evidence Preparation

3.2.1 Segmentation

Inputs exceeding sentence length are segmented using punctuation-based sentence boundary detection prior to normalization and retrieval. This segmentation is applied uniformly to both source inputs and reference materials, ensuring that retrieval operates over comparable linguistic units. Paragraph-level translation is supported by translating segments independently and concatenating outputs; however, we focus on sentence-level translation in this work.

3.2.2 Normalization

Normalization applies lightweight, surface-preserving transformations that reduce orthographic variability in Lakota while maintaining linguistic identity. Lakota sources exhibit substantial variation in the representation of diacritics, nasalization, and digraphs, causing forms that correspond to the same underlying units to fragment across embedding space and weaken retrieval reliability.

To mitigate this effect, we apply a canonicalization procedure that standardizes common orthographic variants without introducing linguistic analysis. Specifically, we perform:

- unification of diacritics for acute vowels and nasalization;
- conversion of digraphs into precomposed characters when available;
- replacement of unambiguous ASCII approximations (e.g., "sh" $\rightarrow$ š, "ng" $\rightarrow$ ŋ);
- consistent casing across sources;
- removal of punctuation artifacts prior to tokenization.

We deliberately avoid operations that alter linguistic structure, such as lemmatization or morphological segmentation. While normalization improves lexical alignment and retrieval consistency, it does not resolve downstream decoding instability on its own; subsequent stages of the cascade further filter misleading evidence based on the model's generative behavior.

3.2.3 Multi-Channel Retrieval

Retrieval is used to assemble two complementary forms of evidence: lexical grounding and structural exemplars. Because the LLM has minimal intrinsic knowledge of Lakota vocabulary, lexical coverage must be provided explicitly, while structural examples serve to constrain morphosyntactic interpretation.

Lexical evidence is drawn solely from dictionary-style entries and word–gloss pairs whose normalized Lakota forms match tokens or subword segments in the input. These entries provide direct lexical grounding and are treated as mandatory when available. Lexical entries are treated as high-precision grounding anchors and are included ex-

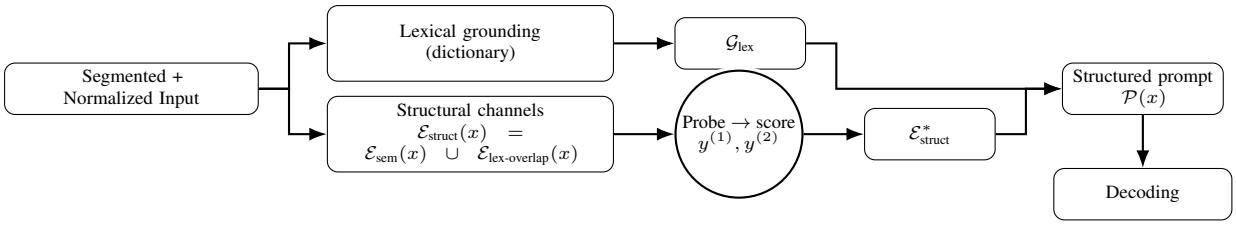

Figure 1: CART pipeline with stability-aware selection. Structural candidates from semantic and lexical-overlap retrieval are probed using instruction-level paraphrases, scored by normalized character edit distance between probe outputs, and filtered to the top-m most stable exemplars. Lexical grounding entries $\mathcal{G}_{\text{lex}}$ are included exhaustively and bypass stability filtering. The final structured prompt $\mathcal{P}(x)$ combines $\mathcal{G}_{\text{lex}}$ with $\mathcal{E}^*_{\text{struct}}$ (along with the instruction and source sentence) for deterministic decoding.

haustively when available. Although they condition generation, they primarily serve to supply vocabulary that the model lacks intrinsically. Stability-aware filtering is therefore applied only to structural exemplars, whose interaction with the generator can induce instability. Lexical grounding entries do not refer to the lexical-overlap retrieval channel as the grounding entries operate at the word level, are included exhaustively, and are not subject to stability-aware filtering; whereas, the lexical-overlap channel retrieves sentence- or phrase-level structural exemplars that are later combined with the output of the semantic channel and filtered based on their effect on decoding behavior.

Structural evidence is drawn from sentence- and phrase-level parallel fragments, glossed examples, and constructional patterns. These items are intended to illustrate morphosyntactic and derivational patterns rather than to provide exhaustive lexical coverage. We encode the normalized input sentence and all reference items using a frozen multilingual sentence encoder (LaBSE) and perform approximate nearest-neighbor search with cosine similarity over a FAISS index. This semantic channel provides high recall but may surface structurally incompatible examples in low-resource settings.

To complement embedding-based retrieval, we additionally include a lightweight lexical overlap channel that surfaces structurally relevant examples sharing stems or fixed expressions after normalization. This channel improves recall for related constructions but does not guarantee compatibility.

The retrieved structural candidates form a high-recall set $\mathcal{E}_{\text{struct}}(x)$ that may contain both helpful and misleading examples. No attempt is made at this stage to assess correctness or compatibility. Instead, structural evidence quality is evaluated in the subsequent stability-aware selection step, which explicitly measures how candidate examples affect the model's decoding behavior.

3.3 Stability-Aware Evidence Selection

The retrieved candidate set $\mathcal{E}_{\text{struct}}(x)$ produced by multi-channel retrieval is intentionally permissive and may contain both supportive and misleading examples. In morphologically rich low-resource settings, similarity-based retrieval alone is insufficient to guarantee that conditioning on a given example will stabilize generation. We therefore introduce a generator-aware evidence selection mechanism that evaluates retrieved candidates based on their effect on the model's decoding behavior. This mechanism operates entirely at inference time and requires no additional training or language-specific analyzers.

3.3.1 Instability Probe Design

We treat instability as a behavioral property of the generator: an example is considered destabilizing if minimal instruction-level paraphrases to the prompt induce divergent outputs. To test this, we design a minimal instability probe that measures the sensitivity of the model's output to minor instruction-level variation.

For each retrieved structural candidate $e_i \in \mathcal{E}_{\text{struct}}(x)$, we construct two probe prompts that differ only in instruction phrasing while preserving identical content. For example, the two prompts may use alternative but semantically equivalent instructions such as "Translate the following sentence:" and "Provide an English translation of:" followed by the same input and candidate example. No additional context or information is introduced.

Each probe prompt is decoded deterministically using greedy decoding with temperature 0 and top-p set to 1, producing two short outputs:

$$y_i^{(1)},\ y_i^{(2)}.$$

Decoding is capped at a fixed maximum length of N tokens to ensure consistent and comparable outputs. Because decoding is deterministic, any divergence between $y_i^{(1)}$ and $y_i^{(2)}$ reflects sensitivity to the prompt perturbation rather than sampling noise. We focus on instruction-level paraphrases rather than input-level noise because they constitute a semantically invariant perturbation that preserves both the source sentence and retrieved evidence. Perturbations to the input or evidence content would confound decoding sensitivity with genuine ambiguity or information loss. Instruction phrasing varies naturally across prompting interfaces and deployment contexts, making it a practical and controlled axis along which to probe behavioral instability induced by retrieved examples.

3.3.2 Instability Metric

We quantify instability using normalized character-level edit distance between the two probe outputs. Character-level (rather than token-level) distance avoids tokenizer artifacts and better captures fine-grained morphological variation. Let $d(\cdot, \cdot)$ denote Levenshtein distance and let $|\cdot|$ denote character length. The instability score for candidate e_i is defined as:

$$D(e_i) = \frac{d\Big(y_i^{(1)}, y_i^{(2)}\Big)}{\max\Big(|y_i^{(1)}|, |y_i^{(2)}|\Big)}.$$

This metric captures surface-level divergence between the two outputs, which is particularly sensitive to morphological variation, stem substitution, and affixal changes common in Lakota translations. Normalization by output length ensures comparability across candidates of different lengths.

Low values of $D(e_i)$ indicate that the candidate induces consistent decoding behavior under minimal perturbation, whereas high values indicate instability. Importantly, instability is used strictly as a negative signal to identify potentially misleading evidence; it is not treated as a proxy for correctness, adequacy, or semantic quality.

3.3.3 Evidence Ranking and Filtering

For each input sentence x, we compute instability scores $D(e_i)$ for all retrieved structural candidates. Candidates are ranked in ascending order of instability:

$$e_{(1)}, e_{(2)}, \ldots, e_{(|\mathcal{E}_{\text{struct}}(x)|)},$$

where $D(e_{(1)}) \leq D(e_{(2)}) \leq \cdots$.

We retain the top m most stable candidates and discard the remainder:

$$\mathcal{E}^*_{\text{struct}}(x) = \{e_{(1)}, \ldots, e_{(m)}\}.$$

Unless otherwise noted, we fix $m = 5$ across all experiments. Threshold-based filtering is a natural alternative but is not explored in this work.

3.3.4 Computational Cost

Stability-aware evidence selection requires additional inference-time computation in the form of short probe decodes. For each input sentence, two short deterministic decodes are performed for each retrieved structural candidate in $\mathcal{E}_{\text{struct}}(x)$. The resulting overhead therefore scales linearly with the retrieval depth $k = |\mathcal{E}_{\text{struct}}(x)|$.

In our experiments, we retrieve $k = 20$ structural candidates and generate probe outputs capped at $N = 30$ tokens, resulting in at most $2kN = 1200$ probe tokens per input. This cost is small relative to full-sequence decoding and dominates neither retrieval nor final generation.

Probe decodes are short, bounded, and independent across candidates, enabling efficient batching or parallelization when resources permit. No high additional cost is incurred at decoding time beyond standard prompting.

3.4 Prompt Construction and Decoding

Prompt construction integrates lexical grounding $\mathcal{G}_{lex}$ and stability-filtered structural evidence $\mathcal{E}^*_{\text{struct}}$ into a compact context that conditions final generation. Lexical and structural evidence serve distinct roles and are presented separately to the model.

Lexical entries matching the input are included as a glossary-style component that provides direct word-level grounding. These entries are included exhaustively when available and are not ranked or filtered by stability, ensuring lexical coverage even when the model lacks prior knowledge of the language.

Structural evidence is drawn from the filtered set $\mathcal{E}^*_{\text{struct}}(x)$ produced by stability-aware selection. A small number of structurally compatible examples are included as parallel fragments or glossed pairs and ordered by increasing instability score, biasing the model toward the most stable evidence while preserving limited exposure to alternative realizations. Structural examples are intended to constrain derivational and argument-structure interpretation rather than to encode explicit rules.

Prompt content is fixed across inputs and does not depend on test-time tuning. Where appropriate, we include brief, generic contextual cues informed by common error patterns observed during development (e.g., stem repetition or misinterpretation of kinship terms). These cues apply uniformly and do not encode language-specific rules.

Decoding is deliberately conservative. Because destabilizing structural evidence has already been removed, the decoder's role is to select a consistent realization rather than explore alternative continuations. We therefore use greedy decoding with temperature set to 0 and top-p set to 1. In preliminary experiments, sampling-based decoding introduced increased semantic drift and derivational variation without clear gains in adequacy.

Post-processing is minimal. Generated outputs undergo only light normalization or formatting adjustments when required for presentation. No rule-based correction or reranking is applied after generation, preserving transparency in the end-to-end pipeline.

4 Experiments

We evaluate CART in a memory-based translation setting designed for low-resource conditions, where no parametric training is performed and all linguistic evidence is supplied at inference time. Experiments focus on isolating the effect of stability-aware evidence selection while controlling for retrieval, prompting, and decoding.

4.1 Data

Our experiments use a heterogeneous Lakota–English parallel reference pool containing approximately 70,000 bilingual items. The data consist of common words and short phrases (∼34k), narrative and verse-style materials drawn from publicly accessible sources (∼18k), and dictionary entries, elicited examples, and pedagogical resources (∼21k). All data are openly available or permissively licensed for non-commercial research use.

Because no large curated parallel corpus exists for Lakota, these materials are consolidated into a single heterogeneous reference pool that serves as retrievable memory for inference-time conditioning, including lexical grounding entries and sentence- or phrase-level exemplars. To prevent evaluation leakage, reference items with close surface overlap to test examples were removed through manual screening.

We construct a held-out test set of 620 Lakota–English pairs drawn from unrelated sources not included in the reference pool. Test items include everyday expressions, short narrative passages, and culturally grounded terms that are challenging for general-purpose language models. Reference data are retrievable at inference time, while test items are not.

4.2 Models and Retrieval Setup

All experiments use frozen models. For retrieval, we employ LaBSE to obtain dense representations of normalized inputs and reference items. Embeddings are indexed using FAISS IVF-Flat for approximate nearest-neighbor search with cosine similarity. For each input, we retrieve the top $k = 20$ candidates, which serve as input to the stability-aware selection procedure described in Section 3.3.

For generation, we use `LLaMA-3-8B-Instruct` as a representative, publicly available instruction-tuned LLM that is large enough to support in-context reasoning while remaining practical for controlled inference-time experiments. No finetuning or parameter updates are performed, and all systems use identical decoding settings for comparability. Unless otherwise noted, we retain the top $m = 5$ most stable examples for prompting, and instability probes generate at most $N = 30$ tokens using deterministic decoding; all hyperparameters are fixed across experiments and are not tuned on the test set.

4.3 Baselines

We compare CART against prompting-based baselines and ablated variants that isolate the contribution of stability-aware evidence selection. All systems use the same underlying LLM and decoding configuration.

Direct Prompting. The model receives only a translation instruction and the source sentence, with no retrieved evidence.

Similarity-Based Retrieval. Retrieved examples are included in the prompt based solely on similarity ranking, without stability-aware filtering.

CART (Full). Our full system applies normalization, multi-channel retrieval, stability-aware evidence selection, and structured prompting prior to deterministic decoding.

We do not compare against pretrained multilingual MT systems (e.g., mBART, M2M-100,

Table 1: Automatic evaluation on the held-out test set. "CART w/o stability selection" removes the stability-aware evidence filtering step (Section 3.3) while keeping retrieval and prompting fixed. ΔBLEU is measured relative to Similarity-Based Retrieval.

	EN→LKT				*LKT→EN*			
System	**BLEU**	**chrF**	**BERTScore**	**ΔBLEU**	**BLEU**	**chrF**	**BERTScore**	**ΔBLEU**
Direct Prompting	3.9	25.6	0.84	–	10.8	41.2	0.89	–
Similarity-Based Retrieval	6.8	30.1	0.88	–	12.6	43.6	0.91	–
CART w/o stability selection	7.4	30.6	0.89	+0.6	13.0	43.9	0.91	+0.4
CART (Full)	**9.7**	**33.0**	**0.91**	**+2.9**	**14.4**	**45.1**	**0.92**	**+1.8**

NLLB), as Lakota is absent from their training data and zero-shot application produces incoherent or degenerate output in preliminary experiments (Costa-jussà et al., 2022). Prompting-based baselines therefore provide the most meaningful comparison in this setting.

4.4 Evaluation Metrics

We report BLEU, chrF, and BERTScore for comparability with prior MT literature. We compute BERTScore using the `xlm-roberta-large` model with IDF weighting and default layer selection, following the standard multilingual configuration provided in the official reference implementation. Given Lakota's rich morphology and flexible word order, we interpret surface-level metrics cautiously and rely more heavily on semantic similarity and complement them with qualitative analysis and stability-focused evaluations in Section 6.

5 Results

Table 1 reports automatic evaluation results on the held-out test set for both translation directions. We compare direct prompting, similarity-based retrieval, and CART, along with an ablated variant that removes stability-aware evidence selection while keeping retrieval and prompting fixed.

Across all metrics and both directions, CART consistently outperforms the baselines. Similarity-based retrieval yields substantial improvements over direct prompting, confirming that access to external evidence is critical in this low-resource setting. However, retrieval alone is insufficient: removing stability-aware selection results in markedly smaller gains, particularly for EN→LKT translation. Incorporating stability-aware evidence filtering yields an additional improvement of +2.9 BLEU over similarity-based retrieval in the EN→LKT direction and +1.8 BLEU in the reverse direction.

Improvements are larger for EN→LKT than for LKT→EN, reflecting the greater difficulty of morphologically grounded generation when translating into Lakota. Gains are consistent across BLEU, chrF, and BERTScore, with BERTScore exhibiting steady increases that suggest reduced semantic divergence. Although absolute BLEU values remain modest, they are consistent with prior observations for severely low-resource, morphologically rich languages evaluated with single-reference metrics.

Overall, these results demonstrate that generator-aware evidence selection provides benefits beyond standard similarity-based retrieval, yielding more robust and consistent translation behavior without modifying model parameters or training data.

6 Analysis and Discussion

This section analyzes how stability-aware evidence selection influences model behavior and explains the empirical trends observed in Section 5. We focus on (i) the contribution of stability-aware filtering, (ii) its effect on decoding robustness, and (iii) qualitative differences in error behavior across systems.

6.1 Stability-Aware Evidence Selection

The ablated variant "CART w/o stability selection" isolates the effect of generator-aware evidence filtering while keeping normalization, retrieval, and prompting fixed. As shown in Table 1, removing stability-aware selection leads to a substantial performance drop, particularly for EN→LKT translation, where BLEU decreases by 2.3 points relative to the full system. The effect is smaller but consistent in the reverse direction.

This asymmetry reflects the differing demands of the two translation directions. When translating into a morphologically richer target language such as Lakota, retrieved examples must better support grounded generation, including stem choice

Table 2: Instruction-level robustness measured as normalized character-level edit distance between final translations produced under two instruction paraphrases. Lower values indicate greater robustness.

	Instability (↓)	
System	*EN→LKT*	*LKT→EN*
Similarity-Based Retrieval	0.41	0.29
CART w/o stability selection	0.36	0.27
CART (Full)	**0.26**	**0.23**

and derivational structure that are absent from the English source. Structurally incompatible examples can therefore destabilize decoding even more prominently. Stability-aware filtering thus better helps mitigate this effect by discarding retrieved evidence that induces divergent behavior under minimal prompt perturbations.

Stability is used strictly as a negative signal: examples that destabilize generation are removed, but stability itself is not treated as a proxy for correctness. The observed gains thus arise from eliminating misleading evidence rather than reinforcing the model's preferred output.

6.2 Instruction-Level Robustness

To quantify robustness directly, we measure the normalized character-level edit distance between final translations produced under the two probe instruction variants. Compared to similarity-based retrieval, CART reduces average instruction-level divergence by 37% for EN→LKT and 21% for LKT→EN, indicating that stability-aware filtering narrows the model's effective hypothesis space.

6.3 Decoding Robustness

Stability-aware selection reduces sensitivity to prompt phrasing by filtering evidence that induces inconsistent short-form outputs. Compared to similarity-based retrieval alone, the full system produces more consistent translations under controlled perturbations, indicating a narrower and more stable hypothesis space.

While instability is not optimized directly, its reduction aligns with improvements in translation quality metrics. Character-level divergence is used as a diagnostic signal to accommodate Lakota's rich morphology, where small orthographic differences may correspond to meaningful morphemic variation. These results support the view that controlling behavioral instability is an effective mechanism for preventing misleading evidence from influencing generation in low-resource settings.

6.4 Qualitative Behavior

Qualitative inspection reveals clear differences between systems. In EN→LKT translation, the full system more reliably selects appropriate derivational forms and avoids overproducing marked morphology. Similarity-based retrieval often surfaces semantically related but structurally incompatible examples, leading to inconsistent stem selection or argument realization.

In LKT→EN translation, improvements primarily reflect more stable sense disambiguation and recovery of participant roles encoded in Lakota verbal morphology. Retrieval-only prompting reduces hallucination but remains sensitive to example mismatch, whereas stability-aware filtering yields more consistent interpretations.

A small round-trip diagnostic further illustrates this effect. English inputs translated into Lakota and then back into English using the same system exhibit substantially less semantic drift under the full system than under prompting-only or similarity-based retrieval. While not a formal evaluation, this diagnostic highlights differences in robustness under compounding uncertainty.

7 Conclusion

We presented CART, an inference-time retrieval–prompting cascade for Lakota–English translation that addresses a key failure mode of LLM-based MT in low-resource settings: unstable decoding under weak lexical and structural grounding. The central contribution of this work is stability-aware evidence selection, which filters retrieved examples based on their effect on the generator's behavior rather than similarity alone.

Our results demonstrate that, in morphologically rich low-resource languages, how evidence is selected can matter more than how much evidence is retrieved. While Lakota serves as a focused case study, the approach is applicable to other settings where parallel data are scarce and example mismatch can actively degrade generation. More broadly, this work suggests a practical path toward MT systems that support Indigenous and endangered languages by leveraging existing lexical resources and community-curated examples to guide modern language models in a controlled and transparent manner.

Limitations

The proposed approach relies on a retrieval memory whose coverage is constrained by the availability of Lakota linguistic resources. Although the system can incorporate heterogeneous materials, including dictionary entries and example translations, many constructions, derivational patterns, and idiomatic expressions remain sparsely represented. As a result, the system may struggle with rare or highly complex forms that are absent from the reference pool.

The approach further assumes access to lexically comprehensive word-level resources. Because large language models exhibit limited intrinsic knowledge of Lakota, lexical grounding is primarily provided by dictionary-style entries rather than by the model's pretrained representations. Stability-aware selection operates on structural examples and does not address lexical coverage, which remains dependent on external resources.

Evaluation is also constrained by the lack of standardized MT benchmarks for Lakota. Reference translations often permit multiple valid realizations, and surface-oriented metrics such as BLEU and chrF therefore capture only coarse trends. While we complement these metrics with qualitative analysis and stability-focused diagnostics, the study does not include targeted evaluation of specific morphological categories or controlled human judgments.

Finally, stability-aware evidence selection introduces additional inference-time cost due to short probing decodes and remains a heuristic mechanism. While it reduces the influence of misleading evidence, it does not guarantee correctness and cannot resolve ambiguity that is genuinely underspecified by the input. Although Lakota serves as a representative case study of morphologically rich, low-resource languages, further empirical evaluation is needed to assess the generality of the approach across typologically diverse settings and larger models.

Acknowledgements

We thank the Lakota speakers and community members who provided guidance, corrections, and valuable discussion throughout the development of this work. Their insights into usage, morphosyntax, and variation were essential in shaping the design and interpretation of system outputs. We also acknowledge the publicly accessible dictionaries, educational materials, and linguistic descriptions that made this project possible.

We also acknowledge the Wordspring Initiative, a nonprofit effort aimed at developing accessible language technologies for Indigenous communities, within which an early prototype of this system was deployed for community feedback.

References

David Adelani, Jesujoba Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Rijhwani, Winston Lewis, Vukosi Marivate, Sebastian Osei, and 1 others. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maximilian Gouws, Aditya Siddhant, Alexandre Conneau, Wei-Jen Wu, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Romadhony, Rahmad Mahendra, Fajri Thoufeq Z. Adipurna, and 1 others. 2024. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning. *arXiv preprint arXiv:2305.13627*.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2023. Improving neural machine translation of indigenous

languages with multilingual transfer learning. *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 73–85.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Liu, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Samuel Frontull and Thomas Ströhle. 2025. Compensating for data with reasoning: Low-resource machine translation with llms. *arXiv preprint arXiv:2505.22293*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Viktor Hangya, Silvia Severini, Radoslav Ralev, Alexander Fraser, and Hinrich Schütze. 2023. Multilingual word embeddings for low-resource languages using anchors and a chain of related languages. *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 85–99.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Surafel Melaku Lakew, Vishrav Chaudhary, and Marcello Federico. 2020. Neural machine translation for extremely low-resource african languages: A case study on bambara. *Proceedings of the 3rd Workshop on African Natural Language Processing*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the americas: An introduction. *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 159–176.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented llm prompting: A study on the mambai language. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): 1st Workshop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TELA 2024)*, pages 1–9.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 215–228.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Surangika Ranathunga, Nisansa de Silva, and Marcos Zampieri. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

R. Ranjith, Vijay Kumar Menon, and Sreelakshmi Joseph. 2023. Neural machine translation: A survey of methods used for low resource languages. *IEEE Access*, 11:52341–52360.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.