Evaluating Frontier LLM Translation Capability for Lakota

Lance Robertson
University of California, San Diego
lrobertson@ucsd.edu

Abstract

We evaluate seven large language models—four proprietary and three open-weight—on bidirectional Lakota–English translation using 200 sentence pairs from the New Lakota Dictionary. Each model is evaluated with and without extended reasoning, where the provider's API permits. The best model (Gemini 3.1 Pro) achieves a mean chrF++ of 59.4 on Lakota→English and 42.6 on English→Lakota; the strongest open-weight model trails the proprietary leaders, and no model produces reliable translation in either direction. Two independent LLM judges from different model families agree substantially (Cohen's $\kappa = 0.75$) that semantic equivalence ranges from 6% (GPT-5.2) to 60% (Gemini), diverging substantially from chrF++ scores. For the open-weight models, enabling reasoning changes refusal behavior far more than translation quality: it surfaces the limitation rather than overcoming it. Diacritic-normalization analysis shows models produce roughly correct base characters but place diacritical marks inconsistently. All results and evaluation code are publicly available at https://github.com/robotson/lakota-translation-benchmark.

1 Introduction

Lakota (*Lakȟótiyapi*) is a Siouan language spoken primarily on reservations in North and South Dakota. UNESCO classifies it as critically endangered (Moseley, 2010), with fewer than 2,000 first-language speakers, most over 65 years of age. Language revitalization efforts are ongoing across multiple tribal and academic institutions. The Lakota Language Consortium (LLC) publishes the New Lakota Dictionary (NLD) and associated pedagogical materials; the NLD serves as the source for this study's evaluation set.

Contemporary large language models (LLMs) support translation across dozens of languages, but the extent of training data for any given language is not disclosed. Users may reasonably expect similar quality across supported languages. For critically endangered languages like Lakota—where digitized text is scarce and no large-scale parallel corpus exists—this expectation is untestable without direct evaluation.

This paper reports the results of a direct evaluation. We tested seven LLMs—four proprietary and three open-weight—on 200 Lakota–English sentence pairs in both translation directions, under two experimental conditions designed to test the impact of extended thinking (chain-of-thought reasoning) on translation quality. We report chrF++ and BLEU scores, diacritic normalization analysis, model self-reported confidence, and refusal behavior. Related work on LLM translation of low-resource languages includes the MTOB benchmark (Tanzer et al., 2024), which evaluated frontier LLMs on Kalamang translation using in-context learning from a reference grammar, and the AmericasNLP shared tasks (Mager et al., 2021; Ebrahimi et al., 2023), which benchmark MT systems on indigenous American languages but do not include Lakota. The only prior Lakota-specific language model is LakotaBERT (Parankusham et al., 2025), a RoBERTa-based model trained for masked language modeling. To our knowledge, no prior study has evaluated zero-shot frontier LLM translation quality for Lakota.

Lakota presents specific challenges for LLM translation. It is polysynthetic, with extensive verbal morphology including person mark-

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 11–21

ing, aspect, mood, and evidentiality encoded through affixes and enclitics. The Standard Lakota Orthography (SLO) uses Latin script with diacritical marks: caron-marked consonants (č, š, ž, ǧ, ȟ), ogoneks on vowels for nasalization (ą, į, ų), acute accent for stress, and the glottal stop marker (’). The velar nasal is written with a separate character, ŋ. These diacritical marks distinguish phonemes—for example, *h* and *ȟ* represent different consonants—making them significant for both translation accuracy and character-level evaluation metrics.

2 Methodology

2.1 Evaluation Set

The evaluation set consists of 200 sentence pairs sourced from published Lakota learning materials for research purposes. Pairs were drawn from conversational examples in the New Lakota Dictionary (NLD) (Ullrich, 2008), published by the Lakota Language Consortium (LLC), and selected from entries appearing under dictionary headwords. The first author filtered pairs for naturalness and conversational relevance. Each pair contains one Lakota sentence and one English translation. Examples range from short questions (*Hé yačhíŋ he?* / “Do you want it?”) to multi-clause sentences with cultural context (*Wóčhičiyakiŋ kte ȟčiŋ.* / “I want to talk to you.”). The evaluation set is conversational register throughout.

The NLD uses the Standard Lakota Orthography (SLO), developed by the LLC. Multiple orthographic conventions have been in use across Lakota language communities, including the Buechel missionary system and the orthography used at Sinte Gleska University (White Hat, 1999). Standardization of Lakota orthography remains a contested issue within Lakota communities (Hauff, 2020). Our use of NLD materials was practical—it is the most accessible published collection of Lakota sentence pairs with English translations available to us—and should not be read as an endorsement of any particular orthographic standard or institutional approach to language documentation.

Each pair includes a conversational score (4–8) assigned during curation, where higher values indicate more everyday, conversational language and lower values indicate more formal or structurally complex sentences. The evaluation set also records the headword context from which each pair was drawn. The same 200 pairs were used for all models and both experimental conditions.

2.2 Models

Seven models were evaluated. Four are proprietary frontier models from three providers: Gemini 3.1 Pro Preview (Google), Claude Opus 4.6 (Anthropic), Claude Sonnet 4.6 (Anthropic), and GPT-5.2 (OpenAI). Three are open-weight frontier models, accessed through Together AI: DeepSeek-V4-Pro, Qwen3.6-Plus, and GLM-5.1. All models were accessed through litellm (v1.63), a unified API wrapper; the open-weight models were routed through an OpenAI-compatible endpoint to preserve the same structured-output contract. Selection criteria: frontier-tier models supporting structured JSON output with schema enforcement. Including open-weight models tests whether the closed–open capability gap observed on higher-resource tasks holds for Lakota.

2.3 Experimental Conditions

Two conditions were run on the same evaluation set:

Baseline: Temperature 0, no extended thinking. Maximum output tokens: 1,024. This condition uses near-deterministic decoding and represents baseline model capability.

Thinking: Extended thinking enabled at maximum provider settings. Maximum output tokens: 16,384 (to accommodate thinking token overhead). Claude models used `thinking.budget_tokens` = 8,192; GPT-5.2 used `reasoning_effort` = high; Gemini used `thinkingLevel` = high.

Gemini 3.1 Pro enables thinking by default and cannot disable it, so its baseline already includes default-level thinking; the baseline→thinking comparison isolates only the increase from default to high, not the enabling of thinking.

For Claude models, enabling thinking fixes temperature at 1 (it cannot be set independently while thinking is on); for

GPT-5.2, high-reasoning configurations ignore sampling parameters, so `temperature=1` likely had no effect. These proprietary baseline→thinking comparisons should therefore be read as "baseline vs. maximum supported reasoning settings" rather than clean temperature-controlled ablations. The three open-weight models (served via Together AI) instead toggle reasoning directly: thinking uses the provider's default reasoning and baseline disables it via the documented control—`reasoning.enabled` for DeepSeek, `chat_template_kwargs.enable_thinking` for Qwen and GLM—holding each model's output-format regime fixed, a cleaner within-model ablation.

2.4 Structured Output Schema

All models produced structured JSON output enforced via provider-native schema validation with three fields: `translation` (best translation or empty string), `confidence` (0.0–1.0), and `refusal_reason` (string or null).

The system prompt was identical across all models and conditions:

> You are a translation system. Translate the input text. Respond as JSON with three fields: "translation" (your best translation, or empty string if you cannot translate), "confidence" (a number from 0.0 to 1.0 representing your confidence in the translation quality), "refusal_reason" (null if you attempted a translation, or a brief explanation if you cannot translate).

User prompts followed the format "Translate from Lakota to English: {text}" and "Translate from English to Lakota: {text}" for the two directions.

2.5 Metrics

chrF++ (a character n-gram F-score extended with word n-grams) was the primary metric, computed via SacreBLEU (Popović, 2015, 2017) with default parameters: character n-gram order 6, word n-gram order 2, $\beta = 2$. chrF++ is preferred over BLEU for morphologically rich languages because it operates at the character level and does not penalize valid morphological variants as harshly as word-level metrics. **BLEU** was computed as a secondary metric for comparability.

Diacritic-normalized chrF++ was computed for English→Lakota translations by stripping Unicode combining marks (categories Mn and Mc) from both candidate and reference via NFD decomposition before scoring. NFD decomposition ensures canonical equivalence between precomposed characters (e.g., U+01CE, ǎ) and combining sequences (e.g., a + U+030C) before mark removal. This normalization removes carons, ogoneks, and stress marks in SLO; the character ŋ (eng) and the glottal stop marker (') are unaffected. The delta between raw and normalized chrF++ isolates orthographic variation attributable to these diacritical marks from broader lexical and morphological differences.

LLM semantic judgment was used to evaluate Lakota→English translations, where chrF++ is known to penalize valid paraphrases against single references. We initially ran BERTScore (Zhang et al., 2020) (roberta-large) on the proprietary models to capture semantic matches that chrF++ misses, but it was uninformative: F1 ranged from 0.906 to 0.959 regardless of actual translation quality. At the pair level, GPT-5.2's translation of "Do you have a car?" as "Do you have a horse?" received BERTScore 0.975. We did not extend the BERTScore pass to the open-weight models added later, having moved to LLM-based semantic judgment as our primary check. We instead used a separate LLM (Gemini 3 Flash Preview) to judge whether each hypothesis–reference pair conveys the same meaning, following the GEMBA framework (Kocmi and Federmann, 2023) but simplified to English-to-English semantic equivalence. Because both strings are English, the judge requires no knowledge of Lakota—it evaluates only whether the model's English output preserves the meaning of the English reference. Each pair was rated `equivalent`, `partially_equivalent`, or `not_equivalent`. All Lakota→English translation attempts across 14 model×condition combinations were judged. To check for family-level self-judging bias, every pair was also judged by Llama-3.3-70B-Instruct-Turbo, an open-weight model from a developer not among those evaluated; agreement is reported in §3.6.

Self-reported confidence (0.0–1.0) was extracted from the structured output.

Model	L→E chrF++ (σ)	E→L chrF++ (σ)
Gemini 3.1 Pro	63.1 (3.3)	33.6 (3.4)
Claude Opus 4.6	45.1 (2.6)	29.8 (4.8)
Claude Sonnet 4.6	42.7 (4.8)	20.8 (4.6)
GPT-5.2	25.5 (5.7)	15.9 (5.8)

Table 1: Pilot variance: mean chrF++ and cross-run standard deviation across 3 runs (20 pairs), four proprietary models (February 2026). The open-weight models, added in May 2026, were not separately re-piloted.

2.6 Variance

A pilot study was conducted before the full thinking evaluation to assess run-to-run variance under the thinking condition's decoding settings. Twenty sentence pairs were sampled and each of the four proprietary models was run 3 times on both directions (480 API calls). Cross-run standard deviation ranged from 2.6 (Opus L→E) to 5.8 (GPT-5.2 E→L). Model rankings were stable across all three runs (Table 1). Based on these results, the full evaluation was conducted as a single run per condition.

2.7 Data Contamination

A central concern with LLM-based evaluation on low-resource pairs is that the evaluation set may appear in pretraining data. The evaluation set here is drawn from the New Lakota Dictionary (Ullrich, 2008), whose contents are findable on the open web in OCR'd reproductions of varying quality; any web-scale training corpus could plausibly include some form of this material.

To bound the empirical risk we searched indexed open-web content for all 200 pairs. The Lakota source string appears verbatim for about 10% of pairs, but the Lakota sentence and its English reference *together* for only 0.5%—almost always in scanned copies of the dictionary itself. This distinction matters: controlled studies find that aligned source–target overlap can substantially inflate translation scores, while source- or target-only overlap produces smaller, less consistent inflation, at the 1B–8B scales tested (Kocyigit et al., 2025). The overlap we observe is almost all the one-sided kind.

The score distribution points the same way: a memorized test set would cluster near ceiling, whereas ours ranges widely (6–61% semantic equivalence across the seven models), consistent with genuine partial capability rather than wholesale recall—though, as §4 notes, the most common phrases are likely memorized.

2.8 Procedure

Each evaluation run called the model API once per (pair, direction) combination: 200 pairs × 2 directions = 400 calls per model, 1,600 calls per condition. A 1.5-second delay was inserted between calls. API timeout was set to 60 seconds for baseline and 180 seconds for thinking. Failed calls were retried up to 3 times with exponential backoff. Proprietary models were evaluated in February–March 2026 and open-weight models in May 2026; exact model identifiers (API versions current on those dates) are listed in the released code. Results were written to per-model JSONL files with per-pair metadata.

3 Results

3.1 Translation Quality

Table 2 reports the best-condition results for each model. For Gemini, baseline and thinking scores are nearly identical (§3.2); thinking is reported here for consistency since it includes confidence data. For all other models, thinking is reported as the higher-scoring condition.

Gemini leads in both directions. Opus and Sonnet score within 2 chrF++ points of each other on L→E (Sonnet slightly ahead), but Opus leads Sonnet by 7 points on E→L. GPT-5.2 trails the proprietary models in both directions. Among the open-weight models, DeepSeek-V4-Pro is strongest (38.0 / 29.2 chrF++), placing between GPT-5.2 and the Claude models, while GLM-5.1 and Qwen3.6-Plus score at or below GPT-5.2; no open-weight model approaches Gemini. Every model scores substantially higher on Lakota→English than English→Lakota. The gap ranges from 6.4 points (GPT-5.2) to 19.1 points (Sonnet). This asymmetry is consistent across all models: Lakota→English (producing English from Lakota input) scores higher than English→Lakota (producing Lakota from

Model	Dir	N	chrF++	σ (pairs)	Median	BLEU
Gemini 3.1 Pro	L→E	199	**59.4**	27.6	56.8	43.5
Claude Opus 4.6	L→E	200	45.9	25.0	43.5	27.2
Claude Sonnet 4.6	L→E	189	46.1	25.9	44.3	28.1
GPT-5.2	L→E	194	30.0	21.8	21.4	15.4
DeepSeek-V4-Pro	L→E	200	38.0	26.0	30.9	23.3
GLM-5.1	L→E	184	29.3	18.8	24.5	14.2
Qwen3.6-Plus	L→E	198	26.1	21.9	18.2	14.4
Gemini 3.1 Pro	E→L	199	**42.6**	22.1	38.9	24.7
Claude Opus 4.6	E→L	200	34.3	16.0	31.8	18.8
Claude Sonnet 4.6	E→L	196	27.0	13.2	24.2	15.7
GPT-5.2	E→L	184	23.6	12.5	20.5	14.3
DeepSeek-V4-Pro	E→L	195	29.2	14.7	26.9	17.0
GLM-5.1	E→L	193	14.6	7.2	12.8	9.3
Qwen3.6-Plus	E→L	164	18.4	8.7	17.2	13.3

Table 2: Translation quality by model and direction. Proprietary models are shown in their best condition; open-weight models (DeepSeek, GLM, Qwen) in their reasoning-enabled condition. N $<$ 200 reflects refusals, empty responses, or API errors excluded from scoring (§3.5). σ is across sentence pairs, not across runs.

Model	Dir	≥80	≥60	≥40	P10
Gemini	L→E	24%	47%	72%	21.6
Opus	L→E	12%	25%	56%	15.9
Sonnet	L→E	12%	25%	55%	16.0
GPT-5.2	L→E	5%	10%	23%	10.4
DeepSeek	L→E	8%	22%	36%	11.0
GLM	L→E	3%	7%	20%	11.8
Qwen	L→E	4%	10%	17%	8.2
Gemini	E→L	7%	20%	47%	17.2
Opus	E→L	1%	6%	29%	16.0
Sonnet	E→L	1%	1%	15%	12.9
GPT-5.2	E→L	1%	1%	8%	12.1
DeepSeek	E→L	1%	4%	21%	13.0
GLM	E→L	0%	0%	0%	7.4
Qwen	E→L	0%	0%	2%	8.9

Table 3: Score distribution by chrF++ threshold (thinking condition). Percentages over scored translations; P10 = 10th percentile.

Model	Dir	Sc. 4	Sc. 5	Sc. 6
Gemini	L→E	52.7	58.2	81.9
Opus	L→E	41.6	41.0	65.6
Sonnet	L→E	40.3	45.1	64.1
GPT-5.2	L→E	27.9	23.7	45.7
DeepSeek	L→E	33.0	33.9	59.6
GLM	L→E	27.3	26.0	41.0
Qwen	L→E	23.3	22.1	41.7
Gemini	E→L	37.1	42.7	61.3
Opus	E→L	31.2	32.7	45.6
Sonnet	E→L	24.7	26.2	35.2
GPT-5.2	E→L	22.2	21.5	30.4
DeepSeek	E→L	29.4	26.3	31.0
GLM	E→L	13.0	16.5	16.3
Qwen	E→L	17.3	17.3	21.9

Table 4: Mean chrF++ by conversational score (thinking condition). Score 4 = more formal/complex; score 6 = most conversational. Scores 7–8 excluded (n = 6 combined).

English input) for every model.

Standard deviations are large—22–28 points for L→E and 13–22 points for E→L—indicating that per-pair quality varies widely. Some pairs score 100.0 (exact character match to reference) while others score below 10. Table 3 reports the distribution of scores across chrF++ thresholds and tail behavior (10th percentile).

Table 4 stratifies chrF++ by the evaluation set's conversational score (§2.1). For Gemini L→E, the most conversational pairs (score 6) average 81.9 chrF++ while more complex pairs (score 4) average 52.7—a 29-point gap. The pattern holds across all models and both directions, suggesting that performance is concentrated on high-frequency phrasebook-like items, consistent with the finding that low-resource LLM translation quality tracks test-set coverage (Aycock et al., 2025).

Even the best model (Gemini L→E) produces translations scoring ≥80 chrF++—roughly usable quality—on only 24% of pairs. For E→L, only Gemini exceeds 5% at this threshold. If "reliable" is operationalized as ≥60 chrF++ on a majority of pairs, no model qualifies in either direction.

Model	Dir	Base	Think	Δ
Opus 4.6	L→E	45.0	45.9	+0.9
Opus 4.6	E→L	27.4	34.3	+6.9
Sonnet 4.6	L→E	45.4	46.1	+0.7
Sonnet 4.6	E→L	24.8	27.0	+2.2
GPT-5.2	L→E	26.4	30.0	+3.6
GPT-5.2	E→L	18.3	23.6	+5.3
Gemini	L→E	58.4	59.4	+1.0
Gemini	E→L	40.4	42.6	+2.2
DeepSeek	L→E	40.1	38.0	−2.1
DeepSeek	E→L	26.4	29.2	+2.8
GLM	L→E	27.8	29.3	+1.5
GLM	E→L	13.6	14.6	+1.0
Qwen	L→E	25.2	26.1	+0.9
Qwen	E→L	15.6	18.4	+2.8

Table 5: Baseline→thinking comparison

Model	Raw	Norm.	Δ
Gemini 3.1 Pro	42.6	48.7	+6.1
Claude Opus 4.6	34.3	40.8	+6.5
Claude Sonnet 4.6	27.0	32.3	+5.3
GPT-5.2	23.6	27.2	+3.6
DeepSeek-V4-Pro	29.2	33.6	+4.4
GLM-5.1	14.6	18.1	+3.5
Qwen3.6-Plus	18.4	21.7	+3.3

Table 6: Diacritic normalization, English→Lakota (thinking condition).

3.2 Effect of Extended Thinking

Table 5 compares baseline (near-deterministic decoding, no extended thinking) and thinking (maximum supported reasoning settings) for each model.

Among proprietary models, thinking improved chrF++ across all model-direction pairs, with E→L gains (2.2–6.9 points) consistently larger than L→E gains (0.7–3.6 points)—reasoning helps more on the harder generation direction. For Gemini, whose baseline already includes default-level thinking, the L→E delta is small (+1.0) while E→L still gains (+2.2). The open-weight models show only marginal chrF++ movement (-2.1 to $+2.8$ points); unlike the proprietary frontier, reasoning does little for their translation quality or semantic equivalence (§3.6) and instead changes their refusal behavior (§3.5).

3.3 Diacritic Normalization

Table 6 reports raw and diacritic-normalized chrF++ for English→Lakota (thinking condition).

Stripping diacritics adds 3.3–6.5 chrF++ points for E→L. The gap indicates that models produce roughly correct consonant and vowel sequences but place diacritical marks—stress, nasalization, and caron-marked consonants—inconsistently. Multiple orthographic conventions are in active use for Lakota (§2.1); models may have been trained on text in several of these systems, and the prompt did not specify an orthographic standard.

Model	L→E	E→L
Gemini 3.1 Pro	0.93 (σ = 0.06)	0.83 (σ = 0.11)
Claude Opus 4.6	0.71 (σ = 0.17)	0.46 (σ = 0.15)
Claude Sonnet 4.6	0.70 (σ = 0.15)	0.32 (σ = 0.11)
GPT-5.2	0.55 (σ = 0.12)	0.45 (σ = 0.11)
DeepSeek-V4-Pro	0.84 (σ = 0.10)	0.68 (σ = 0.13)
GLM-5.1	0.63 (σ = 0.18)	0.56 (σ = 0.22)
Qwen3.6-Plus	0.78 (σ = 0.16)	0.59 (σ = 0.20)

Table 7: Self-reported confidence (thinking condition).

3.4 Confidence

Table 7 reports mean self-reported confidence from the thinking condition.

Among the proprietary models, the confidence ranking largely matches the chrF++ ranking for L→E: Gemini > Opus ≈ Sonnet > GPT-5.2. The open-weight models are the most overconfident: DeepSeek-V4-Pro and Qwen3.6-Plus report L→E confidence of 0.84 and 0.78—above every proprietary model except Gemini—despite scoring well below the Claude models. All models report lower confidence for E→L than L→E.

Gemini reports the highest confidence in both directions (0.93 L→E, 0.83 E→L) and also achieves the highest chrF++ scores, so its confidence is directionally accurate. However, 0.83 confidence on E→L where chrF++ is 42.6 represents significant overconfidence in absolute terms.

GPT-5.2's E→L confidence (0.45) exceeds Sonnet's (0.32) despite scoring lower on chrF++ (23.6 vs 27.0), so E→L confidence does not track quality.

At the per-pair level, confidence is a poor guide to correctness, especially for the open-weight models without reasoning: Qwen3.6-Plus renders *Hamáčhola* ("I am naked") as "Spider" with 0.95 confidence, and the three open-weight models return three different

	L→E		E→L	
Model	**Ref**	**Emp**	**Ref**	**Emp**
GPT-5.2	6	0	16	0
Sonnet 4.6	0	11	2	2
Opus 4.6	0	0	0	0
Gemini	1	0	0	1
DeepSeek-V4-Pro	0	0	5	0
GLM-5.1	16	0	7	0
Qwen3.6-Plus	2	0	36	0

Table 8: Non-translation outcomes (thinking condition). Ref = refusal, Emp = empty response, Err = API error.

confident answers (0.8–0.95) for *Wakȟálya yo* ("Make coffee"). High confidence attached to output that does not preserve meaning is common rather than exceptional, so self-reported confidence cannot be used to filter unreliable translations.

3.5 Refusals and Failures

Table 8 reports non-translation outcomes from the thinking condition.

Among the proprietary models, GPT-5.2 refused the most translations (22 total, 16 on E→L, up from 4 at baseline); among the open-weight models, Qwen3.6-Plus refused far more (38 total, 36 on E→L). Refusal reasons typically cited inability to produce reliable Lakota text. Sonnet produced 11 empty responses on L→E (the model returned an empty translation string with no refusal reason) and refused 2 E→L translations. Opus produced complete translations on all 200 pairs in both directions. Gemini had one L→E refusal and one E→L empty response.

The open-weight models make the role of reasoning in refusal behavior especially clear. With reasoning enabled, Qwen3.6-Plus refuses 36 of 200 E→L pairs and GLM-5.1 refuses 16 of 200 L→E pairs; with reasoning disabled, these fall to 0 and 1 respectively. Disabling reasoning does not improve translation quality (§3.2, §3.6)—it removes the step at which the model recognizes it cannot translate. For these models, reasoning surfaces the limitation rather than overcoming it.

The structured schema lets each model decline with a `refusal_reason`, which could bias models toward refusal over a low-confidence attempt. The non-thinking results argue against this: disabling reasoning makes refusals nearly vanish with no gain in semantic equivalence, so refusals under reasoning track recognized difficulty rather than a prompt artifact.

Among the proprietary models, refusals are concentrated on E→L (16 of 22 GPT refusals, both Sonnet refusals)—the harder generation direction.

Reported scores exclude refusals, the standard conditional-on-attempting convention. Scoring refusals as chrF++ = 0 over all 200 pairs lowers the high-refusal cases (Qwen3.6-Plus E→L from 18.4 to 15.1, GLM-5.1 L→E from 29.3 to 27.0), with no change to the ranking. For the open-weight models it also tempers the apparent thinking gains: Qwen3.6-Plus's E→L thinking score no longer exceeds its baseline once reasoning-induced refusals are counted.

3.6 Semantic Equivalence

Table 9 reports LLM semantic judgments for Lakota→English translations across all seven models and both conditions, rating each hypothesis–reference pair `equivalent`, `partially_equivalent`, or `not_equivalent`. The two judges (Gemini 3 Flash and the independent Llama-3.3-70B) agree substantially (Cohen's $\kappa = 0.75$, 0.89 quadratic-weighted), with disagreement concentrated in the adjacent `partially_equivalent` tier rather than between the polar categories.

The judge reveals that chrF++ is a reasonable proxy for models with higher equivalence rates—Gemini's chrF++ of 58.4 corresponds to 60.4% semantic equivalence—but misleading for models with lower rates. GPT-5.2's chrF++ of 26.4 implies some partial overlap, but only 6% of its translations are semantically correct. The remaining 87% are fluent English that does not preserve the source meaning (Table 10).

Open-weight models do not close the gap. The strongest, DeepSeek-V4-Pro, reaches roughly 19–21% equivalence—between GPT-5.2 and the Claude models—while GLM-5.1 and Qwen3.6-Plus remain at GPT-5.2's baseline tier (7–9%). No open-weight model approaches Gemini.

For the stronger models, 16–24% of translations are *partially equivalent*—capturing the

Model	Cond.	Equiv	Partial	Not	κ
Proprietary					
Gemini	base	60.4	23.4	16.2	0.64
Gemini	think	61.3	23.1	15.6	0.65
Opus 4.6	base	31.5	18.0	50.5	0.72
Opus 4.6	think	37.0	21.0	42.0	0.63
Sonnet 4.6	base	29.0	16.0	55.0	0.69
Sonnet 4.6	think	30.7	23.8	45.5	0.67
GPT-5.2	base	6.0	7.0	87.0	0.63
GPT-5.2	think	13.4	9.8	76.8	0.75
Open-weight					
DeepSeek-V4-Pro	base	20.5	13.0	66.5	0.70
DeepSeek-V4-Pro	think	19.0	14.0	67.0	0.74
GLM-5.1	base	9.0	7.0	83.9	0.68
GLM-5.1	think	9.2	6.5	84.2	0.76
Qwen3.6-Plus	base	7.1	4.5	88.4	0.77
Qwen3.6-Plus	think	7.1	7.6	85.4	0.82

Table 9: LLM semantic judge results (Gemini 3 Flash Preview), L→E; Equiv/Partial/Not are percentages of judged pairs (equivalent / partially equivalent / not equivalent). κ = Cohen's inter-judge agreement with an independent Llama-3.3-70B judge on each row's pairs. `base/think` = reasoning disabled/enabled.

core meaning but missing elements or shifting emphasis—a band that binary scoring would obscure.

Extended thinking provides the largest benefit to the weakest *proprietary* model: GPT-5.2 improves from 6.0% to 13.4% equivalent (2.2×), while Gemini barely changes (60.4% to 61.3%), consistent with the chrF++ thinking deltas (§3.2). The open-weight models behave differently: reasoning leaves their semantic equivalence essentially unchanged (DeepSeek -1.5, Qwen 0.0, GLM $+0.2$ points) while increasing their refusal rates (§3.5).

3.7 Qualitative Error Analysis

Reading the lowest-scoring Lakota→English items reveals a consistent failure mode: compositional translation, rendering idioms and culturally specific vocabulary element by element rather than by conventional meaning. The word glossed "electricity" is rendered "sacred," and the idiom *Wakȟálya yo* ("make coffee") becomes "sanctify it" (Table 10). The starkest of these failures come from the open-weight models; on the same items the frontier models often recover the meaning—Gemini and Opus render *Wakȟálya yo* as "make coffee," and Sonnet translates the electricity sentence (chrF++ 58.2 vs. 18.6). The pattern is partly an artifact of evaluation design: dictionary example sentences over-represent idiomatic usage, and some literal renderings are defensible translations that the single idiomatic reference does not credit (*Wakȟálya yo* → "boil it"). Most failures, however, aren't defensible alternatives like this—the output simply doesn't match any plausible English translation of the source.

4 Discussion

The LLM judge results (§3.6) demonstrate that chrF++ alone cannot reliably characterize model capability on low-resource translation: it closely tracks semantic equivalence for the strongest models but is misleading for the weakest, where fluent English masks near-zero meaning preservation. The BERTScore check we set aside in §2.5 is nonetheless revealing. What was informative was the per-pair correlation between the two: chrF++ penalizes any surface deviation from the reference (a floor), while BERTScore rewards any fluent English (a ceiling), so the two diverge most for models producing fluent output unrelated to the source. For the proprietary models in the baseline condition, this correlation tracks the semantic-equivalence gradient (Gemini $r = 0.91$; Opus and Sonnet $r = 0.83$–0.86; GPT-5.2 $r = 0.79$).

For external calibration, strong systems reach chrF++ in the mid-50s on a high-resource pair such as Chinese→English (Jiao et al., 2023); Gemini's 59.4 on Lakota→English sits at or above that range, yet its 42.6 on English→Lakota and 60.4% semantic equivalence show the capability is far from what that single number suggests. For comparison within indigenous-language MT specifically, the AmericasNLP 2023 shared task reports ChrF of roughly 25–40 for the best fine-tuned systems translating into eleven indigenous languages (Ebrahimi et al., 2023); Gemini's zero-shot English→Lakota result is comparable to the strongest of these systems, though the metric (ChrF vs. chrF++) and system class (fine-tuned NMT vs. zero-shot LLM) differ.

For the proprietary models, extended thinking yields a consistent but small chrF++ gain (1–7 points) that does not change the prac-

Failure mode	Model	Reference	Model output	chrF++
Coined term read literally	DeepSeek	We live in a small trailer house with no electricity.	We are poor in a very small *sacred* tipi.	18.6
Idiom / phonological slide	DeepSeek	Make coffee (put the water to boil).	Sanctify it!	2.9
Defensible literal (dictionary-bound)	Opus	Make coffee (put the water to boil).	Boil it! / Heat it up!	6.5
Kinship confusion	GLM	Did my daughter call?	Did my younger brother come?	27.6
Wrong content word	DeepSeek	Do you have a car?	Do you have a lighter?	65.9
Fluent but unrelated	DeepSeek	I want to talk to you.	I will be very happy.	6.5

Table 10: Representative L→E failure modes across models. chrF++ shows that surface overlap and meaning preservation diverge: a one-word near-miss (65.9) outscores output unrelated to the source. The two *Wakȟálya yo* ("Make coffee") rows contrast a genuine error ("Sanctify it") with a defensible literal translation ("Boil it") that the idiomatic single reference penalizes.

tical assessment; it is largest on the harder English→Lakota direction and for the weakest model (GPT-5.2, $+7.4$ points of semantic equivalence). It is not what separates the leaders: raising Gemini's reasoning from default to high moves Lakota→English by only $+1.0$ (§3.2), suggesting its lead reflects base capability rather than reasoning budget. The open-weight models behave differently again—reasoning leaves their translation quality and semantic equivalence essentially flat while raising refusals (§3.5), declining items rather than improving them. Taken together, these patterns suggest that on a language this far outside the training distribution, the binding constraint is knowledge rather than inference, and reasoning's main contribution is calibration—recognizing what one does not know—rather than capability.

The wide per-pair variance (σ = 22–28 chrF++ for L→E) likely reflects the distribution of Lakota text in training data. Common phrases such as *Mázaská etáŋ luhá he?* ("Do you have any money?") score at or near 100.0 across nearly all models, suggesting these phrases appear verbatim in training data. Culturally specific constructions involving kinship terms, complex verb morphology, or ceremonial language score near zero. The conversational score stratification (Table 4) confirms this pattern quantitatively: for every model and direction, score-6 pairs substantially outperform score-4 pairs. This suggests that model performance is concentrated on high-frequency phrasebook-like items rather than reflecting general capability across Lakota sentence types.

5 Conclusion

As of spring 2026, no frontier LLM, proprietary or open-weight, can reliably translate Lakota. The best Lakota→English performance (chrF++ 59.4) corresponds to 60.4% semantic equivalence on conversational sentences, but the worst model (GPT-5.2) achieves only 6% equivalence despite producing fluent English throughout. Open-weight models do not close the gap: none approaches Gemini, and even the strongest (DeepSeek-V4-Pro) trails the proprietary leaders substantially. The best English→Lakota performance (chrF++ 42.6) is inadequate for unsupervised use. Extended thinking provides a modest improvement of 1–7 chrF++ points for the proprietary models but does not change the practical assessment; for the open-weight models it alters refusal behavior more than translation quality, surfacing the limitation rather than overcoming it. Diacritic normalization analysis shows models produce roughly correct base characters but place diacritical marks inconsistently, possibly reflecting orthographic heterogeneity in training data.

Limitations

Each pair has a single reference translation, and chrF++ against one reference underestimates quality for valid paraphrases; the LLM judge partially mitigates this but cannot fully

solve it. Lakota also marks the speaker's gender through distinct enclitics—the dataset's *Wakȟálya yo* ("make coffee") uses the men's imperative particle, where a woman would say *Wakȟálya ye*—so a single reference can penalize a valid rendering from the other speech register. The evaluation set consists of constructed conversational examples from the New Lakota Dictionary; performance on narrative or spontaneous text may differ.

Although our open-web analysis (§2.7) finds little verbatim overlap between the evaluation pairs and indexed text, contamination cannot be entirely excluded as a contributing factor. Constructing original Lakota–English pairs not derived from any published source would allow it to be ruled out decisively and is a priority for future work.

No fluent Lakota speaker participated in the evaluation. Quality is assessed via established automatic metrics (chrF++) and LLM-based semantic judges (§2.5); while both are standard practice in low-resource MT evaluation, neither substitutes for evaluation by fluent speakers, and qualitative analysis is correspondingly restricted to axes that do not require Lakota fluency. Collaboration with tribal-college language programs and fluent speakers is the most important methodological next step.

This work evaluates only general-purpose frontier and open-weight LLMs without a fine-tuned dedicated MT baseline such as a Lakota-adapted NLLB-200; comparing zero-shot LLM performance against a fine-tuned low-resource NMT system on the same evaluation set would situate these results within the existing low-resource MT literature.

The specific model versions evaluated reflect API availability in February–May 2026; provider model identities and behavior shift over time. The proprietary and open-weight models were evaluated in different months (§2.8); the cross-provider comparison therefore reflects the model snapshots available at each evaluation rather than a synchronized run, and the relative ordering of comparably scoring models should be read with that in mind.

Ethics Statement

The evaluation set was drawn from the New Lakota Dictionary (NLD), a commercially published reference work. The evaluation set is used for research evaluation and is not redistributed. We acknowledge that orthographic standardization is a contested issue within Lakota communities (Hauff, 2020), and that our use of NLD materials reflects availability rather than endorsement of any particular orthographic standard or institutional approach. Future work should evaluate against materials in other actively used orthographies, ideally in collaboration with language programs at tribal colleges and universities.

Acknowledgments

The author used Claude (Anthropic) for editorial feedback, coding assistance, and experimental design discussion.

Author Contributions

L.R. designed the study, curated the evaluation set, ran all evaluations, and made all methodological decisions.

Data Availability

Code and aggregate results are available at `https://github.com/robotson/lakota-translation-benchmark`. The evaluation set was sourced from the New Lakota Dictionary and is not redistributed; see `data/example_pairs.json` for the data schema. Aggregate results are provided in `results/comparison.csv` and `results/llm_judge_summary.csv`. All code requires only API keys and `pip install litellm sacrebleu python-dotenv bert-score`.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv:2409.19151.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, and 1 others. 2023. Findings of the Americas-NLP 2023 shared task on machine translation

into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219.

Tasha R. Hauff. 2020. Beyond numbers, colors, and animals: Strengthening Lakota/Dakota teaching on the Standing Rock Indian Reservation. *Journal of American Indian Education*, 59(1):5–25.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

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 (EAMT)*, pages 193–203.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. ArXiv:2501.18771.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, and 1 others. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 202–217.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Kaushik Parankusham, Rodrigue Rizk, and K.C. Santosh. 2025. LakotaBERT: A transformer-based model for low resource Lakota language. *arXiv preprint arXiv:2503.18212*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Jan F. Ullrich. 2008. *New Lakota Dictionary*. Lakota Language Consortium, Bloomington, IN.

Albert Sr. White Hat. 1999. *Reading and Writing the Lakota Language*. University of Utah Press, Salt Lake City.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.