Language Model Morphology Evaluation on Canadian Indigenous Languages

Duncan Stothers[1,2] [0000−0001−6873−851X]

[1] Harvard University
[2] University of British Columbia

Abstract. We present an evaluation layer for Canadian Indigenous NLP. We combine an auditable Inuktut→English machine translation sanity baseline with morphology-aware probes for Plains Cree (nêhiyawêwin) and Ojibwe (Anishinaabemowin). For morphology, we evaluate reinflection from lemma plus feature bundle to surface form and structured analysis from surface form to plus-delimited segmentation and tags. We then ablate prompts, trivial and heuristic baselines, model choice, and a lightweight hybrid fallback strategy.

The results expose a gap in the current literature. On Inuktut MT, an open NLLB baseline remains brittle even after a six-code configuration sweep: the best configuration reaches BLEU 0.10 and chrF 8.31 on WMT20 dev, with 12.62% null outputs and 48.89% punctuation-only outputs, despite numeric recall of 0.88. On morphology, prompt conditioning sharply improves *analysis* but not *reinflection*. For Cree, analysis rises from 0.00 under the original prompt to 0.32 with two-shot prompting, and the best open model reaches 0.45. For Ojibwe, analysis rises from 0.00 to 0.62. Reinflection is much more resistant to prompting: the best prompt-conditioned language-model scores are 0.17 accuracy for Cree and 0.33 for Ojibwe, while simple heuristics match or outperform these results. A hybrid-lite rescue layer yields selective gains, most notably raising Ojibwe reinflection to 1.00 accuracy by falling back to simple morphology-aware rules.

Keywords: Indigenous NLP · Canadian Indigenous languages · Plains Cree · Ojibwe · Inuktut · Computational morphology · Evaluation

1 Introduction

Canadian Indigenous language technology currently combines two very different forms of progress. On the one hand, some languages now have visible industrial or benchmarked support, especially Inuktut and Inuktitut through the Nunavut Hansard, WMT20, and major platform deployments. On the other hand, several languages have strong linguistic infrastructure, such as analyzers, dictionaries, and corpora, without a correspondingly mature evaluation layer for general-purpose language models. This asymmetry is especially visible in Plains Cree,

where open morphology tooling is already substantial, but there are comparatively few low-friction evaluations of what open language models can actually do with inflectional structure. [17, 5, 21, 15, 30, 13, 2]

This paper addresses that gap by contributing an *evaluation layer* rather than a new model. The central idea is simple: take task forms that are already canonical in computational morphology, make them runnable under unusually strict reproducibility constraints, and align them with the representations already used in community-facing Indigenous-language infrastructure. The resulting benchmark slices are deliberately small, but they are systematic: they support ablations over prompts, models, baselines, and lightweight hybrid rescue rules, and they produce artifacts that can be rerun on a single workstation without hidden dependencies.

The paper has three linked components. First, we establish an open Inuktut→English MT sanity baseline on WMT20 dev using NLLB-200 distilled 600M and show that even a standard multilingual open checkpoint can be brittle in ways not obvious from surface support alone. Second, we introduce morphology-aware probes for Plains Cree and Ojibwe using two tasks: reinflection from lemma plus feature bundle to surface form, and structured analysis from surface form to plus-delimited segmentation and tags. Third, we systematically compare prompts, models, heuristics, and a hybrid-lite fallback layer.

This experimental program is motivated by five questions:

1. Can an open Inuktut MT baseline be trusted as a default starting point under local, reproducible conditions?
2. How much of morphology failure in open language models is genuine linguistic weakness, and how much is prompt-format instability?
3. Are simple heuristics still competitive on low-resource morphology tasks?
4. Do open model choices matter differently for reinflection and analysis?
5. Can a lightweight hybrid layer improve utility without abandoning an open and inspectable workflow?

2 Background and Related Work

2.1 Indigenous-language NLP, infrastructure, and process

A recurring theme in NLP for Indigenous and low-resource languages is that model performance alone is not enough. Global surveys of language representation continue to show severe inequality in data and evaluation, with most languages effectively absent from mainstream pipelines. [18] In Indigenous-language work, this broader problem is often expressed more concretely: useful language technology depends on inspectable infrastructure, collaborative process, and whether tools can be rerun and repurposed in community settings rather than only demonstrated in research environments. [6, 26, 23] This perspective is central to our design. Open-only weights, no new installs, local execution, and machine-readable artifacts are not incidental engineering preferences; they are responses to a well-documented infrastructure gap.

2.2 Canadian Indigenous NLP and the benchmark asymmetry

Within Canada, the most visible Indigenous-language NLP work has centered on Inuktut and Inuktitut, largely because of the Nunavut Hansard parallel corpus and the WMT20 news translation task. The Hansard established an English–Inuktitut resource that could sustain public evaluation, and WMT20 converted that resource into a shared-task benchmark with multiple system descriptions. [17, 5, 1, 21, 15] Later work showed that evaluation design itself matters for this language pair and that character-level metrics correlate well with human judgment in this polysynthetic setting. [20]

Industrial deployments further increased the visibility of Canadian Indigenous language technology. Google Translate added Inuktut, and Microsoft introduced Inuktitut support in Translator and later publicized neural TTS voices. [7, 29, 19] These deployments matter for access, but they do not provide the kind of open, low-friction benchmark layer that researchers and communities can readily rerun.

2.3 Multilingual Indigenous NLP beyond sentence-level MT

Large multilingual efforts such as No Language Left Behind explicitly frame low-resource language support as a response to digital inequity and scale MT to hundreds of languages and tens of thousands of directions. [32, 8] Parallel work in the Americas has expanded Indigenous-language evaluation through shared tasks that move beyond plain MT into educational-material generation through morphological adaptation and translation metrics. [27, 11] Model-centric work such as IndT5 likewise demonstrates that Indigenous-language pretraining can be useful under sparse data. [31]

What remains underdeveloped, especially in the Canadian context, is a morphology-aware evaluation layer for open language models. The current paper addresses exactly that space: not sentence-level MT, not analyzer construction, and not industrial deployment, but a benchmark layer that measures what open LMs can do with linguistically explicit morphology.

2.4 Computational morphology as the methodological base

Our task choice is not ad hoc. Morphological reinflection and related structured tasks have been standard in computational morphology for years through the SIGMORPHON shared-task series. [10, 9, 28] Those tasks were designed precisely to expose whether models can generalize over inflectional structure rather than simply produce plausible fluent text. Our contribution is to port that benchmark logic into a Canadian Indigenous language setting under unusually strong openness and deployment constraints.

2.5 The Plains Cree and Ojibwe computational ecosystem

Plains Cree is not a tooling blank slate. The GiellaLT ecosystem provides finite-state analyzers and generators with shared engineering conventions across many

low-resource languages, including Plains Cree. [30, 13] The *itwêwina* dictionary integrates lexical resources with finite-state morphology so that users can search by inflected form and inspect paradigms. [2] Additional work on the Ahenakew-Wolfart Plains Cree corpus, interactive completion, and word-level prediction shows that Cree morphology is already operationalized in corpora and usable software interfaces. [4, 24, 22]

The same general pattern holds for related Canadian Indigenous languages. OjibweMorph shows how approachable finite-state morphology can support educational and lexicographic applications. [14] Gitksan finite-state work demonstrates that this infrastructure style extends beyond Cree and Ojibwe. [12] ReadAlong Studio offers a complementary lesson from speech technology: practical, licensed, easy-to-use infrastructure is often what makes language technology actionable. [25]

2.6 LLMs and morphology

Recent work on LLMs and morphology helps interpret our results. Multilingual Wug-style evaluations show that LLM performance deteriorates as morphological complexity increases. More recent work on compositional generalization argues that even instruction-tuned models remain weak when asked to systematically realize morphological structure over novel or uncommon combinations. [3, 16] This makes our observed failure modes—lemma copying, conjunct weakness, person confusion, and format instability—look typical of a broader difficulty class rather than idiosyncratic properties of Cree or Ojibwe.

3 Evaluation Program and Experimental Design

3.1 Experimental program

The paper reports five linked experiment families:

1. **Inuktut MT sanity baseline.** We evaluate an open NLLB checkpoint on WMT20 dev and sweep six tokenizer language-code settings to test whether configuration alone rescues the baseline.
2. **Morphology prompt ablations.** We compare zero-shot, stricter formatting prompts, and one-shot/two-shot variants for reinflection and analysis in Cree and Ojibwe.
3. **Heuristic baselines.** We compare open LMs against trivial and lightweight morphology-aware baselines.
4. **Cross-model ablation.** We evaluate multiple open instruction-tuned models under the best prompts selected for each language and task.
5. **Hybrid-lite rescue.** We apply simple validity checks and selective fallback to the best heuristic baseline when the LM output clearly fails structural requirements.

Task	Input	Output	Primary score
Reinflection	ℓ, b	surface $\hat{y}$	exact match, Avg. ED
Analysis	x	analysis $\hat{a}$	Jaccard over tag sets

Table 1. Tasks and scores. Plus-delimited bundles follow GiellaLT conventions. [13, 30]

This design is deliberate. It allows us to separate morphology failure from prompt-format failure, compare learned and rule-based behavior, and test whether a lightweight hybrid layer is already useful before any training or analyzer integration.

3.2 Tasks

We evaluate two morphology tasks.

Reinflection. Given a lemma ℓ and a plus-delimited feature bundle b, generate a surface form $\hat{y}$:

$$(\ell, b) \mapsto \hat{y}.$$

Analysis. Given a surface form x, produce a plus-delimited analysis $\hat{a}$:

$$x \mapsto \hat{a} = m_1 + m_2 + \cdots + \tau_1 + \cdots$$

where the output may contain morphemes and feature tags in a single linearization.

3.3 Data

For morphology, we use compact curated diagnostic sets. Each reinflection item is a triple $(\ell, b, y^\star)$ and each analysis item is $(x, \mathcal{A}^\star)$ where $\mathcal{A}^\star$ is a small set of acceptable analyses. The Plains Cree diagnostic set contains n=6 reinflection items and n=6 analysis items. The Ojibwe replication set contains n=3 reinflection items and n=3 analysis items.

These sets are intentionally diagnostic rather than benchmark-scale. That design is justified by the role they play in the paper: they are *benchmark slices* used to compare prompts, heuristics, models, and hybrid rescue strategies under controlled conditions. Their value lies in interpretability and rerunnability, not in providing a final population-level estimate of morphology competence.

All morphology items are non-sacred and drawn from widely taught paradigms. Feature bundles and segmentation follow GiellaLT-style conventions, and spellings are checked against *itwêwina* where appropriate. [13, 30, 2]

For MT, we use the Inuktut→English WMT20 dev set built from the official SGM files, yielding 5173 sentence pairs.

3.4 Models

The MT sanity baseline uses `facebook/nllb-200-distilled-600M`. The morphology experiments use open causal instruction-tuned models:

- `TinyLlama/TinyLlama-1.1B-Chat-v1.0`
- `Qwen/Qwen2.5-0.5B-Instruct`
- `Qwen/Qwen2.5-1.5B-Instruct`

All decoding is greedy, with no sampling and temperature 0.

3.5 Prompting, baselines, and hybrid-lite

We use short instruction-style prompts that request single-line outputs. The key ablations compare:

- the original zero-shot prompt,
- a stricter formatting prompt,
- a minimal prompt,
- one-shot prompting,
- two-shot prompting.

We also define trivial and lightweight heuristic baselines. For reinflection, these include lemma copying and simple person or conjunct prefix heuristics. For analysis, these include identity outputs, prefix splitting, and prefix splitting with coarse tags.

Hybrid-lite is a simple decision layer. It accepts the LM output when it is structurally valid and otherwise falls back to the best heuristic baseline for that language-task pair. The point is not to simulate a full analyzer-backed system, but to test whether a small amount of explicit structure already changes the tradeoff.

3.6 Normalization, metrics, and artifacts

Before scoring, all outputs are normalized with Unicode NFC and whitespace cleanup, while preserving diacritics and case. Reinflection is scored with exact match accuracy and average Levenshtein distance. Analysis is scored with Jaccard overlap over plus-delimited atom sets. We also track invalid-output rate, prompt-echo rate, and, for analysis, copy-input rate.

Every experiment writes per-item JSONL files, summary JSON files, and a machine-readable runtime configuration. This is one of the paper's practical contributions: the benchmark layer is not just described, it is emitted as a rerunnable artifact set.

Slice	n	BLEU	chrF	Null	Punct.	Num.	Date
All	5173	0.10	8.31	0.13	0.49	0.88	0.01
Short (≤ 5)	677	0.01	9.04	0.18	0.48	0.96	0.00
Medium (6–15)	1700	0.08	6.42	0.06	0.61	0.93	0.09
Long (> 15)	2796	0.09	8.53	0.15	0.42	0.82	0.00
Numeric ref	1518	0.24	11.43	0.02	0.03	0.88	0.01
Date ref	77	0.00	10.10	0.03	0.05	0.83	0.01

Table 2. Slice-level diagnostics for the best open Inuktut→English MT configuration. Null and punctuation-only columns are rates.

4 Results

4.1 Inuktut MT sanity baseline

We begin with the Inuktut MT baseline because it situates the broader evaluation problem. The Section 11 code sweep evaluated six tokenizer language-code settings: `ike_Cans`, `iku_Cans`, `iu_Cans`, `ike_Latn`, `iku_Latn`, and `iu_Latn`. On the subset sweep, all six conditions tied. The best full-run configuration—`sec11_mt_native_01_ike_Cans`—achieved BLEU 0.10 and chrF 8.31 on the full WMT20 dev set, with 12.62% null outputs and 48.89% punctuation-only outputs. Numeric recall remained high at 0.88, but date recall was only 0.01.

This is a useful result even though the absolute translation quality is poor. It shows that an open multilingual checkpoint can preserve superficial structure, especially digits, while still failing badly at sentence-level translation. In other words, apparent language support is not enough; evaluation must also reveal output pathologies.

Figure 1 makes the same pattern visible at a glance. Numeric-heavy lines are relatively more stable, while date-heavy and ordinary sentence slices remain brittle. The middle panel is especially revealing: output failure is not a marginal artifact but a central part of the baseline's behavior.

4.2 Morphology condition landscape

Figure 2 summarizes the full morphology condition space across languages, tasks, prompts, models, baselines, and hybrid-lite. Each panel is column-normalized within task so that clustering highlights *relative* condition behavior rather than raw metric scale. Two broad patterns stand out.

First, reinflection is much less responsive to prompting than analysis. Across both Cree and Ojibwe, many reinflection LM conditions cluster near weak or moderate performance, while heuristic and hybrid-lite rows remain comparatively strong. Second, analysis is markedly more prompt- and model-sensitive. Two-shot prompting and the stronger analysis models occupy visibly better parts of the condition space, especially for Ojibwe.

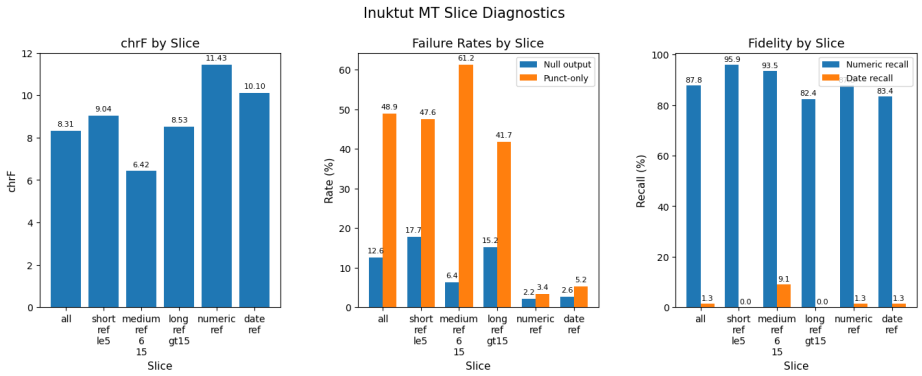

Fig. 1. Inuktut MT slice diagnostics for the best full-run configuration. Left: chrF by slice. Middle: null-output and punctuation-only rates. Right: numeric and date recall. The figure shows that the open baseline preserves digits more readily than sentence content and date structure, and frequently degenerates into null or punctuation-only outputs.

4.3 Prompt sensitivity: analysis is recoverable, reinflection is not

The first major morphology result is that prompt design matters much more for analysis than for reinflection.

For Cree reinflection, the original and strict prompts tie at accuracy 0.17 with AvgED 3.17, and one-shot or two-shot prompting does not improve accuracy. By contrast, Cree analysis improves from 0.00 under the original prompt to 0.21 with one-shot and 0.32 with two-shot prompting.

Ojibwe shows an even sharper split. Reinflection under the best prompts remains at only 0.33 accuracy with AvgED 1.33, while analysis rises from 0.00 under the original prompt to 0.53 with one-shot and 0.62 with two-shot prompting.

This difference is important for interpreting the literature gap. The benchmark is not merely showing that small open LMs fail. It shows *how* they fail. For structured analysis, a substantial portion of the initial error was due to output-format instability and prompt echo. For reinflection, however, prompting does far less. Once the model is asked to realize a specific inflectional bundle as a single surface form, the bottleneck appears to be morphological realization rather than surface formatting.

4.4 Prompt-conditioned LMs versus best heuristics

Table 3 compares the best prompt-conditioned LM result for each language-task pair to the strongest trivial or heuristic baseline.

Two points are especially notable.

First, reinflection remains an area where heuristics are surprisingly competitive. In Cree, the best heuristic baseline matches the LM on accuracy and

Fig. 2. Clustered heatmaps of morphology conditions across Cree and Ojibwe. Each panel summarizes prompt variants, model ablations, baselines, and hybrid-lite under column-normalized metrics. Reinflection panels remain relatively cool except where heuristics dominate, while analysis panels show clear gains from prompt conditioning and, for some settings, model choice.

substantially improves average edit distance. In Ojibwe, a simple person-prefix baseline reaches 1.00 accuracy, clearly outperforming the LM. This suggests that for highly regular, low-resource inflectional patterns, minimal explicit morphology can still dominate small general-purpose LMs.

Second, the story is different for analysis. Here the best prompt-conditioned LM already exceeds the best heuristic in both languages, most clearly in Ojibwe (0.62 versus 0.34). That indicates that structured analysis is a task where small open LMs can provide real added value once prompt formatting is controlled.

4.5 Model ablation

Table 4 reports the best-performing open model for each language and task under the selected prompt conditions.

The ablation shows that model choice matters, but not uniformly. TinyLlama is the strongest reinflection model in both languages, while Qwen 2.5 1.5B is the strongest Cree analysis model. More importantly, bigger is not automatically better. Qwen 2.5 1.5B achieves the best Cree analysis score (0.45), but catastrophically fails on reinflection, yielding 100% invalid outputs for Cree and

Language	Task	Best LM prompt	LM score	Best baseline	Baseline score
Cree	reinflection	strict	Acc=0.17, ED=3.17	conjunct_or_person_prefix	Acc=0.17, ED=1.50
Cree	analysis	two_shot	Jacc=0.32	prefix_split_with_coarse_tags	Jacc=0.24
Ojibwe	reinflection	strict	Acc=0.33, ED=1.33	person_prefix_only	Acc=1.00, ED=0.00
Ojibwe	analysis	two_shot	Jacc=0.62	prefix_split_with_coarse_tags	Jacc=0.34

Table 3. Best prompt-conditioned LM results versus best trivial or heuristic baselines.

Language	Task	Best model	Main score	Invalid	Echo
Cree	reinflection	tinyllama_1.1b_chat	Acc=0.17, ED=3.17	0.00	0.00
Cree	analysis	qwen2.5_1.5b_instruct	Jacc=0.45	0.00	0.00
Ojibwe	reinflection	tinyllama_1.1b_chat	Acc=0.33, ED=1.33	0.00	0.00
Ojibwe	analysis	tinyllama_1.1b_chat	Jacc=0.62	0.00	0.00

Table 4. Best-performing open models by language and task under the selected prompt variants.

effectively unusable reinflection outputs for Ojibwe as well. This is precisely the kind of benchmark behavior the current literature lacks: the evaluation layer distinguishes model families by *task type* and *output stability*, not just aggregate score.

4.6 Hybrid-lite: selective gains, not universal rescue

Table 5 compares raw LM outputs, hybrid-lite outputs, and the best heuristic baseline.

The hybrid layer is useful precisely because it is selective. It helps where explicit morphology is already a strong fallback option, as in Ojibwe reinflection, but it does not improve cases where the raw LM already outperforms the heuristic, as in Cree and Ojibwe analysis. This is a more useful operational result than a blanket claim that hybridization always helps. It shows where simple rule-based rescue is worth keeping in the loop and where it is unnecessary.

4.7 Qualitative findings

The qualitative examples reinforce the quantitative story.

Cree reinflection. The dominant failure is lemma copying. For the conjunct forms `V+AI+Cnj+Prs+1Sg` and `V+AI+Cnj+Prs+2Sg`, the raw LM outputs *mîcisow* instead of the gold *ê-mîcisoyân* or *ê-mîcisoyan*. Hybrid-lite and the heuristic do not fix accuracy here, but they do reduce edit distance by at least inserting the conjunct prefix.

Language	Task	Raw LM	Hybrid-lite	Best heuristic
Cree	reinflection	Acc=0.17, ED=3.17	Acc=0.17, ED=1.50	Acc=0.17, ED=1.50
Cree	analysis	Jacc=0.45	Jacc=0.45	Jacc=0.24
Ojibwe	reinflection	Acc=0.33, ED=1.33	Acc=1.00, ED=0.00	Acc=1.00, ED=0.00
Ojibwe	analysis	Jacc=0.62	Jacc=0.62	Jacc=0.34

Table 5. Hybrid-lite comparison using the best model per language-task pair. The hybrid condition applies simple validity checks and falls back to the best heuristic baseline when appropriate.

Cree analysis. The strongest raw Cree analysis output is not perfect, but it is structurally meaningful. For *mîcisow*, the best model outputs `ni+mîcisow+V+AI+Ind+Prs+4Sg`, which is wrong in person marking yet still overlaps heavily with the gold analysis. This is exactly the kind of behavior a morphology-aware benchmark should expose: the model is not simply random, but it is systematically misaligned at the feature level.

Ojibwe reinflection. Ojibwe reinforces the heuristic point. For *nibimose* and *gibimose*, the raw LM often returns the bare lemma *bimose*. The hybrid and heuristic layers restore the correct person prefix and achieve exact match.

Ojibwe analysis. The strongest raw analysis outputs are strikingly better than the Cree ones. For *bimose*, the best raw LM output is `gi+bimose+V+AI+Ind+Prs+3Sg`, which is wrong in prefix/person but structurally close enough to score Jaccard 0.86 against the gold. This is not perfect morphology, but it is useful evidence that structured analysis is more tractable for small open LMs than reinflection.

5 Discussion

5.1 What the experimental program reveals

Across the whole experimental program, three high-level conclusions emerge.

1. Analysis is more recoverable than reinflection. This is the most robust result in the notebook. In both Cree and Ojibwe, prompt conditioning moves analysis from effectively broken to partially or strongly usable. The same is not true for reinflection, which remains weak across prompts and often loses to heuristics.

2. Heuristics still matter. The benchmark shows this concretely rather than abstractly. In Cree reinflection, a simple heuristic matches the LM on accuracy and beats it on edit distance. In Ojibwe reinflection, a trivial person-prefix heuristic is perfect on the current set. This does not diminish the value of LM-based evaluation; it sharpens it. The benchmark can tell us when learned behavior is actually adding something and when explicit morphology is still the better tool.

3. Model size is not a monotonic predictor of usefulness. Qwen 2.5 1.5B is the strongest Cree analysis model, yet catastrophically unstable for reinflection. TinyLlama is weaker on Cree analysis but more stable overall and the best model for Ojibwe analysis. This is exactly the sort of task-specific benchmark finding that broad sentence-level evaluation often hides.

5.2 Why the small benchmark slices are justified

The morphology slices are intentionally diagnostic. That is not an accidental weakness; it is part of the design logic. The purpose of this paper is to establish an evaluation position and show that it is scientifically and operationally useful. For that purpose, diagnostic slices are enough. They support:

- prompt ablations,
- baseline comparisons,
- model comparisons,
- hybrid rescue evaluation,
- cross-language replication.

All of those comparisons produce stable and interpretable patterns. The contribution is therefore more like a benchmark note or evaluation systems paper than a final leaderboard paper. Future work should scale the slices, but the present scale is already sufficient to make the benchmark layer useful.

5.3 Practical implications

The benchmark supports a simple deployment lesson. If the goal is community-facing tooling under low-resource conditions, analysis and reinflection should not be treated symmetrically. For analysis, prompt-conditioned small LMs may already be worth integrating into experimental tools, especially when outputs are inspected or post-checked. For reinflection, however, lightweight heuristics remain highly competitive and sometimes dominant. A practical system for today would therefore likely be hybrid by default: LM-assisted for structured analysis, rule-anchored for inflection-heavy generation, and explicitly benchmarked before deployment.

6 Conclusion

We introduced a reproducible evaluation layer for Canadian Indigenous NLP that combines an Inuktut MT sanity baseline with morphology-aware probes for Cree and Ojibwe. The contribution is not a new model or a new analyzer. It is a benchmark position that bridges three literatures that are often disconnected in practice: public Indigenous-language MT benchmarks, mature finite-state morphology infrastructure, and the emerging use of general-purpose open LMs.

The results expose a consistent pattern. For Inuktut, a standard open multilingual MT baseline remains brittle even after configuration sweeps, underscoring

the need for explicit local evaluation. For Cree and Ojibwe, prompt-conditioned LMs can recover structured analysis, but reinflection remains much harder and is often rivaled or beaten by simple heuristics. Model choice matters, but not uniformly, and lightweight hybrid rescue can help selectively.

These findings justify the benchmark layer as a contribution in its own right. It makes failure modes visible, makes prompt and model comparisons concrete, and gives communities and researchers an auditable scaffold they can expand with larger datasets, analyzer-backed checks, and more languages. In a field where strong linguistic tools already exist but easy-to-rerun LM evaluations are still scarce, that is a meaningful step forward.

Acknowledgments. This study was supported by funding from iCORD at the University of British Columbia, as well as funded by NVidia through the NVidia Academic Grants Program.

References

1. Shared task: Machine translation of news, wmt20. https://www.statmt.org/wmt20/translation-task.html (2020), includes IU–EN training and dev data links
2. ALTLab, University of Alberta: itwêwina: Plains cree dictionary. https://itwewina.altlab.app/ (2021)
3. Anh, D., Raviv, L., Galke, L.: Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In: Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y. (eds.) Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. pp. 177–188. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.cmcl-1.15, https://aclanthology.org/2024.cmcl-1.15/
4. Arppe, A., Schmirler, K., Harrigan, A.G., Wolvengrey, A.: A morphosyntactically tagged corpus for plains cree. In: Macaulay, M., Noodin, M. (eds.) Papers of the Forty-Ninth Algonquian Conference (PAC49). vol. 49, pp. 1–16. Michigan State University Press, East Lansing, Michigan (2020). https://doi.org/10.14321/j.ctvv417gp.5
5. Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M.R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., Zampieri, M.: Findings of the 2020 conference on machine translation (wmt20). In: Proceedings of the Fifth Conference on Machine Translation. pp. 1–55. Association for Computational Linguistics, Online (nov 2020), https://aclanthology.org/2020.wmt-1.1/
6. Bird, S.: Decolonising speech and language technology. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 3504–3519. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.313, https://aclanthology.org/2020.coling-main.313/
7. Caswell, I.: Google translate learns inuktut. https://blog.google/intl/en-ca/company-news/technology/google-translate-learns-inuktut/ (oct 2024)

8. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., the NLLB Team, et al.: Scaling neural machine translation to 200 languages. Nature (2024). https://doi.org/10.1038/s41586-024-07335-x, https://www.nature.com/articles/s41586-024-07335-x
9. Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M.: CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In: Hulden, M. (ed.) Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection. pp. 1–30. Association for Computational Linguistics, Vancouver (Aug 2017). https://doi.org/10.18653/v1/K17-2001, https://aclanthology.org/K17-2001/
10. Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M.: The SIGMORPHON 2016 shared task—morphological reinflection. In: Elsner, M., Kuebler, S. (eds.) Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 10–22. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2002, https://aclanthology.org/W16-2002/
11. De Gibert, O., Pugh, R., Marashian, A., Vazquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J., Robles, M., Manrique, R., Moreno, O., Lino, A., Coto-Solano, R., Alvarez, A., Agüero-Torales, M., Ortega, J.E., Chiruzzo, L., Oncevay, A., Rijhwani, S., Von Der Wense, K., Mager, M.: Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In: Mager, M., Ebrahimi, A., Pugh, R., Rijhwani, S., Von Der Wense, K., Chiruzzo, L., Coto-Solano, R., Oncevay, A. (eds.) Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). pp. 134–152. Association for Computational Linguistics, Albuquerque, New Mexico (May 2025). https://doi.org/10.18653/v1/2025.americasnlp-1.16, https://aclanthology.org/2025.americasnlp-1.16/
12. Forbes, C., Nicolai, G., Silfverberg, M.: An FST morphological analyzer for the gitksan language. In: Nicolai, G., Gorman, K., Cotterell, R. (eds.) Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 188–197. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.sigmorphon-1.21, https://aclanthology.org/2021.sigmorphon-1.21/
13. GiellaLT: lang-crk: Finite-state and constraint grammar resources for plains cree. https://github.com/giellalt/lang-crk (2025)
14. Hammerly, C., LeVasseur, I., Arppe, A., Stacey, A., Silfverberg, M.P.: Ojibwemorph: An approachable finite-state transducer for ojibwe. Language Resources and Evaluation (2025), https://christopherhammerly.com/publication/ojibwemorph/OjibweMorph.pdf, accepted manuscript; preprint
15. Hernandez, F., Nguyen, V.: The ubiqus english–inuktitut system for wmt20. In: Proceedings of the Fifth Conference on Machine Translation. pp. 213–217. Association for Computational Linguistics, Online (nov 2020), https://aclanthology.org/2020.wmt-1.21/
16. Ismayilzada, M., Circi, D., Sälevä, J., Sirin, H., Köksal, A., Dhingra, B., Bosselut, A., Ataman, D., Plas, L.V.D.: Evaluating morphological compositional generalization in large language models. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.)

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1270–1305. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.59, https://aclanthology.org/2025.naacl-long.59/

17. Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.k., Stewart, D., Micher, J.: The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2562–2572. European Language Resources Association, Marseille, France (may 2020), https://aclanthology.org/2020.lrec-1.312/
18. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the nlp world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020), https://www.aclweb.org/anthology/2020.acl-main.560/
19. Kirk, D.: Amplifying inuktitut voices in the digital age with the power of technology. https://news.microsoft.com/source/canada/features/uncategorized/amplifying-inuktitut-voices-in-the-digital-age-with-the-power-of-technology/ (dec 2024)
20. Knowles, R., Lo, C.k.: Test set sampling affects system rankings: Expanded human evaluation of WMT20 English–Inuktitut systems. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 140–153. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022). https://doi.org/10.18653/v1/2022.wmt-1.8, https://aclanthology.org/2022.wmt-1.8/
21. Knowles, R., Stewart, D., Larkin, S., Littell, P.: Nrc systems for the 2020 inuktitut–english news translation task. In: Proceedings of the Fifth Conference on Machine Translation. pp. 156–170. Association for Computational Linguistics, Online (nov 2020), https://aclanthology.org/2020.wmt-1.13/
22. Kriukova, O., Arppe, A.: Word-level prediction in plains cree: First steps. In: Mager, M., Ebrahimi, A., Rijhwani, S., Oncevay, A., Chiruzzo, L., Pugh, R., von der Wense, K. (eds.) Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024). pp. 15–23. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.americasnlp-1.3, https://aclanthology.org/2024.americasnlp-1.3/
23. Kuhn, R., Davis, F., Désilets, A., Joanis, E., Kazantseva, A., Knowles, R., Littell, P., Lothian, D., Pine, A., Running Wolf, C., Santos, E., Stewart, D., Boulianne, G., Gupta, V., Maracle Owennatékha, B., Martin, A., Cox, C., Junker, M.O., Sammons, O., Torkornoo, D., Thanyehténhas Brinklow, N., Child, S., Farley, B., Huggins-Daines, D., Rosenblum, D., Souter, H.: The Indigenous languages technology project at NRC Canada: An empowerment-oriented approach to developing language software. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 5866–5878. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.516, https://aclanthology.org/2020.coling-main.516/
24. Lane, W., Harrigan, A., Arppe, A.: Interactive word completion for plains cree. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3284–3294. Association for Computational Linguis-

tics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.232, https://aclanthology.org/2022.acl-long.232/
25. Littell, P., Joanis, E., Pine, A., Tessier, M., Huggins Daines, D., Torkornoo, D.: Readalong studio: Practical zero-shot text–speech alignment for indigenous language audiobooks. In: Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. pp. 23–32. European Language Resources Association, Marseille, France (jun 2022), https://aclanthology.org/2022.sigul-1.4/
26. Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., Junker, M.O.: Indigenous language technologies in canada: Assessment, challenges, and successes. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics. pp. 2620–2632. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://aclanthology.org/C18-1222/
27. Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I.V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N.T., Kann, K.: Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In: Mager, M., Oncevay, A., Rios, A., Ruiz, I.V.M., Palmer, A., Neubig, G., Kann, K. (eds.) Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. pp. 202–217. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.americasnlp-1.23, https://aclanthology.org/2021.americasnlp-1.23/
28. McCarthy, A.D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S.J., Heinz, J., Cotterell, R., Hulden, M.: The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In: Nicolai, G., Cotterell, R. (eds.) Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 229–244. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4226, https://aclanthology.org/W19-4226/
29. Microsoft News Center Canada: Microsoft introduces inuktitut to microsoft translator. https://news.microsoft.com/en-ca/2021/01/27/microsoft-introduces-inuktitut-to-microsoft-translator/ (jan 2021)
30. Moshagen, S.N., Pirinen, F., Antonsen, L., Gaup, B., Mikkelsen, I., Trosterud, T., Wiechetek, L., Hiovain-Asikainen, K.: The giellalt infrastructure: A multilingual infrastructure for rule-based nlp. In: Rule-Based Language Technology. NEALT Monograph Series (2023), https://giellatekno.uit.no/publications/6-RBLT-GiellaLT.pdf
31. Nagoudi, E.M.B., Chen, W.R., Abdul-Mageed, M., Çavusoglu, H.: Indt5: A text-to-text transformer for 10 indigenous languages. arXiv preprint arXiv:2104.07483 (2021). https://doi.org/10.48550/arXiv.2104.07483, https://arxiv.org/abs/2104.07483
32. NLLB Team, Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Guzmán, F., Koehn, P., Schwenk, H., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022). https://doi.org/10.48550/arXiv.2207.04672, https://arxiv.org/abs/2207.04672