Visibility as Survival: Generalizing NLP for Native Alaskan Language Identification

Ivory Yang Chunhui Zhang Yuxin Wang Zhongyu Ouyang Soroush Vosoughi
Department of Computer Science, Dartmouth College
{Ivory.Yang.GR, Soroush.Vosoughi}@dartmouth.edu

Abstract

Indigenous languages remain largely invisible in commercial language identification (LID) systems, a stark reality exemplified by Google Translate's LangID tool, which supports over 100 languages but excludes all 150 Indigenous languages of North America. This technological marginalization is particularly acute for Alaska's 20 Native languages, all of which face endangerment despite their rich linguistic heritage. We present GenAlaskan, a framework demonstrating how both large language models and specialized classifiers can effectively identify these languages with minimal data. Working closely with Native Alaskan community members, we create Akutaq-2k, a carefully curated dataset of 2000 sentences spanning all 20 languages, named after the traditional Yup'ik dessert, symbolizing the blending of diverse elements. We design few-shot prompting on proprietary and open-source LLMs, achieving nearly perfect accuracy with just 40 examples per language. While initial zero-shot attempts show limited success, our systematic attention head pruning revealed critical architectural components for accurate language differentiation, providing insights into model decision-making for low-resource languages. Our results challenge the notion that effective Indigenous language identification requires massive resources or corporate infrastructure, demonstrating that targeted technological interventions can drive meaningful progress in preserving endangered languages in the digital age.

1 Introduction

The exclusion of Indigenous languages from mainstream NLP technologies (Littell et al., 2018; Moshagen et al., 2024) reflects a systemic bias in language technology: the prioritization of high-resource languages at the expense of linguistic diversity (Dash, 2024). Nowhere is this marginalization more evident than in Alaska, a North American region home to 20 Native Alaskan languages

Figure 1: An illustration of how we utilize *demonstrations* to generalize LLMs to identify endangered native Alaskan languages.

(Krauss, 2007), all of which are endangered (Grenoble, 2018; Reo et al., 2019). Despite their rich linguistic heritage, these languages contend with limited digital resources and minimal computational support (Jensen, 2020).

This technological invisibility is exemplified by Google Translate's widely used Language Identification (LangID) tool (Caswell et al., 2020), which supports over 100 languages but completely excludes all approximately 150 Indigenous languages of North America. As a result, Native Alaskan languages lack even the most fundamental capability and dignity of being identified online. Addressing this gap demands generalizable[1] NLP models capable of processing languages in extreme low-resource settings (Mager et al., 2018; Yang et al., 2025b). **We posit that acknowledging a language begins with the ability to accurately identify it, as visibility is the first step toward inclusion.**

[1]This paper is positioned in response to ACL 2025 Special Theme: Generalizable NLP Models, as a demonstration piece to raise awareness for endangered Native Alaskan languages.

Findings of the Association for Computational Linguistics: ACL 2025, pages 6965–6979
July 27 - August 1, 2025

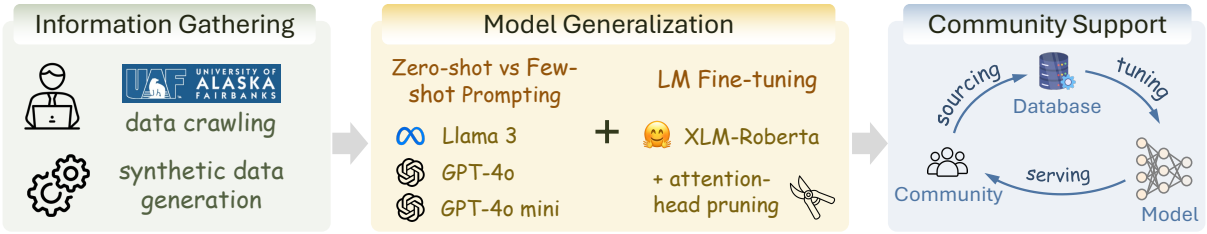


Figure 2: *GenAlaskan*'s three-part contributions: 1. Information Gathering (Manual dataset, synthetic data generation pipeline) 2. Model Generalization (Zero-shot vs Few-shot prompting, LM Fine-tuning) 3. Community Engagement and Support.

To address this challenge, we introduce GenAlaskan, a generalizable framework for identifying all 20 Native Alaskan languages. Our work begins with creating *Akutaq-2k*[2], the **first comprehensive digital dataset** of Native Alaskan languages, comprising 2000 sentences evenly distributed across all 20 languages. Building on this foundation, we develop two complementary approaches (Figure 2): (1) Few-shot prompting with Large Language Models (LLMs) and (2) Fine-tuned classification with XLM-RoBERTa. While initial zero-shot attempts with LLMs shows limited success, our few-shot prompting approach achieves nearly 100% accuracy on proprietary and open-source LLMs, demonstrating effective generalization to extreme low-resource settings. In parallel, our fine-tuned XLM-RoBERTa classifier, enhanced through targeted attention mechanism optimization, achieves robust performance even with limited training data.

By showing that both large-scale LLMs and smaller fine-tuned models can **generalize** effectively to highly endangered languages, we challenge the notion that only large institutions can address Indigenous language identification. Our lightweight, generalizable approach succeeds where mainstream technology communities have overlooked, proving that meaningful progress in Indigenous language technology can be illuminated by an enduring voice.

2 Related Work

Research on NLP applications for Native Alaskan languages remains limited, with most efforts focusing on data documentation (McMillan-Major, 2023) and community-driven revitalization (Dementi-Leonard and Gilmore, 1999; Counceller, 2012; Jennings, 2024) rather than computational modeling (Surma and Truong, 2023). Prior work in the NLP domain has described a repository of example sentences in three endangered Athabascan languages (Koyukon, Upper Tanana, Lower Tanana) (Nordhoff et al., 2016) and how they can be used by researchers and teachers, as well as a new online Akuzipiq-English dictionary (Hunt et al., 2023). Out of the 20 Native Alaskan languages, Yupik appears to be the language with the most amount of existing research interest; prior work includes a case study (Chen, 2019), online dictionary (Hunt et al., 2019), as well as efforts to improve morphological analysis (Chen et al., 2020). Most of the other languages remain unexplored within the context of NLP.

In the broader context of endangered language identification, mainstream LangID tools offer limited or inconsistent support for Native Alaskan languages, as summarized in Appendix D. Notably, only GlotLID (Kargaran et al., 2023) features support for certain Native Alaskan languages, specifically Central Yupik, Gwich'in, Haida, and Inupiaq. Meanwhile, Yang et al. (2025c) introduced a novel approach for detecting Native American languages by leveraging linguistic similarities. Using a Random Forest classifier (Hastie et al., 2009) trained on a 10k dataset of Navajo and 20 languages misidentified by Google Translate's LangID (Caswell et al., 2020), they developed a system that not only identified Navajo with near-perfect accuracy but also generalized to classifying other Athabaskan languages as Navajo. This methodology demonstrated that even with limited training data, statistical learning approaches can achieve high accuracy in endangered language identification, reinforcing the potential for generalizable models in low-resource linguistic settings (Alvarez et al., 2025; Yang et al., 2025a).

[2]Named after akutaq ('auk-goo-duck'), a traditional Yup'ik frozen dessert symbolizing the blending of diverse elements. Just as akutaq nourishes generations, this dataset unifies and preserves Native Alaskan languages in the digital age.

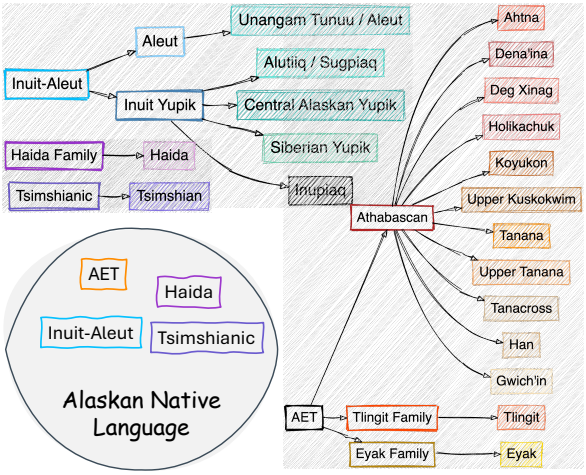


Figure 3: Family tree for Native Alaskan languages, categorized into roughly four families: Inuit-Aleut, AET, Haida, and Tsimshianic.

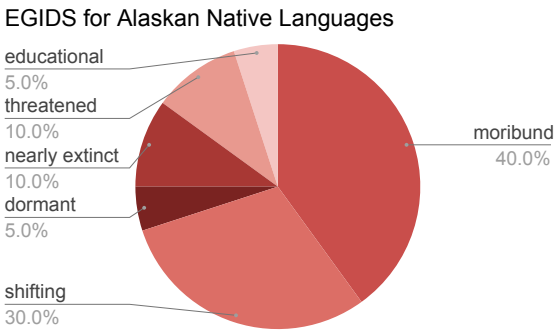


Figure 4: The Expanded Graded Intergenerational Disruption Scale (EGIDS), an attempt to measure language vitality by assessing how the language is used, of 20 Alaskan native languages. Color intensity corresponds to the severity of lost vitality. The severity ranking is *dormant* > *nearly extinct* > *moribund* > *shifting* > *threatened* > *educational*.

3 Native Alaskan Language Landscape

Overview Alaska is home to 20 Indigenous languages (Krauss, 2007), each carrying deep cultural, historical, and epistemological significance (Reo et al., 2019). As shown in Figure 3, these languages can be categorized into four major families: Inuit-Aleut, Athabaskan-Eyak-Tlingit (AET), Haida, and Tsimshianic. The Inuit-Aleut family (Allen and Crago, 1992), spoken across the Arctic regions of Alaska, Canada, and Greenland, includes languages such as Inupiaq and Central Alaskan Yup'ik, which are polysynthetic, encoding rich semantic information within complex word structures. The AET family (Krauss, 1986), which shares linguistic ancestry with Navajo and other Athabaskan languages of the Southwest, includes Gwich'in and Dena'ina, which feature intricate tone systems and highly agglutinative morphology. Haida (Martineau, 2002), spoken in Southeast Alaska, remains linguistically isolated, with debates over its classification, while Tsimshianic languages (Forbes, 2023), such as Coast Tsimshian and Sm'álgyax, exhibit complex verbal morphology and sound systems distinct from neighboring families. Figure 5 shows a overall geographical distribution of this languages on a map of Alaska.

Endangered Status All 20 native Alaskan languages are endangered (Krauss, 1996), as visualized in Figure 4. Eyak is extinct (Naeem-ur Rehman and Abbas, 2024), while Haida, Tsimshian, and Tlingit are critically endangered (Adamou, 2024), spoken by only a handful of elderly speakers. Dena'ina, Gwich'in, Tanacross,

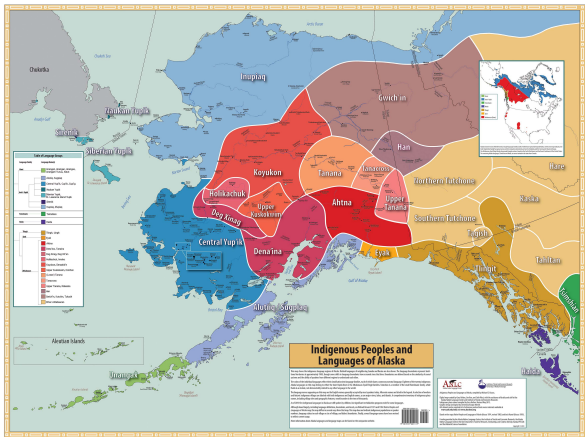


Figure 5: A map of Alaskan languages colored by section. This figure is sourced from https://www.uaf.edu/anlc/languages-move/languages.php.

Ahtna, Hän, Koyukon, Upper Kuskokwim, Tanana, Upper Tanana, Holikachuk, and Deg Xinag are classified as severely endangered (Grenoble and Ignatieva, 2024), with very few fluent speakers remaining, primarily among older generations. Unangax (Aleut), Alutiiq (Sugpiaq), Siberian Yupik, and Inupiaq, though still spoken, are considered endangered due to declining intergenerational transmission (Panova, 2024). Central Alaskan Yup'ik, the most widely spoken with approximately 10,000 speakers (Mithun, 2024), remains threatened as younger generations increasingly shift to English. The obsoletion of any of these languages would represent an irreversible cultural and historical loss (Pakendorf, 2024).

4 Native Alaskan Languages Dataset

Challenges of Existing Resources Existing datasets for Native Alaskan languages are scarce, with most existing linguistic resources being limited to online dictionaries (Hunt et al., 2019, 2023), and small corpora from language preservation initiatives. Some documentation exists through university archives (Coronado and Zavalina, 2024), language revitalization programs (Jia, 2024), and community-driven efforts (Lidubwi and Ndavula, 2025), but these are often fragmented and not readily accessible for computational use. The absence of standardized, publicly available corpora has hindered progress in NLP for these languages.

Manual Curation via Online Community To address this gap, we manually curate a dataset by collecting publicly available sentences from online sources[3] (Indians.org, 2025; Museum, 2025; of Alaska Fairbanks, 2025; Languages, 2025), including linguistic documentation, educational resources, and community-driven language projects. Each language was represented with 100 sentences, ensuring coverage of different linguistic structures. The data was organized into a spreadsheet, with each row containing a Native Alaskan language sentence and its corresponding label.

Synthetic Data Pipeline Given the limited availability of existing data, we initiate the development of a synthetic data expansion pipeline to generate additional high-quality data for endangered Native Alaskan languages, using a combination of few-shot prompting and language-specific tailored instructions[4]. As proof of concept, we test this approach on Ahtna, one of our 20 languages, by using GPT-4o to generate new Ahtna-English sentence pairs based on a few-shot prompting setup. Our process involved validating GPT-generated translations against real Ahtna sentences using Levenshtein similarity scoring, ensuring that the model only initialized synthetic data generation when a similarity of 50% or higher was achieved, as shown in Figure 6. Preliminary results show that GPT-4o can generate Ahtna text with reasonable fidelity, though certain phonetic and grammatical inconsistencies remain, highlighting the need for further refinement and human validation. Moving forward, this pipeline will be extended to other Native Alaskan languages, integrating iterative refinement and community validation to ensure authenticity and usability. Our goal is to create scalable, ethically sourced synthetic data that not only supports model training but also contributes to broader language preservation and revitalization efforts.

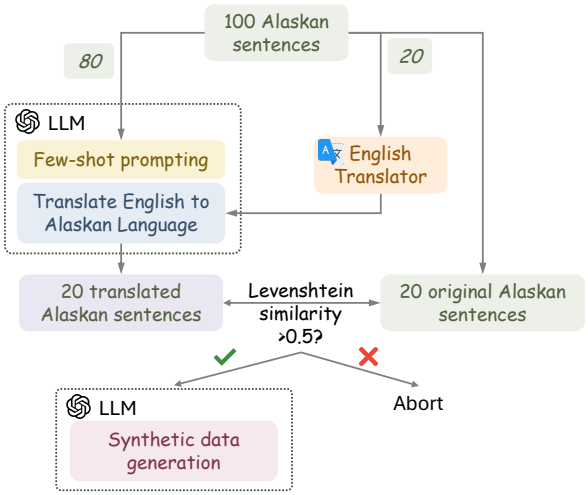


Figure 6: Our synthetic data generation framework.

5 Generalization of LLMs for Native Alaskan Language

5.1 Evaluation Setup

We evaluate whether a few-shot prompting approach using LLMs can significantly improve language identification for Native Alaskan languages compared to their zero-shot performance. Our goal is to determine how few-shot prompting could bridge the gap between large commercial LLMs and specialized solutions for low-resource languages. To achieve this, we test three models, namely GPT-4o, GPT-4o-mini, and LLaMA-3.2-3B, on their ability to identify 20 Native Alaskan languages with minimal prior exposure. The evaluation uses a dataset of 2,000 sentences evenly distributed across 20 languages, with 100 sentences per language. We adopt a **3-4-3 split**, dividing the sentences into three phases: 1. **Zero-shot phase**: Models classify 30 sentences per language without any prior exposure. 2. **Few-shot phase**: Models receive 40 labeled examples per language to learn linguistic patterns and features. 3. **Validation phase**: The models test on 30 new, unseen sentences per language to assess how well they generalize after few-shot prompting. Each sentence is presented in isolation, without metadata or transliteration, forcing the models to rely solely on linguistic features

[3] All dataset citations are provided in the GitHub repository; link provided in Ethics section.

[4] In the context of Ahtna, we enforced SOV word order and leveraged phonetic adaptations.

Language	GPT-4o		GPT-4o Mini		LLaMA3.2-3B	
	Zero-S	Few-S	Zero-S	Few-S	Zero-S	Few-S
Siberian Yupik	0.000	1.000	0.000	1.000	0.000	1.000
Alutiiq	0.100	1.000	0.000	0.967	0.000	1.000
Deg Xinag	0.300	1.000	0.000	1.000	0.000	0.967
Gwich'in	0.167	1.000	0.000	1.000	0.000	0.967
Haida	0.767	1.000	0.000	1.000	0.000	1.000
Holikachuk	0.000	1.000	0.000	0.967	0.000	1.000
Eyak	0.033	1.000	0.000	1.000	0.000	1.000
Tanacross	0.000	1.000	0.000	0.967	0.000	0.967
Han	0.033	1.000	0.000	0.967	0.000	0.933
Lower Tanana	0.000	1.000	0.000	0.900	0.000	1.000
CA Yup'ik	0.067	1.000	0.000	1.000	0.000	1.000
Ahtna	0.000	1.000	0.000	0.967	0.000	1.000
Tlingit	0.600	1.000	0.133	1.000	0.033	0.967
Aleut	0.533	1.000	0.000	1.000	0.000	1.000
Inupiaq	0.067	1.000	0.000	1.000	0.000	1.000
Tsimshian	0.000	1.000	0.000	0.933	0.000	1.000
Dena'ina	0.000	1.000	0.000	0.967	0.000	1.000
Upper Tanana	0.000	1.000	0.000	0.933	0.000	0.967
Koyukon	0.000	1.000	0.000	1.000	0.000	1.000
Upper Kuskokwim	0.000	1.000	0.000	1.000	0.000	1.000

Table 1: Zero-shot (Zero-S) and few-shot (Few-S) classification performance across 20 Native Alaskan languages for GPT-4o, GPT-4o Mini, and LLaMA3.2-3B.

for classification. The prompt is carefully designed to elicit a direct response, and requires the model to return only the language name without explanation or reasoning. By comparing zero-shot and few-shot performance, we quantify the extent to which in-context learning improves language identification in extremely low-resource settings.

5.2 Zero-shot Evaluation

Zero-shot performance across 20 Native Alaskan languages (Table 1) reveals significant variation among GPT-4o, GPT-4o-mini, and LLaMA-3.2-3B. In this evaluation, each model classifies 30 sentences per language without prior exposure, simulating a real-world zero-shot scenario.

GPT-4o demonstrates the best zero-shot performance, achieving moderate accuracy for certain languages such as Haida (0.767), Aleut (0.533), and Tlingit (0.600). This suggests that GPT-4o's broad pretraining data may contain partial exposure to linguistic features related to Native Alaskan languages, allowing it to recognize structural patterns. However, its overall performance remains inconsistent, with many sentences misclassified into more widely spoken languages, highlighting the lack of fine-grained distinctions required for accurate classification in extremely low-resource settings.

GPT-4o-mini, by contrast, struggles significantly, with near-zero accuracy for most languages. Optimized for efficiency and conversational tasks rather than linguistic recall, it frequently defaults to incorrect or generic classifications. This highlights a fundamental limitation of smaller, instruction-tuned models in tasks that require implicit linguistic priors and specialized knowledge.

LLaMA-3.2-3B also struggles in zero-shot settings, achieving performance comparable to GPT-4o-mini. The highest accuracy observed just 0.033 for Tlingit. Despite its poor zero-shot performance, LLaMA-3.2-3B shows remarkable improvement in the few-shot phase, where it achieves accuracy levels comparable to GPT-4o and significantly surpasses GPT-4o-mini. This highlights the potential of open-source models when supplemented with minimal supervision.

5.3 Few-shot Evaluation

Introducing 40 labeled examples per language in the few-shot phase leads to a dramatic improvement in performance across all three models. After this exposure, GPT-4o, GPT-4o-mini, and LLaMA-3.2-3B achieve near-perfect accuracy for most languages (Table 1), demonstrating that even minimal in-context learning is sufficient for these models to generalize effectively.

LLaMA-3.2-3B, in particular, transforms from near-zero performance in zero-shot to matching GPT-4o in few-shot settings. Languages such as Alutiiq, Holikachuk, and Central Alaskan Yup'ik reach perfect accuracy (1.000), underscoring its strong capacity to generalize with limited supervision. *This suggests that open-source models can rival or even surpass commercial LLMs when given the right support.*

GPT-4o-mini also improves significantly but still lags slightly behind GPT-4o and LLaMA-3.2-3B for certain languages (e.g., Deg Xinag: 0.967, Han: 0.933). Despite this, it demonstrates that even smaller models can achieve meaningful results when provided with targeted few-shot prompting.

These results highlight the adaptability of LLMs with minimal examples, proving that few-shot in-context learning can bridge the gap for languages lacking adequate pretraining data. While some misidentifications remain, additional fine-tuning and dataset expansion offer clear paths for improvement. This experiment demonstrates that GPT-4o and LLaMA-3.2-3B can rapidly learn and generalize to all 20 Native Alaskan languages through

Pruned Attention Head	Performances After Pruning			
	ACC	CE	F_1^{mac}	F_1^{mic}
Head 1	0.413	1.767	0.400	0.413
Head 2	0.428	1.857	0.402	0.428
Head 3	0.348	1.982	0.325	0.348
Head 4	0.300	1.951	0.281	0.300
Head 5	0.355	1.902	0.342	0.355
Head 6	0.418	1.798	0.396	0.418
Head 7	0.333	1.939	0.316	0.333
Head 8	0.268	2.077	0.249	0.268
Head 9	0.392	1.819	0.358	0.393
Head 10	0.363	2.097	0.335	0.362
Head 11	0.358	2.031	0.342	0.358
Head 12	0.450	1.750	0.421	0.450

Table 2: Performance of zero-masking different attention heads in XLM-R classifier on 20 Native Alaskan languages. We report Accuracy (ACC), Cross-Entropy (CE), Macro-F_1 (F_1^{mac}), and Micro-F_1 (F_1^{mic}).

a structured few-shot setup, offering a scalable, resource-efficient solution for low-resource language identification. *By leveraging few-shot adaptation, we provide a replicable framework for expanding NLP to other endangered languages, promoting greater linguistic inclusivity worldwide.*

6 Small LM for Native Alaskan Language

6.1 Evaluation Setup

The goal of this experiment is to assess how a multilingual transformer model, XLM-RoBERTa, performs in identifying Native Alaskan languages, and to analyze the importance of specific attention heads in the classification process. In addition to baseline classification, we introduce an attention head pruning mechanism to determine whether certain heads were crucial for distinguishing between these low-resource languages and whether their removal affects model performance. 1. **Dataset and Preprocessing.** The dataset consists of 2,000 labeled sentences, evenly distributed across 20 Native Alaskan languages. This dataset is the same used in the previous zero-shot and few-shot evaluations. To ensure a balanced and representative evaluation, we perform a stratified split, allocating 80% of the data for training and 20% for testing. The dataset is converted into a Hugging Face Dataset format for compatibility with transformer-based training pipelines. 2. **Model and Training.** We use the XLM-RoBERTa-base model, a multilingual transformer pretrained on a wide range of languages, but with no explicit exposure to Native Alaskan languages. The model is fine-tuned as a sequence classifier, where each input sentence is classified into one of the 20 language labels. Tokenization is handled using the `XLMRobertaTokenizer`, with input sequences truncated to a maximum length of 128 tokens. Training is conducted using a learning rate of 1e-5, a batch size of 32, and 100 training epochs to ensure convergence while preventing overfitting.

To analyze the role of individual attention heads in language identification, we implement an attention pruning mechanism by selectively disabling individual heads in all self-attention layers. For each head, the corresponding query, key, and value weights are set to zero, effectively removing its contribution during inference. The model's performance is evaluated before and after pruning each head, allowing us to observe the impact on accuracy, precision, recall, F1 score, and cross-entropy loss. The results of the pruning experiment are presented in Table 2 and Figure 7, highlighting how pruning specific heads affects language classification performance.

6.2 Evaluation Metrics

To quantify the effect of attention head pruning, we evaluate standard multi-class classification metrics: accuracy, macro-F1, micro-F1, and cross-entropy loss. Accuracy provides a general measure of overall performance, while macro-F1 and micro-F1 offer more nuanced insights. Macro-F1 captures per-class performance by giving equal weight to each language class, making it suitable for unbalanced datasets. In contrast, micro-F1 aggregates all instances and reflects the model's overall ability to distinguish between languages. Cross-entropy loss measures the model's confidence in its predictions and penalizes incorrect classifications based on the predicted probability assigned to the true class. Lower cross-entropy loss indicates higher confidence and better performance. Detailed mathematical formulations for these metrics are provided in Appendix A. These definitions include how accuracy, macro-F1, and micro-F1 are computed, along with the formula for cross-entropy loss in this task.

6.3 Pruning Analysis and Model Results

To investigate the role of individual attention heads in language classification, we evaluate the XLM-RoBERTa classifier before and after pruning each head. Table 2 summarizes the impact of pruning

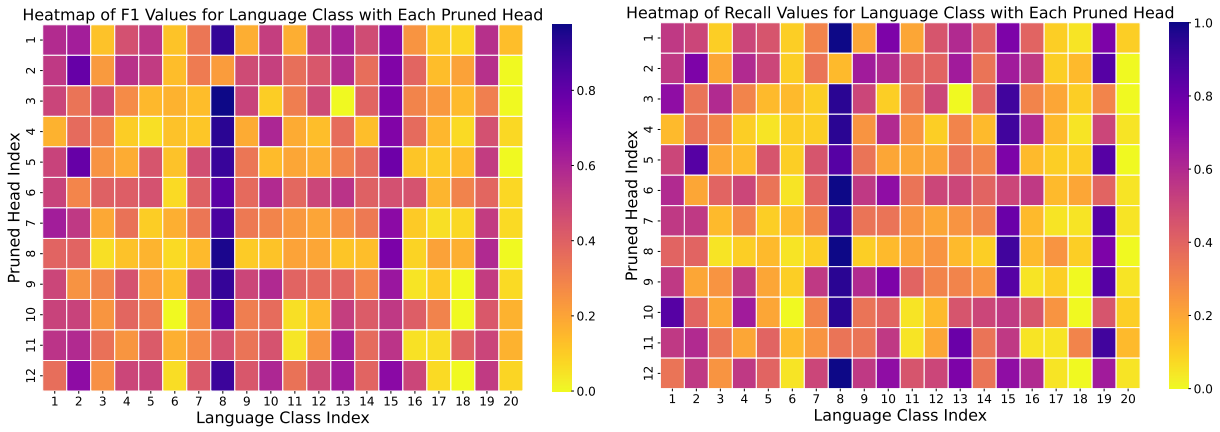


Figure 7: Heatmaps of F_1 scores and Recalls for language classes with each attention head pruned.

each head on accuracy, cross-entropy loss, and F1 scores. The heatmaps in Figure 7 provide a visual representation of F1 and recall scores across all 20 languages for each pruned head. Overall, pruning attention heads does not degrade performance uniformly. For most heads, the model shows resilience, maintaining consistent accuracy and F1 scores. However, specific heads play crucial roles for certain language classes, while others appear to introduce noise. 1. **Head 8** demonstrates the most significant performance drop. After pruning Head 8, accuracy falls to 0.268, Macro-F1 drops to 0.249, and cross-entropy increases to 2.077. This indicates that Head 8 is essential for distinguishing multiple languages, and its removal disrupts key patterns. In contrast, pruning **Head 12** results in the best overall performance, with an accuracy of 0.450 and a Macro-F1 score of 0.421, suggesting that this head contributes minimally or helps stabilize predictions. 2. **Language Class 8 (Tanacross)** shows high sensitivity to **Head 2 and Head 11**. Pruning either of these heads causes a sharp decline in F1 and recall, as reflected in the heatmaps. This confirms that these heads capture critical linguistic features for identifying Class 8. Interestingly, for **Language Class 15**, which generally has poor performance, pruning **Head 6** improves both F1 and recall scores, implying that this head may introduce noise or irrelevant patterns. These results suggest that attention patterns in XLM-RoBERTa are both specialized and distributed. While certain heads are crucial for specific languages, others can be pruned without degrading performance. *This opens up possibilities for using attention pruning as a regularization technique to improve model generalization and reduce complexity in low-resource language settings.*

7 Applicability and Community Feedback

The ability to accurately identify endangered languages extends beyond NLP research. It is fundamental to digital preservation, automated transcription, and educational tools for revitalization. By demonstrating that large language models can recognize Native Alaskan languages with minimal supervision, we lay the groundwork for automatic subtitling, real-time translation, and Indigenous language integration in voice assistants, ensuring their presence in the digital age. Yet, dignity and recognition are just as vital as technological progress (Bird, 2020). Language is identity, and its digital presence affirms its legitimacy and survival. To uphold transparency, ethics, and community respect (Bird, 2024), we engage with Native Alaskan speakers through informal discussions and three formal interviews to seek guidance on our research direction. Interviewees were compensated, and with full permission, one transcript (with identifying information redacted) is included in Appendix H, featuring a Kenaitze Indian Tribe member from Anchorage, AK.

Community members reacted with shock and frustration, assuming that major commercial language technologies would already support at least some Native languages, especially widely spoken ones like Navajo, yet they found no such inclusion. Many express that if a small team with limited resources could develop a working language identification system, then large corporations with far greater infrastructure could do so effortlessly, but have simply chosen not to. One of our team members likened this deliberate exclusion to creating a map while knowingly erasing entire regions where people live. This absence of technological support not only marginalizes these languages but also re-

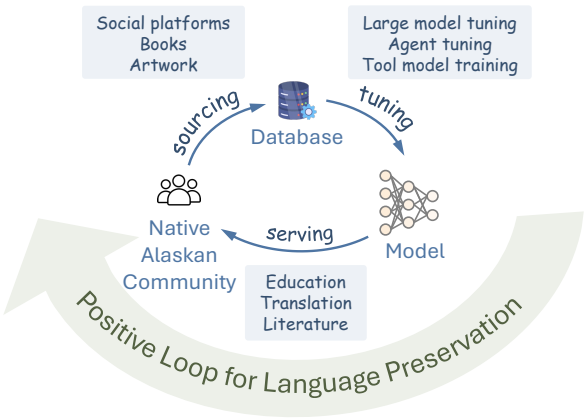


Figure 8: Engagement with the Native Alaskan Community initiates a positive feedback loop.

inforces the systemic neglect that has historically threatened their survival.

Beyond this study, we are in discussions with a commercial language translation company specializing in endangered languages, whose fairly compensated linguists ensure that knowledge is shared with respect and care. While we hope to actively collaborate, human validation remains costly, and sustainable preservation efforts require institutional support. Greater visibility for this work is essential, not just to advance research but to foster the recognition and funding necessary to make meaningful, long-term contributions. By demonstrating the feasibility of lightweight, generalizable approaches, we hope to inspire broader engagement and investment in Indigenous language preservation, ensuring that these languages remain not just studied, but actively supported in the digital age. Some reflection of the real-life applicability of our work is featured in Appendix B. Additional cultural and linguistic context for the Native Alaskan languages covered is provided in Appendix G.

8 Future Work

While this study establishes the effectiveness of few-shot prompting for Native Alaskan language identification, several avenues remain for future exploration. Expanding this approach to other Indigenous and endangered language families, such as Athabaskan languages beyond Alaska or other Arctic and circumpolar languages, would further test its scalability and reinforce its generalizability (Appendix F). Additionally, future work should focus on developing more sophisticated techniques specifically tailored to Native Alaskan languages, such as incorporating linguistic structure, phonetic features, or community-driven validation methods to refine model performance. Beyond identification, integrating LLMs into active language revitalization efforts, including machine translation, speech synthesis, and interactive learning tools, could support broader accessibility and digital engagement. Ultimately, our findings challenge the assumption that resource-intensive methods are necessary for language technology development. Instead, we demonstrate that lightweight, targeted approaches can empower underrepresented languages, paving the way for a more inclusive and linguistically diverse NLP landscape.

9 Conclusion

The exclusion of Native Alaskan languages from commercial language technologies is not due to insurmountable technical barriers but rather a lack of initiative. While discussions on language preservation are widespread, few efforts translate into tangible action. We show that generalizable NLP approaches, such as few-shot prompting with LLMs and fine-tuned classification with XLM-RoBERTa, can effectively identify these languages without requiring extensive data or corporate-scale resources.

Critically, action does not have to be costly or unattainable. In an academic setting, lightweight, targeted approaches can make a significant impact, and our results show that even minimal supervision enables high-accuracy language identification. The continued neglect of these languages in mainstream NLP is a choice, not a necessity. If a small team with limited resources can make meaningful progress, then large-scale institutions and research communities have no excuse for inaction. The level of interest in endangered language technology makes this work not only relevant but highly achievable. Generalizable NLP methods provide a viable pathway for expanding linguistic inclusion, and there is no reason this work should not be extended further.

We call on the NLP community to move beyond discussion and take concrete steps, whether by expanding datasets, refining models, or collaborating with Indigenous speakers, to ensure that these languages are not just studied, but actively supported. By acknowledging and integrating Indigenous languages into technological spaces, we take a necessary step toward recognition, revitalization, and digital survival.

Limitations

While our study demonstrates that few-shot prompting and fine-tuned classification can effectively identify Native Alaskan languages, several limitations remain. Our 2000-sentence dataset, while comprehensive, does not capture the full linguistic diversity of these languages, including dialectal variations. Additionally, LLMs like GPT-4o inherently rely on pre-training data we cannot fully control, making it unclear how much prior knowledge influences their performance. Our XLM-RoBERTa fine-tuning and attention pruning provide insights into model decision-making but do not fully explain the linguistic features driving classification. Finally, while we engaged with Native Alaskan speakers for feedback, further human validation from fluent speakers is needed to assess model outputs with greater linguistic accuracy. Addressing these gaps requires broader collaboration, expanded datasets, and deeper community involvement to ensure computational methods truly support language preservation efforts.

Ethics

Our work prioritizes the ethical and respectful treatment of Native Alaskan languages and communities, recognizing that computational research on endangered languages must be conducted with care. We actively engaged with Native Alaskan speakers, incorporating their insights through informal discussions and formal interviews to ensure our research aligns with community perspectives. Interviewees were compensated fairly, and all shared data was used with full consent and transparency. We acknowledge that language is deeply tied to identity and cultural sovereignty. Any computational approach must serve, not exploit, Indigenous communities. While our models demonstrate technical feasibility, they are not a substitute for human expertise and community-driven revitalization efforts. Future work must prioritize collaborative validation, data ownership, and ethical data collection, ensuring that linguistic technology benefits Indigenous speakers first and foremost. Our manually-curated dataset and code has been made available at https://github.com/ivoryayang/GenAlaskan.

Acknowledgment

This work was partially supported by the CompX Faculty Grant from the Neukom Institute at Dartmouth College.

References

Evangelia Adamou. 2024. *Endangered Languages*. MIT Press.

Shanley EM Allen and Martha B Crago. 1992. First language acquisition of inuktitut. In *Inuit studies occasional papers 4: Proceedings of the seventh Inuit studies conference*, pages 273–281.

Jesus Alvarez, Daua Karajeanes, Ashley Prado, John Ruttan, Ivory Yang, Sean O'Brien, Vasu Sharma, and Kevin Zhu. 2025. Advancing uto-aztecan language technologies: A case study on the endangered comanche language. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 27–37.

Steven Bird. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*, pages 3504–3519. Association for Computational Linguistics (ACL).

Steven Bird. 2024. Must nlp be extractive? In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 14915–14929. Association for Computational Linguistics (ACL).

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.

Emily Chen. 2019. Measuring the value of linguistics: A case study from st. lawrence island yupik. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–33.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved finite-state morphological analysis for st. lawrence island yupik using paradigm function morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2676–2684.

Sergio I Coronado and Oksana L Zavalina. 2024. Digital language archiving: Reuse and adaptation of nonlis learning materials in lis education. In *Proceedings of the ALISE Annual Conference*.

April Counceller. 2012. A decade of language revitalization: Kodiak alutiiq on the brink of revolution. *Journal of American Indian Education*, pages 15–29.

Niladri Sekhar Dash. 2024. Corpus linguistics and language technology. In *Routledge Encyclopedia of Technology and the Humanities*, pages 219–246. Routledge.

Beth Dementi-Leonard and Perry Gilmore. 1999. Language revitalization and identity in social context: A community-based athabascan language preservation project in western interior alaska. *Anthropology & Education Quarterly*, 30(1):37–55.

Clarissa Forbes. 2023. 42 tsimshianic. *The Languages and Linguistics of Indigenous North America: A Comprehensive Guide, Vol. 2*, 13:985.

Lenore A Grenoble. 2018. Arctic indigenous languages: Vitality and revitalization. In *The Routledge handbook of language revitalization*, pages 345–354. Routledge.

Lenore A Grenoble and Vanda B Ignatieva. 2024. Tracking and unlocking the past: Documentation of arctic indigenous languages. In *Library and Information Sciences in Arctic and Northern Studies*, pages 191–207. Springer.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, pages 587–604.

Benjamin Hunt, Emily Chen, Sylvia LR Schreiner, and Lane Schwartz. 2019. Community lexical access for an endangered polysynthetic language: An electronic dictionary for st. lawrence island yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 122–126.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-english dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143.

Indians.org. 2025. Ahtna kenaege' proficiency project.

Meghan Jennings. 2024. The impact of language revitalization efforts on indigenous cultural practices: A case study of the tahltan, cherokee, and lakota nations. *Genocide Studies International*, page e20230021.

Anne M Jensen. 2020. Critical information for the study of ecodynamics and socio-natural systems: Rescuing endangered heritage and data from arctic alaskan coastal sites. *Quaternary International*, 549:227–238.

Wei Jia. 2024. Indigenous language revitalization and preservation in canada: Strategies and innovations. *International Journal of Languages, Literature and Linguistics*, 10(1):97–102.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.

Michael Krauss. 1996. Status of native american language endangerment. *Stabilizing indigenous languages*, pages 16–21.

Michael E Krauss. 1986. A survey of major alaskan language types. In *Language Typology 1985: Papers from the Linguistic Typology Symposium, Moscow, 9 13 Dec. 1985*, volume 47, page 169. John Benjamins Publishing.

Michael E Krauss. 2007. Native languages of alaska. *The vanishing languages of the Pacific Rim*, 406:417.

Haida Languages. 2025. Basic phrases in haida.

Jackline U Lidubwi and John O Ndavula. 2025. Revitalising endangered languages through social media: A case study of olunyore language preservation through facebook in kenya. In *Decolonising Digital Media and Indigenisation of Participatory Epistemologies*, pages 163–174. Routledge.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Joel Barry Martineau. 2002. *Islands at the boundary of the world: Changing representations of Haida Gwaii, 1774-2001*. Ph.D. thesis, University of British Columbia.

Angelina McMillan-Major. 2023. *Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities*. Ph.D. thesis, University of Washington.

Marianne Mithun. 2024. Central alaskan yup'ik. *Clause Chaining in the Languages of the World*, page 370.

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Alutiiq Museum. 2025. Alutiiq grammar: An overview.

Dr Marriam Bashir Naeem-ur Rehman and Ghulam Abbas. 2024. A case study of endangered vocabulary in chilasi dialect of the shina language. *Harf-o-Sukhan*, 8(1):894–909.

Sebastian Nordhoff, Siri Tuttle, and Olga Lovick. 2016. The alaskan athabascan grammar database. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3286–3290, Portorož, Slovenia. European Language Resources Association (ELRA).

University of Alaska Fairbanks. 2025. Useful dena'ina expressions.

Brigitte Pakendorf. 2024. The dynamics of language endangerment: A comparative study. *Sibirica*, 23(1):32–65.

Anastasia Panova. 2024. The chukchi influence on chaplinski yupik: A case study of personal names. *Journal of Language Contact*, 17(1):218–245.

Nicholas J Reo, Sigvanna Meghan Topkok, Nicole Kanayurak, James N Stanford, David A Peterson, and Lindsay J Whaley. 2019. Environmental change and sustainability of indigenous languages in northern alaska. *Arctic*, 72(3):215–228.

Ashleigh Surma and Christina L Truong. 2023. 35 digital tools for language revitalization. *The Languages and Linguistics of Indigenous North America: A Comprehensive Guide, Vol. 2*, 13:789.

Ivory Yang, Weicheng Ma, Carlos Guerrero Alvarez, William Dinauer, and Soroush Vosoughi. 2025a. What is it? towards a generalizable native american language identification system. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 105–111.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025b. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025c. Is it Navajo? accurate language detection for endangered athabaskan languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 277–284.

A Evaluation Metrics Details

This section provides detailed mathematical formulations for the evaluation metrics used to assess the impact of attention head pruning on the XLM-RoBERTa classifier.

A.1 Accuracy

Accuracy is the proportion of correctly classified instances out of the total samples. Mathematically, it is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{M} 1(\hat{y}_i == y_i)}{M},$$

where M is the total number of samples, $\hat{Y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_M]$ is the list of predicted classes, and $Y = [y_1, y_2, ..., y_M]$ is the list of true labels.

A.2 Macro-F1 and Micro-F1

The F1 score is a harmonic mean of precision and recall. For each class c, precision and recall are defined as:

$$P_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \quad R_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}.$$

The Macro-F1 score averages the F1 scores across all classes:

$$\text{Macro-F1} = \frac{1}{N} \sum_{c=1}^{N} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}.$$

Micro-F1 aggregates all instances before computing the F1 score:

$$\text{Micro-F1} = \frac{2 \sum_c \text{TP}_c}{2 \sum_c \text{TP}_c + \sum_c \text{FP}_c + \sum_c \text{FN}_c}.$$

A.3 Cross-Entropy Loss

Cross-entropy loss penalizes incorrect classifications by considering the predicted probability assigned to the true class:

$$\text{Cross-Entropy} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{N} y_{i,c} \log p_{i,c},$$

where $p_{i,c}$ is the predicted probability for class c, and $y_{i,c}$ is the true label for sample i.

B Real Life Applicability

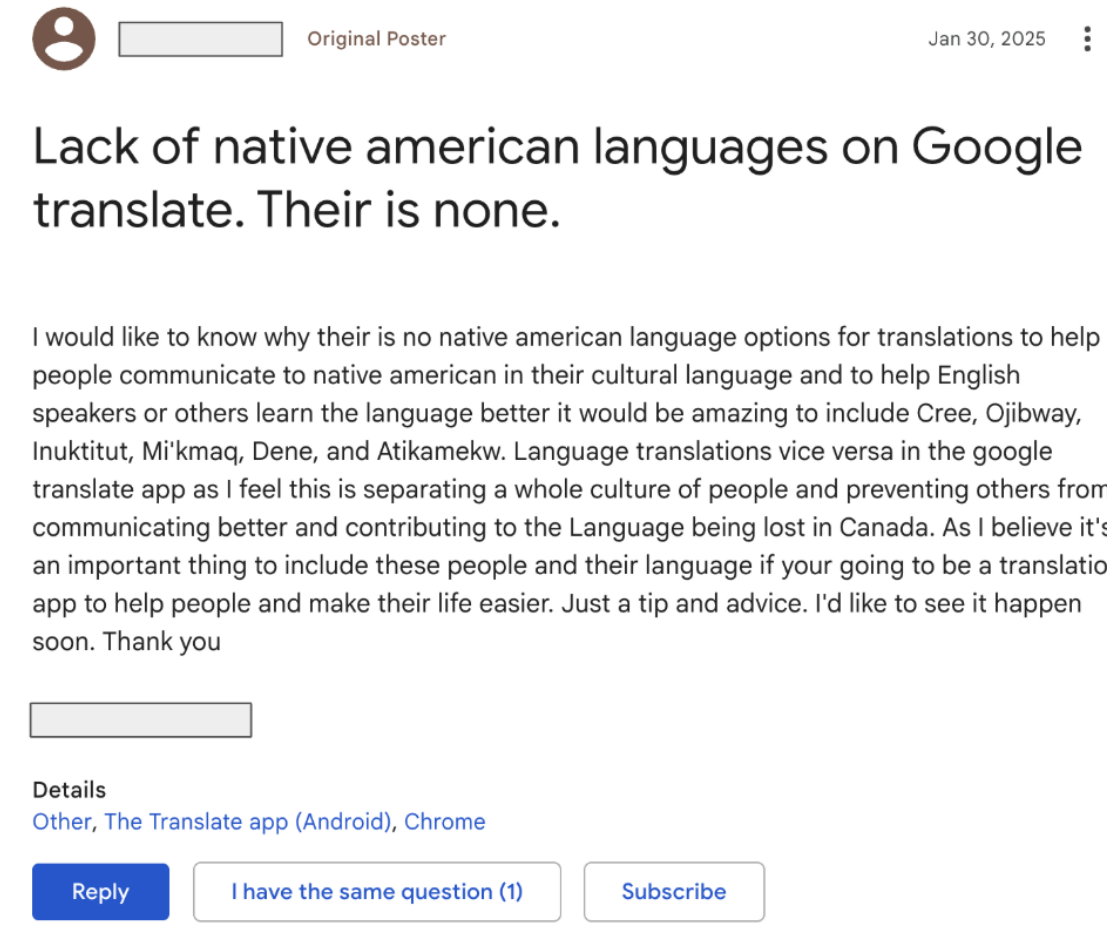
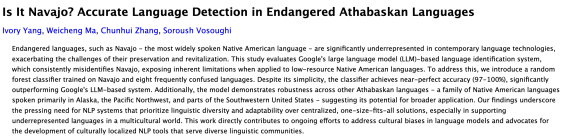


Figure 9: Community member on Google Translate Help Forum questioning the lack of support for Native American Languages. Name is covered for privacy, although this was taken from a public domain.

Figure 10: NBC News article on Indigenous engineers incorporating AI

C Simple Approaches Work Best

Recent NLP research has shown that the simplest approaches often work best. Minimal interventions can yield strong results, making NLP more accessible and reducing reliance on costly resources. By embracing simplicity, we democratize language technology, ensuring that impactful solutions are within reach for all communities. Our work follows this tradition, proving that progress in NLP isn't about complexity, but about what truly works.

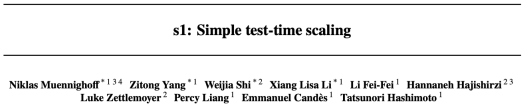


Is It Navajo? Accurate Language Detection in Endangered Athabaskan Languages

Ivory Yang, Weicheng Ma, Chunhui Zhang, Soroush Vosoughi

Endangered languages, such as Navajo – the most widely spoken Native American language – are significantly underrepresented in contemporary language technologies, exacerbating the challenges of their preservation and revitalization. This study evaluates Google's large language model (LLM)-based language identification system, which consistently misidentifies Navajo, exposing inherent limitations when applied to low-resource Native American languages. To address this, we introduce a random forest classifier trained on Navajo and eight frequently confused languages. Despite its simplicity, the classifier achieves near-perfect accuracy (97–100%), significantly outperforming Google's LLM-based system. Additionally, the model demonstrates robustness across other Athabaskan languages – a family of Native American languages spoken primarily in Alaska, the Pacific Northwest, and parts of the Southwestern United States – suggesting its potential for broader application. Our findings underscore the pressing need for NLP systems that prioritize linguistic diversity and adaptability over centralized, one-size-fits-all solutions, especially in supporting underrepresented languages in a multicultural world. This work directly contributes to ongoing efforts to address cultural biases in language models and advocates for the development of culturally localized NLP tools that serve diverse linguistic communities.

Figure 11: NAACL 2025 paper (Yang et al., 2025c) on a simple yet highly effective Random Forest classifier for Navajo, demonstrating its ability to generalize to other endangered Athabaskan languages within the same linguistic family.



s1: Simple test-time scaling

Niklas Muennighoff [*1 3 4] Zitong Yang [*1] Weijia Shi [*2] Xiang Lisa Li [*1] Li Fei-Fei [1] Hannaneh Hajishirzi [2 3]
Luke Zettlemoyer [2] Percy Liang [1] Emmanuel Candès [1] Tatsunori Hashimoto [1]

Figure 12: A recent Stanford paper (Muennighoff et al., 2025) explores the impact of appending "wait" in prompts, an incredibly simple approach that, with just 30 minutes of training and $30, achieved performance on par with o1 models.

D LID System Coverage

LID Type	Native Alaskan Language	Notes
Google LangID	–	97 languages
GlotLID	Central Yupik, Gwich'in, Haida, Inupiaq	2102 languages
MadLAD	–	419 languages
FastText-LID	–	176 languages
WhatLang	–	69 languages
LangDetect	–	55 languages

Table 3: Native Alaskan language coverage across popular LID systems.

E Prompt Details

You are a linguistics expert who knows every single language that exists in this world. What language is this sentence in? Sentence: sentence. Reply with only the language itself and nothing else.

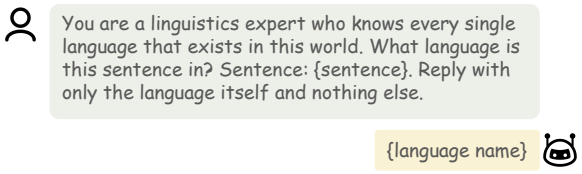


Figure 13: The textual format for zero-shot promoting.

F Generalization to Other Languages

We have conducted preliminary experiments on other small-scale endangered languages, including Native American Apache, with promising early-stage results. These initial findings suggest that our generalizable approach could extend beyond Native Alaskan languages, reinforcing the potential for scalable, data-efficient language identification in other low-resource linguistic settings.

G Cultural Background

Language is deeply intertwined with culture, history, and identity. We believe it is important to share the symbols, stories and imagery and reflect the rich traditions of Native Alaskan communities, many of which are featured in our work above. By including these references, we seek to provide readers with a broader cultural context. Understanding language goes beyond syntax and semantics, it is a gateway to the histories, values, and worldviews of the people who speak it.

H Interview Transcripts

Below is a transcript of one of our interviews with a member of the Native Alaskan community, with their identity and personal information redacted. We received their full permission to release this transcript in its entirety.

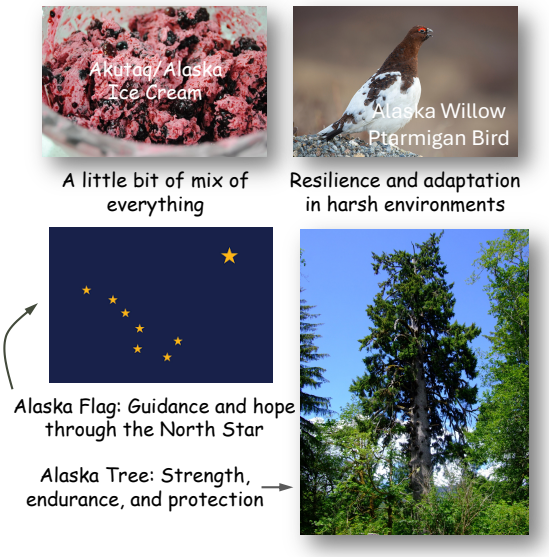
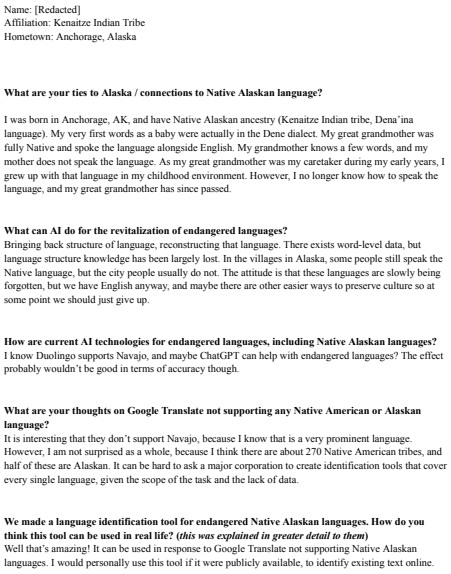


Figure 14: Alaska symbols: Akutaq[1], willow ptarmigan bird[2], Alaska flag, and Sitka Spruce[3].



Name: [Redacted]
Affiliation: Kenaitze Indian Tribe
Hometown: Anchorage, Alaska

What are your ties to Alaska / connections to Native Alaskan language?

I was born in Anchorage, AK, and have Native Alaskan ancestry (Kenaitze Indian tribe, Dena'ina language). My very first words as a baby were actually in the Dene dialect. My great grandmother was fully Native and spoke the language alongside English. My grandmother knows a few words, and my mother does not speak the language. As my great grandmother was my caretaker during my early years, I grew up with that language in my childhood environment. However, I no longer know how to speak the language, and my great grandmother has since passed.

What can AI do for the revitalization of endangered languages?
Bringing back structure of language, reconstructing that language. There exists word-level data, but language structure knowledge has been largely lost. In the villages in Alaska, some people still speak the Native language, but the city people usually do not. The attitude is that these languages are slowly being forgotten, but we have English anyway, and maybe there are other easier ways to preserve culture so at some point we should just give up.

How are current AI technologies for endangered languages, including Native Alaskan languages?
I know Duolingo supports Navajo, and maybe ChatGPT can help with endangered languages? The effect probably wouldn't be good in terms of accuracy though.

What are your thoughts on Google Translate not supporting any Native American or Alaskan language?
It is interesting that they don't support Navajo, because I know that is a very prominent language. However, I am not surprised as a whole, because I think there are about 270 Native American tribes, and half of these are Alaskan. It can be hard to ask a major corporation to create identification tools that cover every single language, given the scope of the task and the lack of data.

We made a language identification tool for endangered Native Alaskan languages. How do you think this tool can be used in real life? (*this was explained in greater detail to them*)
Well that's amazing! It can be used in response to Google Translate not supporting Native Alaskan languages. I would personally use this tool if it were publicly available, to identify existing text online.

Figure 15: Interview Transcript with a member of the Native Alaskan community.

[1] https://sweetstateofmine.blogspot.com/2011/05/alaska-akutaq-eskimo-ice-cream.html
[2] https://www.alaskaphotographics.com/alaska-photo-articles/ptarmigan-photos/
[3] https://nativeplantspnw.com/sitka-spruce-picea-sitchensis/

I K-Fold Sampling

To address concerns regarding the use of a small fixed test set, we conduct a 5-fold cross-validation using our complete dataset (100 sentences per language across 20 Native Alaskan languages). For each fold, we use up to 40 examples per language in the few-shot prompt and evaluated on the remainder. The folds rotated the held-out test set to ensure coverage of the entire dataset.

Language	**Accuracy**	**Language**	**Accuracy**
Siberian Yupik	0.960	Lower Tanana	0.970
Alutiiq	0.920	Central Alaskan Yup'ik	0.980
Deg Xinag	0.970	Ahtna	0.980
Gwich'in	0.990	Tlingit	0.990
Haida	1.000	Aleut	0.950
Holikachuk	0.930	Inupiaq	0.980
Eyak	1.000	Tsimshian	0.900
Tanacross	0.950	Dena'ina	0.990
Hän	0.950	Upper Tanana	0.990
Upper Kuskokwim	0.930	Koyukon	0.980

Table 4: Accuracy scores from 5-fold cross-validation using GPT-4o few-shot prompts.