IndigiEval: Evaluating LLMs in North American Indigenous Languages

Julia Mainzinger
Charles Darwin University
Darwin, Australia

Jacqueline Brixey
University of Wisconsin - Madison
Madison, WI, USA

Abstract

This paper presents IndigiEval, a framework for evaluating the language and cultural proficiency of several commercially available large language models (LLMs) across five North American Indigenous languages (Mvskoke, Choctaw, Cherokee, Cheyenne, and Hawaiian). This framework is a qualitative evaluation method intended for communities with small speaker populations to be able to critically evaluate LLM performance with minimal data and human effort. IndigiEval includes tasks such as answering cultural questions, translation, text generation, and speech recognition. The results of our experiments indicate that no currently available LLM performs well across all evaluation categories, and that LLMs frequently hallucinate orthographies, grammatical structures, cultural knowledge, and vocabulary for all languages and cultures considered. Our proposed evaluation framework is not intended as a comprehensive score, but rather a qualitative and flexible framework to inform language communities about a given LLM's potential as a resource, since each language has unique environments, strengths, and availability of resources.

1 Introduction

With the increasing capabilities of large language models (LLMs) in some low-resource languages, there has been growing interest in how AI, and specifically LLMs might support language revitalization efforts (Moshagen et al., 2024). In the North American context, limited access to first-language speakers leads people to turn to technology to fill gaps (Meighan, 2024).

Although LLMs demonstrate impressive performance in English and other majority languages, the same is not true for minority, Indigenous, and endangered languages (Choudhury, 2023; Stap and Araabi, 2023). Even in English, LLMs are known to make things up, providing both truth and fallacies without qualification (Hicks et al., 2024). These hallucinations and errors are magnified in languages that are underrepresented in the training data (Song et al., 2026a).

Concerns about accuracy are not unique to endangered languages; similar risks also exist in other high-stakes domains where accuracy is essential (Elliott et al., 2025). This includes domains such as law and medicine (Cheong et al., 2024; Quttainah et al., 2024; Singhal et al., 2025), where incorrect information can have profound real-world consequences. Accuracy in teaching language to second-language learners is likewise essential, as learners have no way to judge when mistakes occur, making them vulnerable to learning errors permanently.

Recent scholarship calls for a re-centering of AI research on Indigenous practices, including knowledge-making and storytelling (Lewis et al., 2024), as well as inter-generational language transmission and language learning applications (Hinton, 2013; Pitawanakwat, 2018; Neubig et al., 2020). There is also concern that people will turn to AI over local knowledge holders – the "Ask your Auntie, Not AI" campaign[1] has been created to challenge the idea that AI can replace elders.

Despite these concerns, limited work has been conducted on how to systematically evaluate LLM performance in truly low-resource Indigenous language contexts. Existing benchmarking approaches typically rely on large-scale datasets, such as thousands of evaluation questions or extensive human feedback (Vayani et al., 2025; Myung et al., 2024; Zhang et al., 2023a). For small speaker communities, generating such datasets can be prohibitively labor- and resource-intensive (Wiechetek et al., 2024; Pitawanakwat,

[1] https://www.honorearth.org/

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 82–94

2018). Moreover, given the limited demonstrated benefits of LLMs in low-resource settings, these forms of evaluation may be perceived as extractive, particularly when they require substantial contributions from fluent speakers without clear reciprocity.

For these reasons, we see value in demonstrating a small-scale evaluation that serve as a model for how communities might assess the limitations of LLMs and make informed decisions about whether and how these technologies should be used. In this work, we propose IndigiEval, a framework for exploring language and cultural capabilities and limitations of LLMs in North American Indigenous contexts. Toward this end, we design our framework according to the following principles:

- **Small-scale:** Fits within resource constraints without requiring large-scale datasets
- **Non-extractive:** Minimizes human annotation demands in order to reduce burden on small speaker communities
- **Accessible:** Evaluations can be carried out by community members or those familiar with the language situation without necessarily requiring fluent speakers

As illustrated in Figure 1, we evaluate LLMs across four categories: cultural knowledge, translation, text generation, and speech recognition. We choose tasks that leverage the types of materials commonly available in Indigenous language documentation and revitalization work. However, we also recognize that each community has its own strengths and priorities; as such, this framework is not one-size-fits-all, but can be adapted by other communities to assess whether LLM-based tools align with local needs (Zhang et al., 2022; Liu et al., 2022).

The authors of this work are citizens of the Muscogee Nation (first author) and the Choctaw Nation of Oklahoma (second author). Julia Mainzinger collaborates with the College of the Muscogee Nation on language revitalization projects. Jacqueline Brixey is a learner of the Choctaw language and has built language technology that supports revitalization efforts.

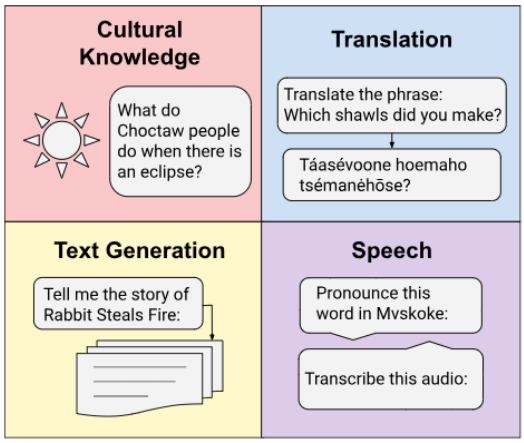

Figure 1: Evaluation categories of IndigiEval

2 Related Work

Several prior benchmarks assess LLM performance for lower-resource languages. Many of these benchmarks use questions from standardized exams, such as grade-school exams (Zhang et al., 2023a) or national language institutes (Song et al., 2026b). Results vary by language and task: one study has found that LLMs could not pass any exam beyond a primary school level in Indonesian (Koto et al., 2023), while another has shown how models can competently handle understanding but not generation tasks at the high school level in Latvian (Darģis et al., 2024).

Benchmark tasks to assess language proficiency often include vocabulary tests (Song et al., 2026b), short text generation (Chaka, 2024), and translation both to and from a given language (Zhang et al., 2023b). Task formats from previous work often favor multiple-choice formats (Darģis et al., 2024; Myung et al., 2024) and open-ended questions (Chaka, 2024), with only a few benchmarks utilizing true/false question formats (Vayani et al., 2025).

The benchmarks All Languages Matter and BLEnD probe LLMs for cultural competency and language proficiency in a diverse array of languages (Myung et al., 2024; Vayani et al., 2025). Other studies target specific contexts, such as FilBench, which evaluates LLM cultural knowledge and language proficiency in Filipino languages (Miranda et al., 2025), and ProverbEval, which tests LLMs for knowledge of proverbs in five African languages (Azime et al., 2025).

Many Indigenous American languages lack standardized tests or large collections of question/answer pairs that could be used in this man-

Language	ISO-3	Speakers	WKP	CC
Mvskoke	mus	300	0	0
Choctaw	cho	1,000	0	0
Cheyenne	chy	344	721	0
Cherokee	chr	2,000	1,003	1,025
Hawaiian	haw	7,000	2,968	6,343

Table 1: Languages included in this study, with speaker count, Wikipedia (WKP) article count, and Common Crawl (CC) count.

ner, and many of our communities lack the manpower and resources to carry out large-scale human evaluation. Our work therefore differs from previous work in that we present a small-scale qualitative evaluation framework. This framework provides representative tasks that could be feasible with small amounts of data, serving as an example of how LLM evaluation might be performed without burdening small speaker populations.

3 Languages

As we are most familiar with Choctaw (cho) and Mvskoke (mus), we evaluate LLMs across all categories for these two languages. We include three additional languages—Cheyenne (chy), Cherokee (chr), and Hawaiian (haw)—in the language proficiency section, as we can reference expertise in language documentation resources to assess responses. We selected the languages because they have similar speaker counts and representation in public datasets (Table 1). All five languages considered in our experiments are indigenous to the present-day United States, and all are endangered.

Choctaw. The Choctaw language is spoken by the Choctaws, who are the third most populous US tribal group, with approximately 223,000 people identifying as Choctaw[2]. However, the language has endangered status (Simons and Fennig, 2018), and it is estimated that there are fewer than 1,000 fluent speakers in the Choctaw Nation of Oklahoma (Rogers, 2021).

The Choctaw language is part of the Muskogean language family (Haas, 1979), and has subject-object-verb word order. The language is relatively well-documented, with numerous grammars (Byington, 1870), dictionaries (The Choctaw Nation of Oklahoma Dictionary Committee, 2016; of Choctaw Indians, 2026), and printed learning materials. We refer to these materials for our experiments.

Mvskoke. The Mvskoke (also known as Muscogee or Creek) language is spoken by members of the Muscogee and Seminole tribes. It is estimated that fewer than 300 first-language speakers remain, and nearly all are over the age of 60[3].

The language is synthetic and agglutinative, with a traditional orthography of 20 Latin letters. The orthography is relatively transparent and allows for spelling variations (Martin, 2011). We reference the dictionary by Martin and Mauldin (2000) and a collection of texts by Haas et al. (2015).

Cheyenne. Cheyenne is an Algonquian language, spoken by the Cheyenne people, a Great Plains Tribe (Leiker and Powers, 2011). There are two federally recognized tribes, now located in Montana and Oklahoma[4]. The language is highly agglutinative and polysynthetic (Badhorse, 2023). There are approximately 300 fluent speakers today (Littlebear, 2024). We draw vocabulary from Fisher et al. (2023) and sentences from Leman (1980).

Cherokee. The Cherokee people originate from the Southeastern United States, from North Carolina to Georgia (Justice, 2025). They were removed largely to Oklahoma, with one federally recognized tribe remaining in North Carolina [5]. The language is Iroquoian (Julian, 2010), with a unique 85-character syllabary[6]. There are an estimated 2,000 first-language Cherokee speakers (Zhang et al., 2022). We reference vocabulary from a collection of Cherokee dictionaries[7], and collect sentences from Howard and Eby (1990).

Hawaiian. Hawaiian is spoken by people indigenous to the Hawaiian Islands. It is in the Eastern Polynesian branch of the Austronesian language family (Parker Jones, 2018). Although there was no large-scale removal from ancestral

[2] https://archive.ncai.org/tribal-vawa/sdvcj-today/the-choctaw-nation-in-oklahoma

[3] This estimate is from personal communication with a member of Ekvn-Yefolecv, a community of Mvskoke people.

[4] http://www.cheyennenation.com and https://www.bia.gov/regional-offices/southern-plains/concho-agency

[5] https://www.doi.gov/tribes/cherokee

[6] https://georgiahistory.com/wp-content/uploads/2023/11/NIE-2017-web.pdf

[7] https://www.cherokeedictionary.net/about

lands as with many US continental tribes, linguistic loss occurred due to 'English-only' language ideology, social injustice, and population decline from disease (Brenzinger and Heinrich, 2013). Revitalization of the language began in the 1970s, and through school and immersion programs, there are now around 7,000 fluent speakers, including new first-language speakers (Brenzinger and Heinrich, 2013). We reference vocabulary from Pukui and Elbert (2003) and Leo and Kuamo o (2003), and sentences from Bardwell et al. (2020).

4 Evaluation categories

The tasks included in this framework are not meant to be comprehensive or exclusive. Rather, they are intended to demonstrate gaps in language and cultural competency of the LLMs. Failures observed at a small scale can indicate limitations in the ability to support community language needs. We present one task that evaluates cultural knowledge, and three tasks that evaluate language proficiency.

4.1 Cultural Knowledge

Studies have found that LLMs often exhibit Western biases (Myung et al., 2024). These types of biases can harmfully reproduce stereotypes towards Indigenous people with cultures distinct from Western ones (Hanson, 2025), or propagate misinformation about Indigenous communities.

To evaluate LLMs' cultural knowledge, we developed 10 multiple-choice questions in English about Mvskoke and Choctaw culture. These questions assess knowledge and values, including material knowledge, historical figures, events, and traditional food. Each multiple-choice question has four choices presented with only one correct answer. We determined the answer to these questions based on our own knowledge or by referencing published sources. We did not pose any questions that would be unacceptable to share with those outside of our communities. As both of our communities are fairly open to outsiders and many of the cultural details asked about are documented online, we hypothesized that the LLMs would be able to correctly answer many of these questions.

4.2 Language Proficiency

The language proficiency category comprises machine translation, vocabulary questions, text generation, and speech tasks.

4.2.1 Machine Translation

Previous research has shown that LLM-based machine translation for low-resource languages does not outperform traditional neural machine translation (Ebrahimi et al., 2022; Stap and Araabi, 2023; Robinson et al., 2023). Because almost all commercial LLMs do not publicly share their datasets, it is unclear exactly how much language-specific data is contained in a given model. However, performance is often directly correlated with the number of Wikipedia pages in a given language (Robinson et al., 2023). Additionally, for languages known to have low representation in the LLM's training data, the output is almost entirely nonsensical (Zhang et al., 2022). Table 1 provides the resource counts in the evaluation languages. While Mvskoke and Choctaw do not have any pages entirely in these languages on Wikipedia, we have found that many LLMs nevertheless attempt to translate to and from them.

In this task, we selected 10 sentences for translation in each of the five languages. The ten sentences display a range of syntactical structures and vocabulary. We utilize chrF++ as our scoring system.

It should be noted that we are not suggesting that LLMs should be used for Indigenous language translation over traditional machine translation methods. Here, we use translation as a task because it has a well-established scoring system, and we can utilize language documentation for high-quality references. This provides a more objective measure of language proficiency. We also anticipate that many language learners might ask for translation assistance from an LLM.

4.2.2 Vocabulary

In this task, we prompt the LLMs to translate single English terms into each language. While this is a sub-task of MT, prompting for isolated words enables us to assess specific vocabulary, such as modern vocabulary that may not be formalized in digitized corpora. Additionally, it helps us to identify hallucinated words more easily.

We translated 50 English words into each language by referencing the relevant dictionaries. We also consider several additional words that may not have a formal translation, such as "microwave" and "yogurt". Some of these words have locally-defined terms but are not present in

dictionaries. We do not include these as part of the scoring, but rather to observe the behavior when a model is asked for a word that may not exist in the language.

We prompt each model for a translation of the given English word. During scoring, we also reference these same dictionaries, allowing for synonyms or closely related terms to be marked correct.

4.2.3 Text Generation

Prompting for open-ended text generation allows us to consider the capabilities of the language model of the LLM by evaluating the sensibility, errors, and hallucination of the generated text (Chaka, 2024). In this task, we prompt the LLMs to tell a traditional story in Mvskoke and Choctaw and then compare their responses with published versions. The published versions are part of language documentation that has been written or reviewed by fluent speakers. We examine the overall fluency and contrast the outputs with the language documentation. We only prompt for one story for each model per language, as the stories are several paragraphs long, and the long-form generation gives a thorough sense of the models' command of the language.

4.2.4 Speech

Speech technologies can be meaningful tools for low-resource languages. Automatic speech recognition (ASR) is a valuable technology, as it can be used for automatic captioning, voice typing, and improving transcription efficiency (Ćavar et al., 2016). Additionally, in the context of North American Indigenous languages, orthographies are often varied, and standardization may not be consistently adopted. As a result, text-to-speech (TTS) can be helpful both for language learners and for overcoming challenges in writing systems. To this end, we include **ASR** and **TTS** tasks in our evaluation framework.

5 Setup

5.1 Prompting

To test our framework, we developed a Python script for each evaluation category that makes API calls to each of the LLMs. Each API call includes an instruction and a prompt. The instruction gives the LLM context for the prompt, such as "You are a helpful assistant knowledgeable in the [target] language." This is included along with the prompt for each task. The prompts are zero-shot with no web search functionality.

5.2 Models

For the text-based tasks, we tested large commercially available LLMs with the highest performance at the time of testing: OpenAI's GPT gpt-5.2-2025-12-11 (Singh et al., 2025), DeepSeek deepseek-reasoner (Guo et al., 2025), Anthropic's Claude claude-opus-4-6 (Anthropic, 2026), and Google's Gemini gemini-3.1-pro-preview (Team et al., 2025). For speech tasks, we tested gemini-3.1-pro-preview and gpt-4o-transcribe[8]. We leave it to future work to test other LLMs, including open source models.

6 Results

6.1 Cultural Knowledge

All of the LLMs tested are largely able to answer basic questions about popular sports, foods, and traditional clothing in the Mvskoke and Choctaw cultures. The results are in Table 2. GPT-5.2 and DeepSeek make more mistakes overall than Claude and Gemini, missing questions such as "How did clans get their names in Mvskoke culture?" and "What is the name of a traditional Choctaw dance?"

Interestingly, Claude and Gemini can correctly answer detailed and specific questions about stories documented in the books Haas et al. (2015) and Gouge et al. (2004), showing that these resources, which are two of the most extensive collections of written Mvskoke language documentation, are likely contained in the training data for these two models.

The most-missed question, which all four models missed, is "In the Dawes Commission short film by Bob Hicks, what does the grandmother tell the little girl she used to use for dancing?" This film is publicly available on YouTube, and the majority of it is in Mvskoke with English subtitles. The second most common mistake was a tie for a question about a Mvskoke historical figure, and a question about when the Okla Chahta Clan of California holds its annual gathering. These mistakes demonstrate that despite the vast training data, these LLMs are not infallible sources of cultural information.

[8]https://developers.openai.com/api/docs/models/gpt-4o-transcribe

	GPT	DS	Claude	Gemini
mus	4	5	9	8
cho	9	8	9	9
Total	13	13	18	17
	65%	65%	90%	85%

Table 2: Number and percent of correct answers on 20 multiple-choice cultural knowledge questions.

6.2 Machine Translation

We performed a zero-shot evaluation on 10 sentences per language. All of the examples come directly from language documentation or language learning textbooks and were thus produced and/or verified by fluent speakers. Because of this, the MT dataset is not multiparallel (as the vocab quiz is), but rather high-quality samples of each language. An example output of the task is given in Table 3. Outputs are evaluated by the chrF++ score with β=2 (Popović, 2015).

Overall, Gemini outperforms every other model. Our findings agree with Zhu et al. (2024) and Enis and Hopkins (2024) that the LLMs generally perform better from the language to English. The results for this task, illustrated in Figure 2, can be summed up by a few broader points.

The LLMs have data in every language. Even though Mvskoke and Choctaw are not listed in Common Crawl or other public datasets, it is clear that all the LLMs have at least minimal training data for these languages, as they produced text in the correct orthographies. In the lower-resourced languages, the responses range from complete gibberish (GPT-5.2) to somewhat understandable sentences (Gemini).

Language underrepresented online. Hawaiian's representation in publicly accessible datasets such as Common Crawl and Wikipedia (see Table 1) shows that more representation in the training data via larger representation online increases accuracy, as every model performs best for Hawaiian in the language -> X direction. The lesser-represented languages generally perform more poorly, with some variability in performance perhaps due to linguistic complexity or small sample size.

Nonsensical output. Even though chrF++ scores for our evaluation are on par with what would be expected from a baseline neural model (De Gibert et al., 2025), one issue is that LLMs often generate nonsensical output when producing

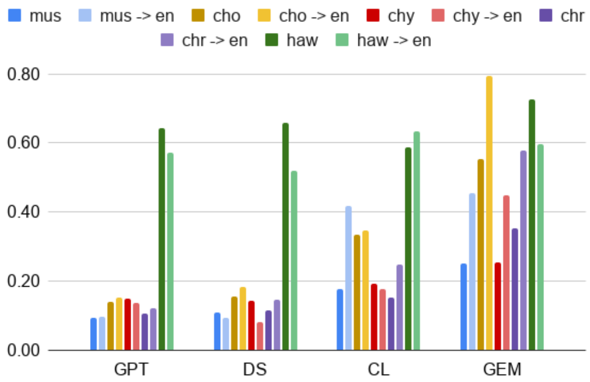

Figure 2: MT Results: chrF++ scores for translations in two directions, En -> X (darker color) and X -> En (lighter color), for four models (GPT 5.2, DeepSeek Reasoner, Claude, and Gemini 3.1 Pro).

translations from English into the given language. Table 3 shows an example of hallucinated words. However, when translating in the direction from the Indigenous language to English, there are no nonsensical words.

6.3 Vocabulary Quiz

We prompt the models to translate 50 words from English into the target language. We hand-graded the responses by looking up each response in the corresponding dictionary. Synonyms or commonly accepted alternate spellings are counted as correct answers. Because we are only familiar with Mvskoke and Choctaw, we are unable to provide value judgments on alternative answers in languages not represented in the dictionaries. Because of this, the results for Mvskoke and Choctaw may be higher than those of the other three languages. Nonetheless, some interesting trends can be observed. Figure 3 shows the results of prompting the models for individual words. Overall, Gemini Pro 3.1 outperforms other models in every language, getting a perfect score for Hawaiian.

Nonsense and unrelated words. Many responses, especially in the lower-performing models, are simply jumbles of letters in the given orthography. Other responses are words that do exist in the language, but are unrelated to the term. For example, Claude provides Cherokee "ᎤᏬᏚᎯ" (pretty) for "to know".

Inventing modern words. We also separately tested for modern words that are less formalized in the languages, such as "computer", "microwave", and "yogurt". These are not counted as part of the vocab quiz, but are intended to ex-

Source:	Yv likat mēkken omat hēr̄ēs.
Ref Translation:	This one sitting here would be a good king.
Mus->En:	It would be good if this one sitting here were king.
En->Mus:	Heyv likat meco here tvres.
Back translation:	This one sitting here [unknown] good will be.

Table 3: Example output from Gemini for Mvskoke translation. The word "meco" is nonsensical, and the sentence is grammatically incomplete.

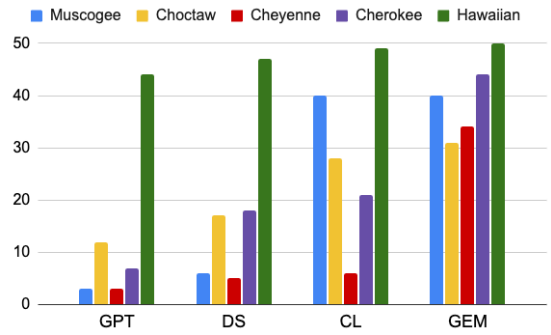

Figure 3: Vocabulary Quiz: Number of correct answers out of 50 for five models (GPT 5.2, DeepSeek Reasoner, Claude, and Gemini 3.1 Pro).

amine model behavior. Some of these terms have been normalized by community usage. A few of the words have published translations, such as Cherokee's "ᎠᎦᏙᎯᏍᏗ ᎠᏍᎪᎯᏰᎯᏍᏗᎩ" for "computer". No model was able to produce this term. For words that do not exist in a language, we found that the LLM will still try to come up with something. Sometimes the inventions, such as "esvhonkvtkv" (*thing you count with*) as Mvskoke for "computer", are not ideal but still reasonable. Others are completely nonsensical letter sequences.

Answer refusal. A notable behavior is Claude's refusal to provide responses for words with very low certainty. For example, when prompted to provide the word for yogurt, Claude responded "I'm not aware of a specific, established Muscogee (Creek) word for 'yogurt.' This is a modern food item that may not have a standardized translation in the Muscogee language. You may want to consult the Muscogee (Creek) Nation's language department for an official term." In terms of preventing hallucinations and errors, this might be preferred over providing low-quality translations. The only answer refusals by Claude were contained to the "modern" words, and did not affect the vocabulary quiz results.

6.4 Text Generation

Text generation performance is similar for Mvskoke and Choctaw languages for each LLM. GPT-5.2 and DeepSeek make use of some related terms, but the grammar and word choices overall are largely not understandable. Claude and Gemini can use both related terms and the correct grammar form (distant past) to conjugate verbs. Furthermore, some general aspects of the stories align with traditional stories, such as similar characters and storylines. Even so, the outputs contain many hallucinated words and grammatical mistakes, making the text difficult to read and, at times, meaningless. See Table 4 for an example of text generation in Mvskoke.

6.5 Speech

DeepSeek and Claude do not natively support TTS or ASR. Thus, we conduct the speech category evaluation only for GPT-4o and Gemini 3.1 Pro.

ASR. We test two short audio clips for each language Mvskoke and Choctaw. The samples are of clear, read speech recorded during language documentation (Gouge et al., 2004; Haas et al., 2015). The word error rates and character error rates are given in Table 5. Further details can be found in Appendix A. The WERs for both models are not better than previously developed ASR models in either language (Brixey and Traum, 2022; Mainzinger and Levow, 2024). One notable finding is that the models do not demonstrate competence in using the languages' orthographies for this task, unlike results found in other evaluation categories.

TTS. We prompt each model with a short phrase in Mvskoke and Choctaw to be said aloud. We provide the target language in both the instruction and the prompt. Based on listening to the audio, we determine that GPT and Gemini are not able to produce audio that sounds like these languages. GPT does not remotely follow the ex-

Output	Back-translation
Hofonof, ponvttv sulkē omof, kaspes. Totkv sekon. Ponvttv vtekat kvapvtes. Uewv tvpvlv, este hopvye totkv ocvtes. Mv este totkv vcayecvtes. Cufe makvtes, "Totkv horkoparēs. Vnheckv herēs, ce!"	A long time ago, when animals many, [unknown, possibly an attempt at "cold"]. There is no fire. All the animals [unknown]. Across [misspelled] the water, person far away had fire. That person stored the fire. Rabbit said, "I will steal the fire. My appearance is good!"

Table 4: Sample of Gemini response to the prompt, "Tell me the story about Rabbit Steals Fire. Tell it in Muscogee, in the style of a Muscogee traditional story." In the traditional story, the rabbit travels across a great body of water to steal fire. There are several instances of hallucinated words (red), grammatical issues, and awkward wording. The word order is very English-like - for example, putting "the rabbit said" before the quote, when a Mvskoke person would always put it after.

	GPT		Gemini	
	WER	CER	WER	CER
mus	0.92	0.47	**0.91**	**0.34**
cho	**0.82**	**0.26**	0.98	0.35

Table 5: Word Error Rates (WER) and Character Error Rates (CER) for GPT 4o Transcribe and Gemini 3.1 Pro for the ASR task.

pected phonemes for the orthographies – most egregiously, it pronounces the Mvskoke 'v' as /v/ instead of the vowel /ə/. Gemini produces this vowel correctly but makes other phonetic mistakes that render the audio unintelligible.

7 Conclusions and Future Directions

In this work, we demonstrated tests that an Indigenous language community could perform to evaluate a given LLM's acceptable language proficiency in a specific language without extensive labor or large-scale datasets. While some of our samples for certain tests were limited, the tests included in our framework provide a holistic understanding of a given LLM's proficiency that could be generalizable to other communities. Studies that demand significant effort from communities with small speaker populations without offering reciprocal benefits can be perceived as extractive. Our framework, which can be conducted by a second-language speaker, is thus less extractive, as comparisons can be made with language documentation, and consultation with elders could be limited to a subset of output for review.

We present four categories of evaluation: cultural knowledge, language proficiency, text generation, and speech. Overall, the results reveal substantial limitations across all models for the five Indigenous languages considered. Cultural knowledge, especially, is incomplete and inconsistent across the four LLMs evaluated, which, for minority communities, is essential for preventing harmful stereotypes and misinformation. In terms of language proficiency, performance is consistently higher in Hawaiian than in the other languages. GPT and DeepSeek show very low proficiency in all languages except Hawaiian. Claude and Gemini show slightly better performance, but their outputs still contain frequent errors and hallucinations, especially in the four lower-resourced languages. Text generation for Mvskoke and Choctaw shows that LLMs can use correct orthography and some limited grammatical constructs. However, the outputs still exhibit frequent hallucinations and grammatical errors. We found that ASR underperforms when compared to traditional neural models. Finally, text-to-speech capabilities are not able to produce intelligible audio in the given languages.

These evaluation tasks are not meant to be comprehensive or exclusive. Rather, they are offered as a starting point for communities to adapt according to their own priorities and goals. Communities with more existing linguistic resources may be able to curate larger datasets for evaluation. More-resourced language groups may find that LLMs are performant enough to use as part of an end-to-end system. For example, performance in Hawaiian might be good enough for use with a RAG system, as demonstrated by the Kumu Connect implementation (Baker-Ramos et al., 2025).

Future work could include fine-tuning models to better estimate the extent of performance improvement achievable with limited data, although our findings suggest that existing models may already incorporate much of the available

printed data for the five languages we reviewed. More granular methods, such as quantifying the number of hallucinations and mistakes (for example, Chaka (2024)), could offer additional insight. At the same time, these directions raise ethical and practical considerations. Curating data for LLMs requires significant effort, often from fluent speakers whose time may be better spent on community-based activities such as teaching (Wiechetek et al., 2024). Moreover, hallucinations are an inherent limitation of current LLM architectures and are unlikely to be fully eliminated, even in high-resource languages (Hicks et al., 2024).

These findings suggest that while LLMs may offer limited utility in certain contexts, they may not be appropriate for endangered-language situations where second-language learners are unable to identify errors in AI-generated outputs. Careful, community-driven evaluation remains essential in determining whether and how these technologies should be used.

8 Ethical Considerations

By prompting the LLMs with data from language documentation, there is the risk that these models may store and use that data without permission. However, it is likely the LLMs already scrape most of the language documentation from online sources, without regard for copyright, intellectual property rights, or tribal data sovereignty.

At present, none of the AI companies that develop these LLMs offer methods for speakers of any language to correct language proficiency issues or cultural misunderstandings. Additionally, to our knowledge, our tribes were not contacted to be included in any current LLM versions, nor were any culturally appropriate, modern, and representative linguistic data requested from our communities by LLM companies for inclusion. Unregulated use of such data by AI companies risks undermining community language authorities, and may harm the vitality of Indigenous languages.

Finally, while we have designed this framework to be as least extractive as possible, there may be necessary involvement from elders or fluent speakers, which may distract from other important language revitalization activities.

Limitations

The findings presented here are not equally applicable across all Indigenous languages. While North American Indigenous languages may share some broad characteristics, each language has distinct linguistic features, and communities differ in terms of available resources and speaker populations. As such, the relevance of our evaluation framework may vary significantly across contexts.

We recognize that English is implicated in all evaluation tasks, and as a result, this study is not a within-culture monolingual evaluation. We also acknowledge that "culture" is not equivalent to language, and we do not intend to define any singular or essentialized notion of culture. Rather, we recognize that there is diverse cultural, social, and linguistic variation within language communities.

Our evaluation is also limited by our own experiences in our respective language communities. Many additional questions could be asked during evaluation, including different tasks, use cases, and linguistic considerations. The selected tasks represent only a subset of possible evaluations and should not be interpreted as exhaustive. Likewise, the datasets used in this study are small; as a result, our findings should be interpreted as indicative rather than definitive and do not provide statistically precise estimates of model performance.

Finally, our analysis is shaped by our own knowledge and positionality as researchers and community members. We offer insight into the Mvskoke and Choctaw communities, but we are not elders and do not claim authoritative perspectives on every aspect. Our inclusion of three additional Indigenous languages is intended for comparative purposes only, and we do not claim authority on those languages. There may be linguistic, cultural, or contextual nuances that are not fully captured in our evaluation.

Acknowledgments

Thank you to our elders and knowledge holders. We acknowledge those who have gone before us, who have carried and passed down our languages, and those who continue to sustain them. Thank you to Steven Bird for advising, and to Tad Hosford and Ian Iglesias for reviewing Mvskoke output. Finally, we thank the anonymous reviewers

for their helpful feedback.

References

Anthropic. 2026. Claude opus 4.6.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266. Association for Computational Linguistics.

Rosalia Badhorse. 2023. *Tsetsėhestâhese and So'Taeo'o (Cheyenne) Language: Grammar Sketch for Learners*. Ph.D. thesis, University of Arizona.

Rachel Baker-Ramos, Will Gelder, Leah Cho, Jahnavi Kolakaluri, and Josiah Hester. 2025. Kumu connect: Design thinking for place-based generative educational technology in hawaiian immersion schools. In *Proceedings of the 2025 Conference on Research on Equitable and Sustained Participation in Engineering, Computing, and Technology*, pages 186–195. ACM.

Anita Bardwell, Joe Bardwell, Lopaka Weltman, and Hoaloha Westcott. 2020. *He Papa Kuhikuhi Pilina'ōlelo: Reference Grammar of the Hawaiian Language*. University of Hawai'i.

Matthias Brenzinger and Patrick Heinrich. 2013. The return of hawaiian: language networks of the revival movement. *Current Issues in Language Planning*, 14(2):300–316.

Jacqueline Brixey and David Traum. 2022. Towards an automatic speech recognizer for the choctaw language. In *Proc. S4SG 2022*, pages 6–9.

Cyrus Byington. 1870. Grammar of the Choctaw language. *Proceedings of the American Philosophical Society*, 11:317–367. Edited by Daniel Garrison Brinton. Also published as a monograph by McCalla and Stavely, Philadelphia, 1870.

Chaka Chaka. 2024. Currently available GenAI-powered large language models and low-resource languages: Any offerings? wait until you see. *International Journal of Learning, Teaching and Educational Research*, 23:148–173.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2454–2469, New York, NY, USA. Association for Computing Machinery.

Monojit Choudhury. 2023. Generative AI has a language problem. *Nature Human Behaviour*, 7:1802–1803.

Roberts Darģis, Guntis Barzdins, Inguna Skadiņa, Normunds Gruzitis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Marc T J Elliott, Deepak P, and Muiris Maccarthaigh. 2025. Evolving generative ai: entangling the accountability relationship. *Digital Government: Research and Practice*, 6(1):1–13.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude.

Louise Fisher, Lenora Holliman, Wayne Leman, Leroy Pine Sr., and Marie Sanchez. 2023. *Cheyenne Dictionary*. Chief Dull Knife College.

Earnest Gouge, Edited, Translated by Jack B. Martin, and Juanita McGirt. 2004. *Totkv Mocvse / New Fire: Creek Folktales*. Norman: University of Oklahoma Press.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and

175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Mary R. Haas. 1979. Southeastern languages. In Lyle Campbell and Marianne Mithun, editors, *The Languages of Native America: Historical and Comparative Assessment*, pages 299–326. University of Texas Press.

Mary R. Haas, James H. Hill, Jack B. Martin, Margaret McKane Mauldin, and Juanita McGirt. 2015. *Creek (Muskogee) Texts*. University of California Publications.

Zachary Arao Hanson. 2025. Indigenous (mis)representation in emerging LLM research methodologies. *UC Riverside Undergraduate Research Journal*, 19.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*, 26(2):1–10.

Leanne Hinton. 2013. *Bringing our languages home: Language revitalization for families*. Heyday Books.

Gregg Howard and Rick Eby. 1990. *Introduction to Cherokee*. Various Indian Peoples Publishing Co.

Charles Julian. 2010. *A History of the Iroquoian Languages*. Ph.D. thesis, University of Manitoba.

Daniel Heath Justice. 2025. *Our Fire Survives the Storm: A Cherokee Literary History*. University of Minnesota Press.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374. Association for Computational Linguistics.

James N Leiker and Ramon Powers. 2011. *The northern Cheyenne exodus in history and memory*. University of Oklahoma Press.

Wayne Leman. 1980. *A reference grammar of the Cheyenne language*. Chief Dull Knife College.

Aha Pūnana Leo and Hale Kuamo o. 2003. *Māmaka Kaiao*. University of Hawai i Press.

Jason Edward Lewis, Hundefinedmi Whaanga, and Ceyda Yolgörmez. 2024. Abundant intelligences: placing ai within indigenous knowledge frameworks. *AI & Society*, 40(4):2141–2157.

Richard Littlebear. 2024. Neneehove'tanonėstse tsehe'enėstsestse; tsehe'enėstsetse neneehove'tanone: We are our languages; our languages are us. *Tribal College*, 35:1–3.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Jack B Martin. 2011. *A grammar of creek (Muskogee)*. UNP - Nebraska.

Jack B Martin and Margaret Mckane Mauldin. 2000. *A Dictionary of Creek/Muskogee*. University of Nebraska Press.

Paul J Meighan. 2024. Indigenous language revitalization using *TEK-nology* : how can traditional ecological knowledge (TEK) and technology support intergenerational language transmission? *Journal of Multilingual and Multicultural Development*, 45:3059–3077.

Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. FilBench: Can LLMs understand and generate Filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.

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.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah

Moeller, Jeffrey Micher, and 5 others. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

Mississippi Band of Choctaw Indians. 2026. Choctaw language dictionary.

Ōiwi Parker Jones. 2018. Hawaiian. *Journal of the International Phonetic Association*, 48(1):103–115.

Brock Pitawanakwat. 2018. Strategies and methods for anishinaabemowin revitalization. *The Canadian Modern Language Review*, 74(3):460–482.

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

Pukui and Elbert. 2003. *Hawaiian Dictionary*. University of Hawai i Press.

Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, and Shlomo Mark. 2024. Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: Narrative review and qualitative study. *JMIR AI*, 3:e51834.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Mike Rogers. 2021. Choctaw Nation members talk about impact of losing native speakers to COVID-19. *News 12*.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-first edition. SIL International, Dallas, Texas.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo' Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2026a. Are small language models the silver bullet to low-resource languages machine translation? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 1–26. Association for Computational Linguistics.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo' Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2026b. Are small language models the silver bullet to low-resource languages machine translation? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 1–26. Association for Computational Linguistics.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

The Choctaw Nation of Oklahoma Dictionary Committee. 2016. *Chahta Anumpa Tosholi Himona: New Choctaw Dictionary*, 1st edition. Choctaw Print Services.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating LMMs on culturally diverse 100 languages. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575. IEEE.

Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and

Sjur Nørstebø Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 15922–15931. European Language Resources Association (ELRA) and ICCL.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language. In *ACL 2022*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781. Association for Computational Linguistics.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, ASR. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4004–4011. European Language Resources Association (ELRA).

A Appendix

This section provides additional detail about the ASR task.

mus	
reference	Este hvmket likvyvntvs, makesasvtēs, mahokvnts.
prediction	Este hvmket legayantis mvgisazetis mahagets.
cho	
reference	Haklo hattak mʊt kanima il ia chi ka pim anoli
prediction	Haklo, hattakbat kanimma il ia chinka pim anoli.

Table 6: ASR example output of short samples given by GPT 4o transcribe. Punctuation is stripped before error rate calculation. Output contains characters that are outside of the orthographies.

The ASR experiment comprised of two audio samples per language, one short and one longer sample. Each sample is from high-quality recordings from language documentation. Table 7 gives the length of each audio clip and the corresponding WER and CER of the outputs. Table 6 shows an example output from a short clip.

		GPT 4o transcribe		Gemini 3.1 pro	
lang	length	WER	CER	WER	CER
mus	12	0.83	0.41	1.00	0.45
mus	37	1.00	0.52	0.83	0.23
cho	5	0.70	0.15	1.00	0.26
cho	38	0.94	0.37	0.96	0.45

Table 7: Audio clip lengths (in seconds) and ASR error rates per model.