Bottlenecks of In-Context Learning for Fieldwork ASR: A Case-study of Panãra

Siyu Liang[1,2], **Myriam Lapierre**[3,2], **Gina-Anne Levow**[2]
[1]Department of Linguistics, Rice University
[2]Department of Linguistics, University of Washington
[3]Department of Linguistics, McGill University
{liangsy, levow}@uw.edu, myriam.lapierre2@mcgill.ca

Abstract

In-context learning (ICL) enables ASR models to transcribe unseen languages by conditioning on a handful of audio-transcript pairs at inference time, with no fine-tuning. This is appealing for language documentation, where transcribed data is scarce and recording conditions vary across sessions. We evaluate ICL on Panãra (Northern Jê, Brazil), a language with a complex practical orthography in which diacritics encode phonemic contrasts, across seven fieldwork recordings varying in speaker, narrative, and recording context. We find substantial within-language variation in transcription accuracy unexplained by any single recording-level factor, and show that diacritics are a systematic bottleneck with pronounced differences across diacritic types. An orthographic manipulation experiment further shows that how diacritics are represented in context transcriptions substantially affects model performance. These results highlight orthographic complexity and recording-level variation as key practical challenges for ICL-assisted fieldwork transcription.

1 Introduction

Manual transcription remains a severe bottleneck in linguistic fieldwork: a single hour of audio in a newly documented language can require up to 50 hours of expert effort (Shi et al., 2021), and the volume of untranscribed recordings continues to far outpace transcription capacity (Bird, 2020b). Automatic speech recognition (ASR) promises to accelerate this process, but traditional supervised approaches require substantial transcribed training data, creating a circular dependency for under-documented languages. Two paradigms—fine-tuning and in-context learning (ICL)—have emerged to address this (Section 2). In ICL, the model is given a small set of audio–transcript pairs as context at inference time and adapts to an unseen language with no parameter updates, lowering the annotation requirement from minutes of training data to a handful of utterances. Yet both approaches typically report aggregate metrics averaged over test sets and languages, obscuring how orthographic and recording-level factors interact within a single language.

We address this gap by evaluating ICL on seven Panãra recordings spanning different narratives and speakers in distinct recording contexts. Panãra (Northern Jê, Brazil) has a large phonological inventory represented in a diacritic-rich practical orthography (Lapierre, 2023b), making it an ideal test case for studying the interaction between orthographic complexity and ICL performance.

We make three contributions. First, we provide a per-recording evaluation of ICL for a single endangered language across seven recordings, revealing substantial within-language variation driven by recording-level factors, with diacritics constituting a systematic bottleneck. Second, we present a linguistically grounded error analysis breaking down diacritic accuracy by phonological category, showing that different diacritic types are reproduced at vastly different rates. Third, we show through an orthographic manipulation experiment that diacritics are a key driver of transcription error: providing diacritic-stripped context exemplars lowers CER in most recordings, but at the cost of losing phonological distinctions essential to Panãra (e.g., vowel nasalization and height contrasts), highlighting a fundamental limitation of current ICL approaches for languages with complex orthographies.

2 Background

2.1 Fine-Tuning for Low-Resource ASR

Self-supervised multilingual models such as MMS (Pratap et al., 2024) and XLS-R (Babu et al., 2021) can be adapted to new languages with small amounts of transcribed data. Adapter-based fine-tuning—introducing small, trainable layers while keeping the base model frozen—has proven partic-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 157–166

ularly efficient for fieldwork languages. (Houlsby et al., 2019; Mainzinger and Levow, 2024; Guillaume et al., 2022; Nowakowski et al., 2023). Systematic benchmarks show that as little as 10 minutes of data can yield usable CER with MMS, though no single architecture consistently outperforms others under extreme data scarcity (Liang and Levow, 2025; Jimerson et al., 2023).

2.2 In-Context Learning for ASR

In-context learning (ICL) requires even less annotation: models like Omnilingual ASR (Keren et al., 2025) accept a handful of transcribed utterances as context at inference time, adapting to unseen languages with no parameter updates. ICL for speech has shown promise for speaker and variety adaptation (Roll et al., 2025) and multilingual ASR (Zheng et al., 2025; Fathullah et al., 2023). Cross-language evaluations on fieldwork languages find that context selection by acoustic similarity reduces CER by 25% relative to random selection , though ICL does not match fine-tuned performance (mean CER 0.47 vs. 0.29 with 10 minutes of fine-tuning data; Liang and Levow, 2025).

2.3 Orthographic Complexity

A key challenge for both paradigms is orthographic complexity. Taguchi and Chiang (2024) show across 25 languages that orthographic complexity—not phonological complexity—predicts ASR accuracy, with logographic and diacritic-heavy writing systems posing the greatest difficulty. However, these cross-language studies report aggregate metrics that obscure how orthographic factors interact with recording-level variation within a single language. Our work addresses this gap by isolating the role of diacritics across multiple recordings of language with a diacritic-rich transcription system.

3 Data

3.1 Panãra

Panãra (also written Panará or Panära; ISO 639-3: kre) is a Jê language spoken by approximately 700 people in Mato Grosso, Brazil (Lapierre, 2023b). The language has an exceptionally large phonological inventory: 17 consonant and 28 vowel phonemes, with oral/nasal and short/long contrasts in both inventories, as well as complex patterns of segmental alternation (Lapierre, 2023b). Consonant clusters include the obstruent-approximant sequences /pr, pj, tw, sw, kr, kj/—which give rise to excrescent vowels (De Falco et al., 2026)—as well as surface complex segments of the type nasal consonant-obstruent ([mp, nt, ns, ŋk]) which result from a process of nasal consonant post-oralization before a phonemically oral vowel (Lapierre, 2023c). Tables 1 and 2 present the consonant and vowel inventories; phonemes are given in /slashes/ with corresponding graphemes in ⟨angled brackets⟩.

	Bilabial		**Dental**		**Palatal**		**Velar**	
	I	O	I	O	I	O	I	O
Short obstruents	/p/	⟨p⟩	/t/	⟨t⟩	/s/	⟨s⟩	/k/	⟨k⟩
Long obstruents	/pː/	⟨pp⟩	/tː/	⟨tt⟩	/sː/	⟨ss⟩	/kː/	⟨kk⟩
Short nasals	/m/	⟨m⟩	/n/	⟨n⟩	/ɲ/	⟨j̃⟩	/ŋ/	⟨j̃⟩
Oralized nasals	[mp]	⟨np⟩	[nt]	⟨nt⟩	[ns]	⟨ns⟩	[ŋk]	⟨nk⟩
Long nasals	/mː/	⟨mm⟩	/nː/	⟨nn⟩				
Approximants	/w/	⟨w⟩	/ɾ/	⟨r⟩	/j/	⟨j⟩		

Table 1: Panãra's consonant inventory. **I** = IPA; **O** = Orthography.

	Short						**Long**					
	Front		**Central**		**Back**		**Front**		**Central**		**Back**	
	I	O	I	O	I	O	I	O	I	O	I	O
Oral vowels												
High	/i/	⟨i⟩	/ɯ/	⟨y⟩	/u/	⟨u⟩	/iː/	⟨ii⟩	/ɯː/	⟨yy⟩	/uː/	⟨uu⟩
Mid	/e/	⟨ê⟩	/ɤ/	⟨â⟩	/o/	⟨ô⟩	/eː/	⟨êê⟩	/ɤː/	⟨ââ⟩	/oː/	⟨ôô⟩
Low	/ɛ/	⟨e⟩	/a/	⟨a⟩	/ɔ/	⟨o⟩	/ɛː/	⟨ee⟩	/aː/	⟨aa⟩	/ɔː/	⟨oo⟩
Nasal vowels												
High	/ĩ/	⟨ĩ⟩	/ɯ̃/	⟨ỹ⟩	/ũ/	⟨ũ⟩	/ĩː/	⟨ĩĩ⟩				
Mid	/ẽ/	⟨ẽ⟩			/õ/	⟨õ⟩	/ẽː/	⟨ẽẽ⟩			/õː/	⟨õõ⟩
Low			/ã/	⟨ã⟩					/ãː/	⟨ãã⟩		

Table 2: Panãra's vowel inventory. **I** = IPA; **O** = Orthography.

Transcriptions follow a practical orthography developed collaboratively with Panãra teachers (Lapierre, 2024). Diacritics are pervasive: circumflexes mark mid oral vowels (*â*, *ê*, *ô*), and nasalization is marked with tildes or dieresis[1] (*ã*, *ẽ*; *ä, ë*). These orthographic marks directly reflect phonemic contrasts, so diacritic errors in ASR output correspond to phonologically meaningful mismatches. This large and complex inventory, with roughly 45

[1]Tilde is used to represent vowel nasality in phonetic and phonological transcriptions. In Panãra orthography, both tilde and dieresis are used—tilde was the sole convention until 2019, but because several characters do not support it easily on phone keyboards (i, y, u), dieresis became the more practical and widely adopted alternative. At present, both diacritics are commonly used in orthography; see Table 4 for the distribution across recordings.

ID	Year	Genre	#Utt	Dur(s)	Mean(s)	Diac%
sykja	2017	narrative	208	428	2.06	17.9
karansa	2018	procedural	304	807	2.65	15.3
sokriti	2019	historical	221	731	3.31	16.1
turen	2022	myth	78	276	3.54	17.9
krenpy	2023	myth	216	1228	5.69	18.8
kjarasa	2024	myth	52	274	5.26	18.3
patti	2024	narrative	295	871	2.95	17.2
Total			**1374**	**4615**	**3.36**	**17.2**

Table 3: Panãra recordings. **#Utt** = utterances after filtering (0.5–30s duration; metalinguistic annotations removed). **Dur** = total audio duration in seconds. **Mean** = mean utterance duration. **Diac%** = percentage of combining diacritic characters relative to all NFD characters (excluding spaces), computed over all utterances in the recording.

ID	Chars	Tilde	Dieresis	Circumflex
sykja	4684	660	90	393
karansa	6885	512	159	563
sokriti	7243	69	906	499
turen	2214	13	291	174
krenpy	9143	0	1286	476
kjarasa	2167	20	287	134
patti	5869	8	675	463

Table 4: Diacritic token counts across all utterances per recording (full corpus, not restricted to train/test split). **Chars** = total Normalization Form C (NFC) characters excluding spaces. Tilde and dieresis both mark vowel nasalization; their distribution reflects a shift in orthographic convention from tilde to dieresis beginning around 2019.

distinct phone categories (comparable in size to English, e.g., ∼39 phones in TIMIT), poses particular challenges for cross-lingual ASR due to the additional dimensions of vowel nasality and height encoded through diacritics (Ahn et al., 2024).

3.2 Recordings

We use seven recordings from documentary fieldwork, summarized in Table 3. The recordings vary considerably in length, ranging from 52 to 304 utterances and from 274 to 1228 seconds of audio. Utterances with parenthetical metalinguistic comments or slash-marked alternative transcriptions were excluded entirely. Utterances consisting solely of bracketed non-linguistic content (e.g., laughter, sound effects) were also removed. Partially bracketed material (e.g., false starts, Portuguese loanwords) was retained with brackets removed, since this content is present in the audio.

The recordings differ not only in speaker and content, but also in transcription conventions: the *sykja* recording uses Unicode combining diacritics (ã, ẽ) while others use precomposed characters (ä, ê). This orthographic variation is typical of fieldwork data collected across multiple years, reflecting the natural evolution of the documentation process—the gradual increase in knowledge and the ongoing standardization of conventions.

Table 4 shows the distribution of the three main diacritic types across recordings. Older transcriptions (*karansa*, *sykja*) predominantly use tilde for nasalization, while more recent ones (*krenpy*, *patti*, *kjarasa*, *turen*) favor dieresis, reflecting the shift in orthographic convention described above.

Kjarasa is a traditional myth told by Kjârasâ Panãra (female, ∼75 y.o.) about the origin of the Panãra people. *Patti* is a personal narrative told by Pâtti Panãra (male, ∼85 y.o.) about how he hunted a jaguar in his young adulthood. *Krenpy* is a traditional myth told by Kreenpy Panãra (female, ∼70 y.o.) about a woman who gave birth to a snake. *Karansa* is a procedural narrative by Karãsâ Panãra (female, ∼30 y.o.), describing traditional work and daily tasks carried out by Panãra women. *Sykja* is a personal narrative by Sykjã Panãra (male, ∼50 y.o.), recounting his path to becoming a shaman and aspects of witchcraft. *Turen* is a traditional myth told by Turẽn Panãra (female, ∼70 y.o.) about how the sun burned the moon's belly. Finally, *Sokriti* is a historical narrative told by Sokriti Panãra (male, ∼70 y.o.), recounting the first contact between the Panãra community and non-Indigenous Brazilians in the 1970s.

While the recordings exhibit some variation in genre, they all reflect the monologic discourse style characteristic of traditional Panãra storytelling—a common genre in societies where knowledge is primarily transmitted orally.

The recordings are available online in archival collection #2017-12 of the California Language Archive (Lapierre, 2017).

4 Methodology

4.1 Model

We use Omnilingual ASR 7B zero-shot (Keren et al., 2025), an encoder-decoder model supporting over 1,600 languages. For unseen languages, the model requires exactly 10 audio-transcript context examples at inference time as an architectural constraint; if fewer are available, examples are repeated to fill all 10 slots. Panãra is not in the

model's training data.

4.2 Within-File ICL Design

For each recording independently, the first 10 utterances serve as ICL context. We evaluate on 20 test utterances randomly sampled from the remaining utterances in the recording. We run the following two experiments.

Experiment 1 (baseline). We evaluate 10-shot ICL on each recording under two context selection strategies: (a) *sequential*—the first 10 utterances in recording order, simulating the natural fieldwork scenario of transcribing from the beginning of a session; and (b) *random*—10 utterances sampled uniformly at random, where each of the 20 test utterances receives its own independently sampled context from the non-test pool.

Experiment 2 (orthographic representation). Using the sequential split, we compare three orthographic representations applied consistently to both context transcriptions and evaluation references:

- *Original*: Normalization Form C (NFC)-normalized fieldwork transcription (as in Experiment 1).
- *Stripped*: All diacritics removed, retaining only base letters.
- *Expanded*: Each diacritic character replaced by a two-letter ASCII digraph motivated by its phonological value. Tilde and dieresis are unified since both mark nasalization (e.g., ã/ä→an), and circumflex is expanded to reflect mid vowel quality (â→ah).

Experiment 2 tests whether presenting the model with transcriptions that use only ASCII characters—eliminating the unseen diacritic inventory from context—affects recognition accuracy.

4.3 Evaluation Metrics

We report Character Error Rate (CER) as our primary metric, computed after lowercasing and whitespace normalization. We report both mean and median CER: the median is more robust to occasional hallucination errors where the model generates substantially more text than the reference (CER > 1.0). For Experiment 1, we additionally compute *base CER* by stripping diacritics from both prediction and original reference after scoring, isolating the model's accuracy on the consonant-vowel level from its handling of diacritics. We note that base CER and Experiment 2 stripped CER measure different things: base CER applies post-hoc stripping to model output generated under the original orthography, whereas stripped CER evaluates a model that was conditioned on stripped context transcriptions.

Recording	Sequential	Seq. Median	Random
sykja	.770	.587	.548
karansa	.470	.458	.405
sokriti	.520	.537	.474
turen	.514	.530	.562
krenpy	.669	.690	.626
kjarasa	.577	.565	.654
patti	.619	.618	.653

Table 5: Experiment 1 results: 10-shot ICL, original orthography, 20 test utterances per recording. **Sequential**: mean CER, sequential context (first 10 utterances). **Seq. Median**: median CER, sequential. **Random**: mean CER, random per-utterance context.

5 Results and Analysis

5.1 Experiment 1: Per-Recording Baseline

Table 5 reports mean and median CER for each recording under 10-shot ICL with the sequential context strategy, along with the mean CER under random context selection.

ICL achieves moderate accuracy on most recordings, but mean CER varies substantially across the seven recordings. Five recordings cluster between 0.47 and 0.62, while *krenpy* (0.67) and *sykja* (0.77) show higher error rates. *Sykja* has high per-utterance variance ($\sigma = 0.96$; note that its median CER of 0.59 is substantially lower than the mean of 0.77, indicating a few severely hallucinated utterances inflate the average). Standard deviations within recordings range from 0.11 to 0.96. The sequential and random context strategies yield similar overall CER, with neither consistently outperforming the other across recordings.

5.2 Cross-Recording Variation

Despite similar aggregate CER, the recordings differ in per-utterance variability and which factors drive error. Utterance duration shows a positive association with CER for *kjarasa* ($\rho = 0.57$, $p = 0.009$) and *krenpy* ($\rho = 0.67$, $p = 0.001$), where longer utterances tend to receive higher CER, consistent with error accumulation over longer sequences. Vocabulary overlap (the fraction of unique words in a test utterance that appear in the context) between context and test utterances

Text	Original	Stripped	Expanded
sykja	.770	**.475**	.504
karansa	.470	.392	**.369**
sokriti	.520	**.464**	.478
turen	.514	**.416**	.447
krenpy	**.669**	.817	.720
kjarasa	.577	.492	**.478**
patti	**.619**	.648	.657

Table 6: Experiment 2: mean CER under three orthographic representations (10-shot sequential, same test set as Exp. 1 sequential). **Bold** = best per row.

shows a negative association with CER for *turen* under random context ($\rho = -0.64$, $p = 0.002$), suggesting that test utterances whose words appear in the context receive lower CER; however, this effect is not observed consistently across other recordings. Given the small per-recording test sets ($n = 20$), these correlations should be interpreted as exploratory.

Genre does not appear to be a strong predictor of model performance. Notably, the procedural text *karansa*—despite containing relatively little repetition, a rhetorical strategy more typical of narratives and often associated with elder speakers—nonetheless receives the lowest CER. Female speakers tend to receive lower CER overall than male speakers, and *karansa* is told by a female speaker who is also substantially younger than the others in the dataset. With only seven recordings, however, speaker, age, gender, and genre are heavily confounded—*karansa* is the sole procedural text and its speaker is the only one in her age range—so we can only flag age and gender as candidate drivers for follow-up work on a larger sample, not as established effects.

5.3 Experiment 2: Orthographic Representation

Table 6 compares mean CER across three orthographic representations under 10-shot sequential ICL.

Stripping diacritics from context transcriptions consistently reduces CER in five of seven recordings, with improvements ranging from 0.06 (*sokriti*) to 0.30 (*sykja*). The largest improvement occurs for *sykja* (0.77→0.48), which also has the highest baseline CER. However, diacritic stripping *degrades* performance for *patti* (+0.03) and *krenpy* (+0.15), the latter also exhibiting 5% hallucination (heuristically defined as CER > 1.0).

This CER reduction is partly a scoring effect: since both context and reference are stripped, diacritic errors are no longer penalized. However, it also reflects a genuine improvement in base-character accuracy: when the model is no longer required to produce unfamiliar diacritic characters, it can focus on the consonant-vowel skeleton where its cross-lingual priors are stronger. Comparing stripped CER against the base CER from Experiment 1 (which applies post-hoc stripping to output generated under original orthography) would isolate this effect, and we leave this as future work.

Phonologically motivated digraph expansion—unifying tilde and dieresis as nasalization (`an`) and encoding circumflex as mid vowel quality (`ah`)—yields a more nuanced picture. For two recordings, expanded *outperforms* both original and stripped: *kjarasa* (0.48 vs. 0.49 stripped) and *karansa* (0.37 vs. 0.39 stripped). For most other recordings, expanded falls between original and stripped (e.g., *sokriti* 0.48, *turen* 0.45). However, expanded still hurts performance for *patti* (0.66) and *krenpy* (0.72) relative to original, mirroring the pattern seen with stripping for these two recordings. Additionally, the longer token sequences produced by expansion (e.g. `mahmah` or `hapoooo`, arising from repeated diacritic vowels) may exceed the model's expected character-sequence distribution, triggering abnormal outputs.

These results suggest that when using ICL for Panãra fieldwork data, both stripped and expanded representations could improve over original for most recordings, but practitioners should test both on a small held-out set, as the benefit is recording-dependent.

5.4 Diacritic Analysis

Table 7 isolates the diacritic contribution to CER by comparing the full CER from Experiment 1 with a base CER computed by post-hoc stripping of diacritics from both model output and reference. This differs from Experiment 2 stripped CER: here, the model was conditioned on diacritics, but we evaluate only the base-letter skeleton.

Diacritics contribute 0.04–0.10 to CER across recordings. The diacritic penalty varies across recordings: *turen* has the largest Δ (0.101) while *patti* has the smallest (0.038). This variation suggests that recording-specific factors—speaking rate, acoustic clarity, or utterance length—affect diacritic reproducibility independently of how many diacritics appear in the transcription.

Table 8 breaks down model accuracy by dia-

Text	Diac%	Full CER	Base CER	Δ Diac
sykja	17.9	.770	.693	+.077
karansa	15.3	.470	.380	+.090
sokriti	16.1	.520	.437	+.083
turen	17.9	.514	.413	+.101
krenpy	18.8	.669	.608	+.061
kjarasa	18.3	.577	.507	+.070
patti	17.2	.619	.581	+.038

Table 7: Full CER vs. base CER (post-hoc diacritic stripping applied to Exp. 1 sequential predictions). **Diac%** from Table 3. **Δ Diac** = full − base.

Type	Correct	Base-sub	Other-sub	Deleted
Tilde	32%	24%	36%	8%
Circumflex	5%	16%	45%	34%
Dieresis	24%	18%	24%	34%

Table 8: Model accuracy on diacritic characters by phonological category (aggregate over all recordings, Exp. 1 sequential). **Correct**: diacritic reproduced exactly. **Base-sub**: substituted with base letter (e.g., ã→*a*). **Other-sub**: substituted with a different character. **Deleted**: omitted entirely.

critic category. Circumflex vowels (â, ê, ô; marking mid oral vowels) are reproduced correctly only 5% of the time, and nearly half of all circumflex errors are substitutions with a *different* diacritic rather than the base letter—most commonly â→ä or ê→ẽ. This suggests the model perceives that a diacritic is needed but mis-selects the category.[2] Tilde (vowel nasalization: ã, ẽ, ĩ) is reproduced most reliably (32% correct), while dieresis (vowel nasalization: ä, ë, etc.) is intermediate (24% correct). The most frequent individual substitution is ï→*i* (20 occurrences) and ã→*a* (17 occurrences), confirming base-letter drop as the dominant diacritic error mode.

5.5 Error Patterns

Character-level Levenshtein alignment of model outputs against references reveals consistent patterns across recordings. Substitutions are the dominant error type (84–194 per recording), with deletions highly variable (32 in *kjarasa* to 469 in *krenpy*) and insertions ranging from 20 to 278. The high deletion count in *krenpy* likely reflects the model generating systematically shorter output than the reference for longer utterances (mean 5.69s). *Sykja* has unusually high insertions (278), consistent with its hallucination-prone outlier utterances. Across recordings, 29–47% of substitutions involve diacritic characters, confirming that diacritics are a disproportionate source of character-level confusion. The most common non-diacritic consonant confusions are r→n (13 occurrences) and j→i (11 occurrences), both acoustically plausible cross-lingual approximations—the palatal approximant ⟨j⟩ and rhotic ⟨r⟩ frequently map to nasal or glide-like segments in the model's output.

[2] Anecdotally, this error type is also common among native speakers learning to write Panãra.

6 Discussion

6.1 Linguistic interpretation of results

Diacritic omission. As noted in §5.4, the most frequent individual substitution is ï→*i* (20 occurences across all recordings). This is unsurprising from a phonological perspective: [i] is a common epenthetic vowel in Panãra whose nasality is non-contrastive, assimilating instead to that of the adjacent consonant (Lapierre, 2023a). Epenthetic [i ∼ ĩ] appears word-initially when the root-initial consonant is a geminate or post-oralized nasal, in roughly 20% of words in the lexicon. Because the presence or absence of the nasal diacritic on this segment is phonemically vacuous and never distinguishes lexical items, the model effectively learns to treat it as irrelevant—hence the high frequency of this substitution.

Diacritic stripping. Stripping diacritics was found to improve accuracy in five of the seven recordings (*sokriti, kjarasa, karansa, turen, sykja*), with the largest improvement occurring for *sykja*. However, stripping was instead found to reduce performance accuracy for *patti* and *krenpy*. This pattern mirrors the transcription accuracy timeline: *sykja* was transcribed earliest, when orthographic conventions—including diacritic use—were still being established, while *patti* and *krenpy* are among the most recent transcriptions, reflecting greater linguistic knowledge and more consistent phoneme-to-grapheme mapping. The benefit of stripping diacritics is therefore greatest where diacritic use in the reference transcription is itself least reliable.

Digraph expansion. Digraph expansion yielded a more nuanced picture: it outperforms both original and stripped for *kjarasa* and *karansa*, falls between the two for most other recordings, but hurts performance for *patti* and *krenpy*. One important complication in this strategy is that the digraph encoding for nasal vowels (e.g., ã/ä→an) interacts with existing phonemic contrasts in Panãra: since

/an/ and / a/ are distinct, and the language further contrasts /VN/, /ṼN/, and /ṼT/ structures, introducing nasality digraphs risks neutralizing contrasts that are present in Panãra's phonological grammar.

Circumflex substitutions. Circumflex vowels (â, ê, ô) were reproduced correctly only 5% of the time, with the majority of errors involving substitution with a different diacritic. This likely reflects the token-level frequency of diacritized vowels types in the corpus: across all recordings, nasal-diacritic vowels (tilde and dieresis combined) account for approximately 65% of all diacritic tokens, with circumflex vowels making up the remaining 35% (see Table 4). Since the decoder emits each diacriticized vowel as a single character token rather than composing a base vowel and a diacritic in two steps, this distributional skew translates directly into output behavior: when the model emits any diacriticized vowel, it is more likely to be a nasal vowel than a circumflex one—consistent with both the corpus frequencies and the relative inventory sizes of nasal vs. mid oral vowel phonemes (10 vs. 6).

Consonant confusions. The most common non-diacritic consonant substitution is r→n, another pattern that can be explained via language-specific facts: /n/ frequently undergoes lenition to [ɾ̃] in the onset of unstressed syllables (Lapierre, 2023b), and in such contexts may be orthographically represented as ⟨r⟩. The model thus likely encountered word-initial /n/ transcribed inconsistently as both ⟨n⟩ and ⟨r⟩, yielding the observed confusion pattern.

Excrescent vowels. Panãra permits obstruent-approximant onset clusters (/kr, kj, pj, sw, pr/) that are typologically uncommon and absent from most of the model's training languages. The model systematically breaks these clusters by inserting a vowel, producing CV.CV sequences that conform to cross-linguistically common syllable structure. Examples include: *kre→kare* (*karansa* utt. 165), *pjâ→pija* (*karansa* utt. 168), and *swa→suwa* (*sokriti* utt. 76). This pattern mirrors the phonological process of excrescent vowel insertion described for Panãra (Lapierre, 2023b; De Falco et al., 2026), in which such vowels surface phonetically but are not represented orthographically. The model's cross-lingual priors thus impose CV syllable structure where the target orthography expects CC clusters, consistent with commonly reported native speaker intuitions about syllable structure.

6.2 Which Recording Properties Predict ICL Success?

Utterance length shows a moderate effect *within* recordings: duration correlates significantly with CER for *kjarasa* ($\rho = 0.57$) and *krenpy* ($\rho = 0.67$), suggesting that longer utterances are harder to transcribe within a given session. However, mean utterance duration does not predict CER *across* recordings—*sykja* has the shortest mean duration (2.06s) yet the highest CER (0.77), though this is largely driven by a single hallucinated utterance (CER 4.79); its median (0.59) is comparable to other recordings.

Vocabulary overlap between context and test shows a significant effect only for *turen* under random context ($\rho = -0.64$, $p = 0.002$), but is not consistent across recordings. This suggests that while direct reuse of context transcriptions (where the model reproduces words it has already seen in the context examples) may contribute to ICL success in some cases, it is not the dominant mechanism. Recordings with more repetitive or formulaic language (typical of Panãra storytelling) may benefit more from ICL, as utterances in these texts are more likely to share vocabulary with context examples.

Diacritic density (15–19%) does not strongly predict the size of the diacritic penalty across recordings (Δ Diac 0.04–0.10), suggesting recording-specific acoustic factors modulate diacritic reproducibility more than the raw frequency of diacritic characters. The diacritic *type* distribution matters more than overall density: recordings with more circumflex vowels face systematically higher diacritic error rates given the model's 5% accuracy on that category.

6.3 Practical Recommendations

Based on our results, we offer some guidance for fieldworkers considering ICL for Panãra transcription.

Consider stripped orthography in context transcriptions. Providing the model with base-letter-only transcriptions as context reduces CER (by up to 0.30). However, the benefit is not universal: two recordings (*patti*, *krenpy*) performed better with original orthography. We recommend testing both on a small held-out set before committing to a strategy for a given recording. Post-editing can then restore diacritics with reference to the audio.

Consider digraph encoding selectively. Phono-

logically motivated ASCII digraphs (ã→an, â→ah) outperform stripping for some recordings (*kjarasa*, *karansa*) but underperform for others. The longer token sequences and potential ambiguity with existing vowel-nasal contrasts make the benefit recording-dependent. We recommend testing on a small held-out set before adopting this strategy.

Expect high per-utterance variance. Even within a single session, CER ranges from near-zero to well above 1.0 in some cases. ICL drafts should be treated as variable-quality suggestions: useful as a starting point but requiring careful review rather than light correction. A staged workflow where ICL serves as a rapid first-pass tool, with post-edited outputs accumulating into training data for subsequent fine-tuning, is a promising direction. Our per-recording results support this: the ICL drafts we obtain (CER 0.47–0.77, median across recordings ≈0.57) require only 10 transcribed utterances rather than the minutes of annotation needed for fine-tuning.

Short utterances are unreliable. Very short recordings with brief utterances are poor candidates for ICL: in preliminary tests, a recording with mean utterance duration under 1.5s exhibited 20% of test utterances with CER > 1.0 (our heuristic for hallucination, where the model generates substantially more text than the reference). Where possible, segmentation should avoid single-word utterance parses, instead combining them with adjacent material.

6.4 Utility for the language documentation workflow

Language documentation typically aims to produce a corpus of naturalistic speech that is transcribed, morpheme-segmented, glossed, and translated into a lingua franca. Within this workflow, the principal bottleneck is transcription into the target language and translation into the lingua franca. For Panãra, this process requires on the order of 30 hours per hour of recording, even for the second author, who has functional conversational proficiency and over a decade of experience working with the language. Even partial automation would therefore represent a meaningful advantage—but only if the ASR output is close enough to the target that correction is faster than transcribing from scratch. The present results suggest this threshold has not yet been reached for naturalistic speech.

An ASR output that approximates phonetic content without reliably identifying morphemes and morpheme boundaries offers limited practical benefit, since the goal of transcription in language documentation is not merely rendering the phonetic signal but interpreting the meaning and glossing the morpho-syntactic content of utterances. Worse, near-but-not-quite transcriptions may actively interfere with perception and interpretation of the speech signal—especially for non-native transcribers—potentially increasing error rates rather than reducing effort.

That said, the model's stronger performance on short utterances points to one promising application: phone segmentation of target words with limited string length (roughly 4–10 characters). This is precisely the setting of a controlled phonetics experiment, where target words are known in advance, the set of items to be transcribed is limited, and the target words generally follow a templatic shape. Since the goal of phone segmentation is to demarcate boundaries between consonants and vowels rather than to interpret meaning or identify morphemes, the model's tendency to approximate phonetic content without capturing higher-level structure is less problematic.

For Panãra, current semi-automated phonetic data processing typically involves: (i) segmenting target utterances or words, (ii) supplying a phone-segmentation model (e.g., MFA) with the expected phone sequence for each interval, (iii) generating the phone-aligned output, and (iv) manually correcting the output. For low-resource languages, this workflow can require anywhere from dozens to hundreds of hours, depending on the number of speakers and target items. An ASR model could bypass the first two steps entirely, substantially reducing processing time and bringing workflows for low-resource languages closer to the more automated pipelines available for high-resource languages such as English, French, or Korean. We expect both accuracy and practical utility to be considerably higher in this setting than in naturalistic transcription, and leave systematic evaluation of this hypothesis to future work.

A useful framing for assessing partial automation is to ask what the post-edit task would look like if the model produced perfect base characters but no diacritics—collapsing transcription to a re-diacritization task. Our Experiment 2 stripped condition approximates an upper bound on this scenario: stripped CER values of 0.39–0.49 on five recordings indicate that even the base-character skeleton is not yet reliable enough to make re-

diacritization the only remaining work. A more direct measurement would be to time fieldworkers post-editing ICL drafts (under each orthographic condition) against transcribing the same utterances from scratch, at varying CER levels; we view this human-factors evaluation as the natural next step for establishing the practical utility threshold.

7 Conclusion

We evaluated in-context learning for ASR on Panãra, a low-resource and endangered language with a diacritic-rich practical orthography, across seven fieldwork recordings varying in speaker, genre, and recording context. Our results show that ICL can produce useful first-pass transcriptions with only 10 context utterances, but accuracy varies substantially across recordings in ways that no single factor fully explains. Diacritics are a systematic bottleneck: circumflex in particular is almost never reproduced correctly, and providing diacritic-stripped context improves performance for most recordings. These findings suggest that orthographic representation and recording-level heterogeneity deserve more attention in evaluations of ASR for low-resource languages, and that aggregate metrics mask important variation relevant to fieldworkers deploying these tools in practice.

Limitations

We evaluate on a single language with seven recordings (20 test utterances each), and our findings about speaker and recording effects may not generalize to other languages or larger corpora. The small test sets limit statistical power for per-recording claims; our Spearman correlations should be treated as exploratory. Our results are also based on a single ICL-capable system (Omnilingual ASR 7B); whether the diacritic and recording-level patterns we observe generalize to other ICL ASR models is an open question. We do not evaluate extrinsic utility, such as how much ICL-generated drafts actually speed up human post-editing. Our diacritic type analysis uses automatic Levenstein alignment, which may misattribute some errors at segment boundaries.

Ethical Considerations

The Panãra recordings used in this study were collected during long-term fieldwork with community consent for linguistic documentation and research, including all relevant community-issued authorization documents and IRB approvals. Any deployment of ASR tools in endangered language communities should involve community consultation and respect data sovereignty principles (Bird, 2020a). In addition, deploying ASR tools trained on majority languages for endangered language data carries broader risks: linguistically incorrect outputs misrepresent the language, and any derivative products should not be used for linguistic analysis without careful post-hoc correction with a speaker of the language.

References

Emily P. Ahn, Eleanor Chodroff, Myriam Lapierre, and Gina-Anne Levow. 2024. The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra. pages 1505–1509.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs, eess].

Steven Bird. 2020a. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2020b. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Ella De Falco, Myriam Lapierre, and Katherine Guild. 2026. Excrescent vowels in panãra: Evidence for gestural coordination in consonant clusters. Poster presented at the 20th Conference on Laboratory Phonology, Montréal.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Prompting Large Language Models with Speech Recognition Abilities. *arXiv preprint*. ArXiv:2307.11795 [eess].

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. *arXiv preprint*. ArXiv:2511.09690 [cs].

Myriam Lapierre. 2017. Panãra field materials, 2017–12. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley. Online archival collection.

Myriam Lapierre. 2023a. The phonology of Panãra: A prosodic analysis. *International Journal of American Linguistics*, 89(3):333–356.

Myriam Lapierre. 2023b. The Phonology of Panãra: A Segmental Analysis. *International Journal of American Linguistics*, 89(2):183–218.

Myriam Lapierre. 2023c. Two types of [nt]s in panãra: Evidence for temporally ordered subsegmental units. *Glossa: a journal of general linguistics*, 8(1).

Myriam Lapierre. 2024. Orthography development in the amazonian indigenous context: The case of panãra. In *2024 Annual Meeting of the Society for the Study of Indigenous Languages of the Americas*, New York City.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. 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 170–176, Bangkok, Thailand. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. 2025. In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4412–4426, Suzhou, China. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint*. ArXiv:2101.10877 [eess].

Chihiro Taguchi and David Chiang. 2024. Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15493–15503, Bangkok, Thailand. Association for Computational Linguistics.

Haolong Zheng, Yekaterina Yegorova, and Mark Hasegawa-Johnson. 2025. TICL: Text-Embedding KNN For Speech In-Context Learning Unlocks Speech Recognition Abilities of Large Multimodal Models. *arXiv preprint*. ArXiv:2509.13395 [eess].