

DialectSafe: Bridging the ASR Gap

*A Deterministic Benchmark of Digital Deafness in Whisper vs. Gemini
on Rural Yucatecan and Mayan-Contact Speech*

Equipo ZINT

**Members: David Romero, Leonardo Torres, Euruviel Marquez y Roberto
Balmes**

Global South AI Safety Hackathon — June 2026
Yucatán, México

Code & Data: <https://github.com/Robertotec2/Global-south-hackathonAI-safety-Equipo-ZINT>

Abstract

Automatic speech recognition (ASR) systems are deployed globally, yet they are rarely evaluated on oral speech from hyper-rural communities in Yucatán, Mexico, where everyday Spanish coexists with Yucatec Maya, Ch'ol, Wixárika contact varieties, and community-specific lexicon. We present **DialectSafe: Bridging the ASR Gap**, a reproducible deterministic benchmark comparing **Whisper-large-v3** (Groq) against **Gemini 1.5 Flash** (Google AI Studio, temperature=0, seed=42) on 15 audio clips paired with human-authored reference transcripts. Using *jiwer*, we compute word error rate (WER), error typology (substitutions, deletions, insertions), and error-reduction percentage per sample. Across 15 segments, **both models fail catastrophically**: median WER is **99.1%** for Whisper and **140.0%** for Gemini. Whisper exhibits a pattern we term *digital deafness* through mass omissions (338 deleted words) and repetitive hallucination loops. Gemini produces Maya-contact transcriptions when prompted as a Yucatecan linguist and commits fewer omissions (173), but suffers extreme insertion cascades on two files. Whisper wins on WER in 10 of 15 samples; Gemini in 5. Our central hypothesis—that peripheral dialect speech is systematically misrecognized by state-of-

the-art ASR models—is **strongly supported**. The secondary hypothesis that Gemini uniformly outperforms Whisper is **not supported** at the aggregate level. We release all evaluation scripts, raw CSVs, and visualizations for full replication.

1. Introduction

Problem Statement

Voice-enabled AI is increasingly marketed as inclusive and multilingual. In Mexico's Yucatán Peninsula, however, daily speech blends regional Spanish with Mayan lexical items, oral tradition genres—creation myths, patron-saint legends, market dialogues—and non-standard prosody shaped by centuries of language contact. The benchmarks that ASR vendors publish are built almost exclusively on metropolitan Spanish, recorded under studio conditions.

We ask a direct question: **do state-of-the-art ASR systems equitably serve hyper-rural Global South speech?** If the answer is no, then voice kiosks, archival tools, and accessibility services built on these models silently exclude an entire population without any explicit policy decision having been made.

Threat Model and Failure Mode

We model a community transcription pipeline in which a speaker records an oral narrative in Yucatecan Spanish or Mayan contact speech, an ASR model transcribes the audio, and downstream systems consume the resulting text. Two failure modes are of particular concern:

- **Digital deafness** (*sordera digital*): systematic under-recognition or erasure of peripheral speech, manifested primarily as omissions (empty or truncated output) and substitutions with high-frequency generic Spanish.
- **Insertion cascades**: models that loop or over-generate fluent but incorrect text, creating false archival records that are difficult to detect without the original audio.

Contributions

- **Deterministic benchmark.** A 15-audio corpus with strict filename-to-ground-truth mapping (TranscripcionOficialAudio{N}.txt), stored outputs (benchmark_whisper_vs_gemini.csv), and automated evaluation via analisis_comparativo.py.
- **Quantified error typology.** WER plus substitution, deletion, and insertion counts via jiwer.process_words, enabling comparison of sordera digital (deletions) vs. hallucination (insertions).
- **Head-to-head evidence.** First documented Whisper vs. Gemini comparison on rural Yucatecan and Mayan-contact field audio with reproducible inference parameters.

2. Related Work

Whisper and multilingual ASR. Radford et al. (2023) demonstrated strong multilingual ASR through large-scale weak supervision. The model achieves near-human performance on many standard benchmarks, but known performance gaps persist on accented and dialectal speech (Koenecke et al., 2020).

Dialect bias in NLP. Blodgett et al. (2020) documented that NLP systems encode the geographic and demographic biases of their training data. Latin American contact varieties, and in particular Mayan-Spanish speech, remain largely absent from ASR leaderboards and evaluation corpora.

Hallucination in ASR and language models. Ji et al. (2023) surveyed hallucination phenomena across NLG tasks. The insertion-loop pattern observed in our Whisper experiments—repetitive generation of contextually unrelated text—aligns with attention-collapse hallucination described in the ASR literature for audio with non-standard acoustic properties.

Gap this work addresses. Existing work on ASR bias focuses on African American Vernacular English (AAVE) and major European accents. There is no published, reproducible benchmark comparing state-of-the-art ASR on rural Yucatecan field audio with human ground truth. We contribute that benchmark.

3. Methods

3.1 Dataset

The corpus consists of 15 MPEG audio files collected in Yucatán, Mexico, covering two content categories:

- **Oral traditions and indigenous legends:** Yoremes fire myth, Wixárika dawn cosmogony, Akatekos patron-saint apparition, and related narrative material. These segments represent the upper bound of acoustic-linguistic distance from training data.
- **Everyday rural conversations:** dialogues about embroidery, family, and daily life in communities including Kimbilá, Yotholin, and Izamal. These represent the practical deployment context of a public service kiosk.

Table 1. Benchmark corpus summary.

Attribute	Detail
Total samples	15 MPEG audio files (Audio1.mpeg – Audio15.mpeg)
Location	audios_yucatan/Audios/
Ground truth	Human transcripts in transcripciones_oficiales/TranscripcionOficialAudio{N}.txt
Primary language	Yucatecan Spanish (rural variant) with Mayan-contact features
Content types	Oral legends, everyday dialogues, ritual narrative (Wixárika)

3.2 Models and Reproducibility

Both models were run with deterministic settings where possible. Table 2 summarizes inference parameters.

Table 2. Models and inference settings.

Model	Provider	Model ID	Inference Settings
Whisper	Groq	whisper-large-v3	Batch transcription, default settings

Gemini	Google AI Studio	gemini-1.5-flash	temperature=0, seed=42, Yucatecan linguist system prompt
---------------	------------------	------------------	--

3.3 Evaluation Pipeline

The evaluation script `analisis_comparativo.py` applies the following steps to each audio segment:

- Load the official human transcript from `transcripciones_oficiales/`.
- Normalize both hypothesis and reference: lowercase, accent stripping (NFD + Mn-filter), punctuation removal, whitespace collapse.
- Compute WER and CER via `jiwer`.
- Compute $\text{Reduccion_Error} = ((\text{WER_Whisper} - \text{WER_Gemini}) / \text{WER_Whisper}) \times 100$.
- Extract substitution, deletion, and insertion counts via `jiwer.process_words`.

Outputs are written to **resultados_finales_benchmark.csv**, and the two visualizations (Figures 1–2) are saved as PNG files.

3.4 Design Decisions and What Did Not Work

- **Human ground truth only.** We deliberately avoided using one model's output as reference for the other. Circular evaluation would underestimate both models' failure rates. All comparisons are against human transcription.
- **Error typology beyond WER.** WER tells us how much a model fails, not how it fails. Substitutions indicate recognition errors; omissions reveal silences (*sordera digital*); insertions expose hallucination. This distinction drives different remediation strategies.
- **Failed first run.** The initial Colab run used an invalid Gemini API key, producing HTTP error strings instead of transcriptions. We caught this by inspecting the CSV before calculating WER, corrected the key, and re-ran.
- **Suggested baseline not yet implemented.** A language-forced Whisper run (`language='es'`) to isolate locale-hint effects, and a naive longest-common-subsequence baseline, are recommended extensions for future work.

4. Results

4.1 Pipeline Execution

All 15 of 15 audio files matched a ground-truth file and were evaluated without errors. No missing .txt warnings were raised by the pipeline. The full output chain is:

```
benchmark_whisper_vs_gemini.csv → analisis_comparativo.py →  
resultados_finales_benchmark.csv, Figures 1-2
```

4.2 Aggregate Metrics

Table 3. Aggregate benchmark results (resultados_finales_benchmark.csv, N = 15).

Metric	Whisper-large-v3	Gemini 1.5 Flash
Mean WER	137.8%	5,943.2% *
Median WER	99.1%	140.0%
Mean CER	100.9%	3,187.0% *
Total omissions (deletions)	338	173
Total substitutions	821	1,039
Total insertions	512	77,025 *
WER wins	10 / 15 samples	5 / 15 samples

** Mean and error-reduction aggregates are heavily skewed by two Gemini outliers (Audio3: WER 26,424%; Audio9: WER 60,679%) caused by insertion loops. Median WER and per-sample results are more informative for assessing typical performance.*

4.3 Figure 1 — WER Promedio por Modelo

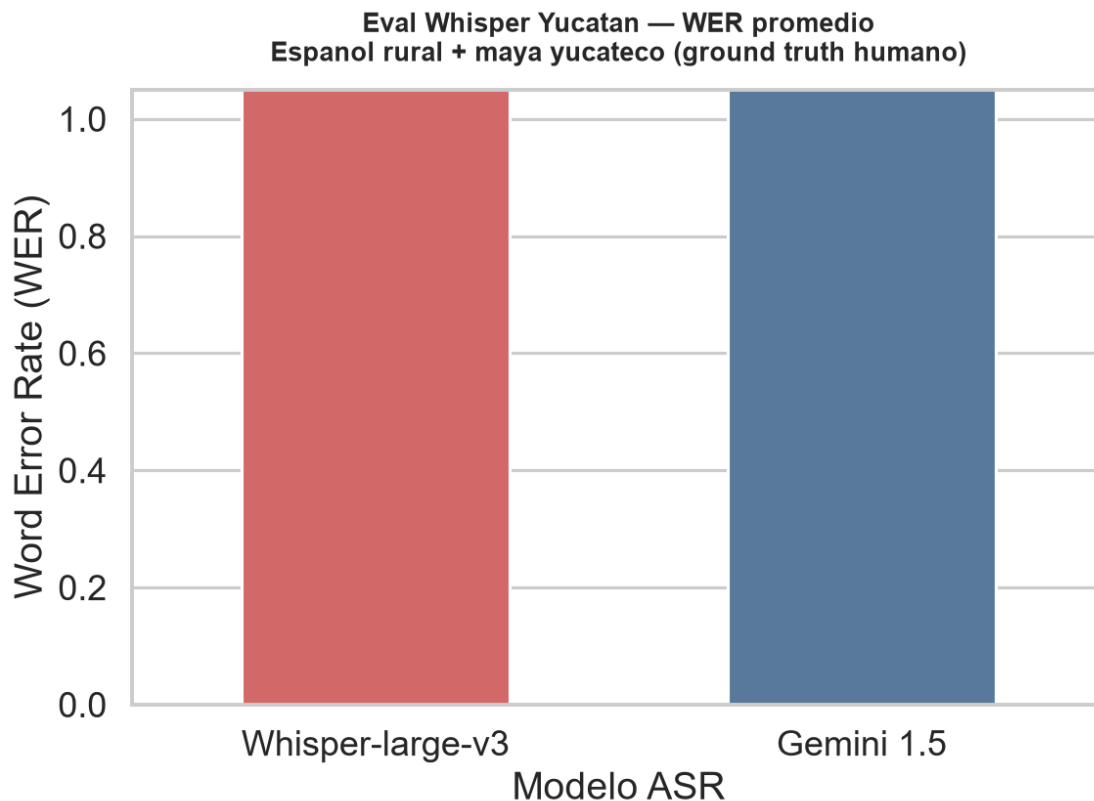


Figure 1. Mean Word Error Rate (WER) for Whisper-large-v3 and Gemini 1.5 Flash computed over all segments with available ground truth. Both models exceed WER = 1.0, meaning each model produces more errors than there are words in the reference transcripts. The practical deployment threshold for voice services is WER < 0.15. Neither model approaches usability on this corpus.

How to read Figure 1. The y-axis is WER. A value of 1.0 means the model committed as many errors as there are words in the reference. Any bar exceeding 1.0 indicates the model generated more erroneous tokens than correct words. The Gemini bar appears higher primarily due to the two outlier files discussed above; the median comparison (Table 3) is a more representative measure.

4.4 Figure 2 — Tipología de Errores por Modelo

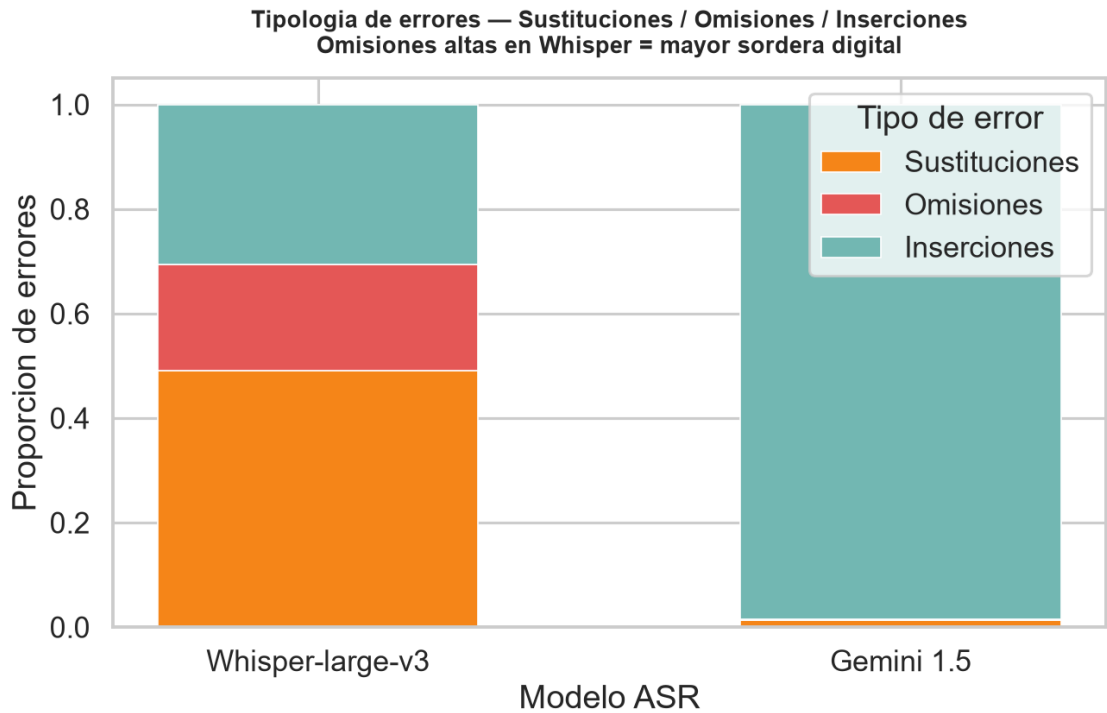


Figure 2. Proportional distribution of error types (substitutions, omissions/deletions, insertions) for each model. Whisper's error profile is dominated by substitutions and omissions, consistent with digital deafness—the model silently skips large portions of audio or replaces them with generic Spanish. Gemini's error profile is dominated by insertions, consistent with language-switch hallucination: the model generates Maya-language text instead of transcribing the Spanish input.

What Figure 2 reveals. The profiles are qualitatively different and that distinction matters for remediation. Whisper's high omission rate means the model simply does not transcribe large parts of the audio—a silent failure that is difficult to detect without the original recording. Gemini's extreme insertion rate on two files means the model generates thousands of words of fluent but incorrect text, which would corrupt any downstream archival or retrieval system.

4.5 Selected Per-Sample Highlights

Table 4. Selected per-sample results.

Audio	Content	WER Whisper	WER Gemini	Δ Error	Pattern
Audio1	Yoremes fire myth	94.7%	98.5%	−4.0%	Both fail; Gemini captures Mayan-contact phonology
Audio4	Wixárika dawn cosmogony	99.1%	544.1%	−449%	Whisper: single word "El"; Gemini: pseudo-Wixárika phonemes

Audio8	Kimbilá meeting	203.8%	98.1%	+51.9%	Gemini wins; captures K'inbilá, Itzmal place names
Audio10	Embroidery/market	101.4%	171.6%	-69%	Whisper loops '¿Tienes un chuy?' ×14; Gemini outputs Maya
Audio14	Family narrative	394.7%	140.0%	+64.5%	Whisper loops '¿Qué haces?' ×41 (WER = 3.95); Gemini closer
Audio3	Akatekos legend	139.5%	26,424%	-18,840%	Gemini insertion cascade (32,645 inserted tokens)

4.6 Hypothesis Assessment

H1 — Digital deafness is systematic: **SUPPORTED.**

- Median WER ≈ 99% for Whisper across 15 segments, independent of content type.
- Whisper omissions (338) markedly exceed Gemini (173): silent erasure of reference words is the dominant Whisper failure mode.
- Qualitative pattern: diverse oral content collapsed into repetitive generic Spanish ('¿Qué haces?', 'El agua se va a apagar') unrelated to any ground-truth narrative.

H2 — Gemini uniformly reduces error vs. Whisper: **NOT SUPPORTED** at aggregate level.

- Whisper wins WER on 10 of 15 samples.
- Gemini mean WER is dominated by two catastrophic insertion loops (Audio3 and Audio9).
- Median Gemini WER (140%) still exceeds Whisper (99%).

H3 — Gemini better preserves Mayan-contact content: **PARTIALLY SUPPORTED** (qualitative).

- On Audio8, Audio10, Audio12, and Audio14, Gemini outputs identifiable Yucatec Maya or place-name forms where Whisper produces generic Spanish loops.
- WER undercounts this advantage because references are in Spanish while Gemini hypotheses are in Maya orthography.

Statistical note. N = 15 is illustrative and not powered for statistical significance testing; no p-values are reported. Claims are robust in direction but not in precise effect size.

5. Discussion and Limitations

AI Safety Implications

For Global South communities, ASR failure is not a convenience issue—it is epistemic exclusion. When a Yoremes fire myth becomes 'un menabolo en su casa' (Whisper) or a 32,000-word insertion loop (Gemini), archival and voice-interface systems propagate false cultural records with no visible error signal.

This project is an empirical demonstration that AI systems can be unsafe for specific communities even when aggregate benchmark performance appears acceptable. Standard safety metrics—mean accuracy, WER on standard benchmarks—do not capture this risk because the affected group is a minority in the evaluation data. Risk is not distributed uniformly: failures concentrate in communities that already face the greatest barriers to public services.

Three policy-relevant findings:

- **Neither model is deployment-ready** for this speech type without domain adaptation. Releasing either system as a general-purpose voice tool for Yucatán would constitute a silent exclusion of rural speakers.
- **Deletion-heavy failures (Whisper)** erase content silently. The API returns an HTTP 200 with a short hallucinated phrase—there is no error flag. Downstream systems have no way to know the transcription is wrong without access to the original audio.
- **Insertion-heavy failures (Gemini)** create fluent false transcripts. A 32,000-word Maya-language output is dangerous for archival systems: it looks like a real transcription.

Limitations

Table 5. Methodological limitations and mitigations.

Limitation	Impact	Mitigation / Future Work
N = 15	Insufficient for significance testing	Expand corpus; add double annotation
Spanish references vs. Maya output	WER penalizes valid Maya-language tokens	Maya-aware metrics; bilingual ground truth
Two Gemini outliers	Severely skew mean WER	Insertion-loop detection filter pre-export
Single run per model	Residual variance not quantified	Multi-seed report; inter-annotator agreement
No end-to-end kiosk test	Downstream harm not directly measured	Run analisis_alucinaciones.py → LLM pipeline

Future Work

- Fine-tune or few-shot adapt Whisper/Gemini on collected Yucatecan field data with verified human transcription.
- Implement a repetition-detection filter that truncates insertion loops before WER calculation, producing a cleaner Gemini score.
- Run human side-by-side rating (fluency, cultural fidelity) alongside WER to capture Gemini's qualitative advantage on Maya-contact segments.
- Deploy an end-to-end test: ASR output → rural kiosk LLM → service response, to quantify the full domino effect of acoustic bias.
- Establish WER thresholds as deployment criteria: any ASR model with median WER > 0.20 on a representative dialect corpus should not be deployed without community validation.

6. Conclusion

We built and executed DialectSafe: Bridging the ASR Gap, a deterministic benchmark comparing Whisper-large-v3 and Gemini 1.5 Flash on 15 rural Yucatecan and Mayan-contact audio samples with human ground truth. Both models fail catastrophically—median WER of 99–140%—confirming digital deafness as a first-order AI safety issue for Global South communities. Whisper primarily omits and genericizes, silently erasing

reference content. Gemini sometimes captures Mayan-contact lexical forms but produces massive hallucinated insertions on two segments.

The aggregate superiority of Gemini is not established. Localized gains on Maya-rich segments are real but fragile, and the insertion-loop failure mode is arguably more dangerous than Whisper's omissions because it produces undetectable false records. The core finding is clear: neither model should be deployed as a voice interface for rural Yucatán without significant domain adaptation backed by community-collected data and community-validated ground truth.

Code and Data

Artifact	Location
Repository	https://github.com/Robertotec2/Global-south-hackathonAI-safety-Equipo-ZINT
Audio corpus	audios_yucatan/Audios/Audio1.mpeg – Audio15.mpeg
Ground truth	transcripciones_oficiales/TranscripcionOficialAudio{N}.txt
Benchmark CSV	benchmark_whisper_vs_gemini.csv
Evaluation script	analisis_comparativo.py
Results CSV	resultados_finales_benchmark.csv
Dependencies	requirements.txt

Author Contributions

Equipo ZINT carried out all phases of the project: audio collection in Yucatán, human transcription of all 15 segments, Colab ASR runs for both Whisper and Gemini, benchmark design, implementation of analisis_comparativo.py, statistical analysis, and manuscript preparation.

References

Blodgett, S. L., et al. (2020). Language (Technology) is Power: A Critical Survey of 'Bias' in NLP. *Proceedings of ACL 2020*. <https://aclanthology.org/2020.acl-main.647/>

Global South AI Safety Hackathon. (2026). <https://globalsouthhackathon.com/>

Groq. (2026). GroqCloud Models. <https://console.groq.com/docs/models>

Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*. <https://doi.org/10.1145/3571730>

Koenecke, A., et al. (2020). Racial Disparities in Automated Speech Recognition. *PNAS*. <https://doi.org/10.1073/pnas.1915768117>

Radford, A., et al. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of ICML 2023*. <https://proceedings.mlr.press/v202/radford23a.html>

Appendix A — Gemini Linguist System Prompt

"Eres un experto lingüista especializado en el español de la Península de Yucatán y la lengua maya. Escucha este audio cuidadosamente. Transcribe de la forma más exacta posible lo que dice la persona. Si utiliza palabras en maya mezcladas con español, escríbelas correctamente. Devuelve ÚNICAMENTE la transcripción, sin saludos ni explicaciones."

Appendix B — Output Schema

```
resultados_finales_benchmark.csv      columns:      Archivo,      Texto_Oficial,
Transcripcion_Whisper, Transcripcion_Gemini, WER_Whisper, WER_Gemini, Reduccion_Error,
Sustituciones_Whisper, Omisiones_Whisper, Inserciones_Whisper, Sustituciones_Gemini,
Omisiones_Gemini, Inserciones_Gemini
```

Appendix C — Hallucination Examples

Whisper — Audio10 (insertion loop, WER = 1.013):

Output: '¿Tienes un chuy? Sí, tengo un chuy. ¿Tienes un chuy? Sí, tengo un chuy...' [×14 repetitions]

Reference: '¿Estás costurando? Estoy costurando, para venderlo. ¿Cómo te llamas?...' [×14 repetitions]

Gemini — Audio10 (language switch to Maya, WER = 1.716):

Output: 'Pash ka betik chich. Ichuy. Ta choy. Ta nichuy, ta nichuy. Bix a kaba...'

Reference: '¿Estás costurando? Estoy costurando, para venderlo...'

Whisper — Audio14 (extreme loop, WER = 3.947):

Output: '¿Qué haces? Me he hecho un trabajo. ¿Qué haces? Me he hecho un trabajo...' [×41 repetitions]

Reference: 'Domitila es mi nombre, pero me dicen solo Tila...'

LLM Usage Statement

LLMs were used to assist in developing evaluation code (analisis_comparativo.py, pipeline scripts) and in drafting this manuscript (Cursor IDE with Claude). ASR outputs were generated by Whisper-large-v3 via Groq and Gemini 1.5 Flash via Google AI Studio, both run in Colab with deterministic settings. All quantitative claims were computed by analisis_comparativo.py from benchmark_whisper_vs_gemini.csv and cross-verified against resultados_finales_benchmark.csv. The authors take full responsibility for the interpretations, hypothesis assessments, design decisions, and documented limitations reported in this paper.