

Low-Resource Automatic Speech Recognition for Algerian Dialectal Arabic: Fine-Tuning, Regional Generalization, and Evaluation Against Multimodal LLMs

Khier Charif

khier.charif@univ-bba.dz

University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj

Lyazid Sabri

University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj

Research Article

Keywords: Automatic Speech Recognition, Algerian Dialect, Multimodal LLMs, Gemini, GPT-4o, Fine-tuning, Regional Bias

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10022445/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Low-Resource Automatic Speech Recognition for Algerian Dialectal Arabic: Fine-Tuning, Regional Generalization, and Evaluation Against Multimodal LLMs

Khier Charif^{1*} and Lyazid Sabri¹

^{1*} Intelligent Systems and Cognitive Computing Laboratory, Faculty of Mathematics and Information Technology, Mohamed El Bachir El Ibrahimi University of Bordj Bou Arréridj, Bordj Bou Arréridj, Algeria .

*Corresponding author(s). E-mail(s): khier.charif@univ-bba.dz;
Contributing authors: lyazid.sabri@univ-bba.dz;

Abstract

Purpose: Whether massive zero-shot models can supplant targeted fine-tuning for severely under-resourced dialects remains open. This work addresses this for Algerian Arabic (Darja)—a linguistically complex, data-scarce North African dialect—by building and rigorously evaluating purpose-built ASR systems under authentic real-world conditions.

Methods: We fine-tuned three models (MMS-1B, XLS-R-53, Whisper Small) on 800 minutes of spontaneous Algerian speech, integrated with a 5-gram KenLM. Models were evaluated on a challenging two-hour in-the-wild benchmark from social media, spanning four diverse geographic macro-regions. Results were benchmarked against zero-shot acoustic baselines and frontier Multimodal LLMs (Gemini 3.1 Pro, GPT-4o Transcribe). An ablation study examined the regional specialization trade-off.

Results: Fine-tuning reduced WER by up to 38.77%. Achieving 61.58% overall WER, MMS-1B + KenLM outperformed GPT-4o Transcribe by 30 absolute points and matched Gemini 2.5 Flash, yielding ground only to frontier LLMs like Gemini 3.1 Pro. Centre and West dialects posed the greatest challenge across all architectures due to dense French code-switching. Furthermore, the ablation study revealed catastrophic forgetting under narrow regional augmentation, proving geographic diversity in training data outweighs regional depth.

Conclusion: Targeted dialectal fine-tuning on limited, well-curated data remains indispensable, delivering competitive accuracy with full deployment

autonomy. Code-switching persists as a universal bottleneck, and geographic diversity is identified as the primary driver of cross-regional robustness.

Keywords: Automatic Speech Recognition, Algerian Dialect, Multimodal LLMs, Gemini, GPT-4o, Fine-tuning, Regional Bias

1 Introduction

The landscape of Automatic Speech Recognition (ASR) has undergone a fundamental paradigm shift, moving from traditional hybrid systems toward robust End-to-End (E2E) architectures [Prabhavalkar, Hori, Sainath, Schlüter, and Watanabe \(2024\)](#); [Tabrej, Deb, Hakim, Goswami, and Nayeem \(2025\)](#). Recently, multilingual foundation models—such as wav2vec 2.0 [Baevski, Zhou, Mohamed, and Auli \(2020\)](#), MMS [Pratap et al. \(2023\)](#), and Whisper [Radford et al. \(2023\)](#)—have set new benchmarks by leveraging large-scale pre-training. While these systems have achieved remarkable accuracy for high-resource languages, a significant performance gap remains for under-resourced regional dialects where labeled data is scarce. This discrepancy limits the deployment of inclusive digital accessibility and highlights the inadequacy of off-the-shelf solutions when applied to complex linguistic environments.

Within this landscape, Arabic stands out as a particularly challenging domain due to its high dialectal variance [Alhaksan, Al-Turaiki, and Al-Twairesh \(2025\)](#). In this context, the Algerian Dialect (AD) represents one of the most complex frontiers, characterized by rich phonetic diversity and deep variations across different geographical regions. Standard ASR models, typically trained on Modern Standard Arabic (MSA), often fail to accurately transcribe spontaneous Algerian speech. Developing robust ASR for AD necessitates addressing three fundamental constraints: (1) extreme data sparsity, as labeled audio resources are severely scarce compared to MSA [Berrimi \(2024\)](#); [Salhab, Sait, Abusheikh, and Abusheikh \(2025\)](#); [Talafta et al. \(2024\)](#); (2) intra-sentential code-switching, which introduces a complex hybrid linguistic structure [Lachemat et al. \(2024\)](#); and (3) the absence of orthographic standards, leading to high spelling variability [Elleuch, Saidi, Mdhaffar, Estève, and Bougares \(2025\)](#); [Haboussi, Oukas, Zerrouki, and Djettou \(2025\)](#).

To address these limitations, this study presents a multi-dimensional contribution to Algerian dialect ASR that goes beyond mere benchmarking. Our work encompasses the full pipeline: from data collection and system construction to real-world deployment evaluation and systematic ablation.

Concretely, our key contributions are fourfold:

- **System Construction:** We fine-tune three acoustically distinct models—MMS-1B, XLS-R-53, and Whisper Small—on an 800-minute corpus of spontaneous Algerian speech, augmented with a custom 5-gram KenLM language model. This provides a fully reproducible, locally deployable ASR pipeline for Algerian Darja.
- **Real-World Regional Evaluation:** We introduce a challenging in-the-wild evaluation framework spanning over two hours of social media audio across four

geographically and linguistically distinct macro-regions (East, West, Centre, and South). This provides the first systematic geo-linguistic stress test for Algerian dialect ASR.

- **LLM Benchmarking in Context:** We benchmark our systems against leading zero-shot acoustic baselines and frontier Multimodal LLMs (e.g., Gemini 3.1 Pro, GPT-4o Transcribe). This contextualizes the role of targeted fine-tuning against systems trained on orders-of-magnitude more data, explicitly demonstrating where scale alone is insufficient.
- **Ablation on Regional Adaptation:** Through a targeted augmentation experiment using a dedicated 5-hour Eastern corpus, we empirically quantify the trade-off between regional specialization and cross-dialectal generalization. This reveals a clear catastrophic forgetting effect, establishing geographic diversity as a critical design principle for low-resource dialectal ASR.

By providing this integrated analysis, this work establishes a reproducible performance baseline for Algerian speech technologies and offers empirical guidelines for practitioners building ASR systems for North African and similarly under-resourced dialectal contexts.

2 Materials and Methods

2.1 Datasets

To train and robustly evaluate the ASR models, we utilized three distinct datasets: a primary volunteer-sourced dataset for base training, a localized regional augmentation corpus for controlled ablation testing, and an “in-the-wild” social media dataset designed to assess real-world generalization.

2.1.1 Primary Training and Evaluation Dataset

A total of approximately 1,000 minutes of Algerian dialect speech was collected from over 50 volunteer native speakers using their personal smartphones. To mitigate the challenges posed by the lack of standardized orthography for the Algerian dialect, volunteers transcribed the audio themselves adhering to strict phonetic and segmentation guidelines. Specifically, the protocol required segmenting audio into chunks of 3 to 9 seconds, spelling out numbers exactly as pronounced locally (rather than using digits or Standard Arabic forms), transliterating occasional Latin loanwords entirely into Arabic script, and completely omitting punctuation marks, alongside several other specific transcription directives. **Demographically**, the dataset reflects the characteristics of the volunteer recruitment pool, which was primarily conducted within a university setting in the Central Algerian region. Consequently, the age distribution is concentrated between 18 and 22 years, with a notable gender imbalance; female speakers contribute approximately 12.5 hours (748 minutes), while male speakers contribute around 2.8 hours (170 minutes). The remaining 1.9 hours (116 minutes) comprise audio data sourced from the Nasri model repository [Nasri \(2024\)](#).¹ To capture the natural

¹<https://huggingface.co/MohammedNasri/whisper-algerian-dialect>

flow of the dialect, the content consists of unscripted discussions about daily activities, educational experiences (e.g., Baccalaureate and university life), personal hobbies (e.g., cooking, football) and the narration of well-known stories.

Geographical Distribution and Core Hubs: Although the recruitment process was centered around a university pipeline within the Central Algerian region, a detailed demographic and linguistic mapping of the participating volunteers revealed a distinct geographical clustering. The vast majority of the speakers either originated from or retained deep dialectal roots in two specific sub-regions: the northern area of Bordj Bou Arreridj (representing the Eastern macro-region) and the province of Biskra (serving as a gateway to the Southern macro-region). This implicit regional concentration provides a critical baseline characteristic of the primary corpus, which is essential for understanding the model’s regional exposure during the training phase.

This primary dataset was partitioned into 800 minutes for fine-tuning, 100 minutes for validation, and 100 minutes for an *internal, in-domain baseline test*.

2.1.2 Eastern Algerian Augmentation Corpus

To support the targeted regional adaptation experiments, we compiled a specialized 5-hour (300 minutes) Eastern speech corpus. Importantly, this corpus was intentionally designed to focus on the *Far-East* border wilayas (such as Tébessa, El Tarf, Annaba, and Guelma). This targeted collection strategy was adopted because the baseline 800-minute training corpus already possessed a strong representation of the *Near-East* dialect (centered around the Bordj Bou Arreridj student recruitment pool, as detailed in Section 2.1.1). By explicitly reinforcing the Far-East phonetic variations, we aimed to address the deep dialectal gap near the border.

The dataset is divided into two balanced streams:

- **Algerian Street Speech (approx. 3 hours / 183.41 mins):** Sourced directly from YouTube field-journalism programs, where reporters interview citizens across various far-eastern wilayas. This setup guarantees highly spontaneous, unscripted speech, spanning diverse age groups, topics (political, cultural, daily life), and environments. The exact distribution across these 9 wilayas is detailed in Table 1.
- **Cross-Border Tunisian Data (approx. 2 hours / 116.59 mins):** Sourced from the open-access TUNIFRA dataset [Bougares \(2022\)](https://huggingface.co/datasets/fbougares/TUNIFRA)² on Hugging Face. This cross-border data injection is highly justified by the deep phonological and lexical proximity between the Far-East Algerian border dialects (e.g., Tébessa and El Tarf) and Tunisian Arabic, serving as an effective low-resource transfer learning strategy.

2.1.3 In-the-wild Evaluation Dataset

This primary “in-the-wild” evaluation dataset serves as the central benchmark of this study, upon which all fine-tuned models, Foundational ASR baselines, and multimodal LLMs are rigorously evaluated, and from which all final empirical results are reported. To mirror the authentic linguistic and acoustic complexities of daily Algerian life, this macro-regional corpus spanning over 2 hours (comprising approximately 200 audio

²<https://huggingface.co/datasets/fbougares/TUNIFRA>

Table 1 Distribution and durations of the Algerian street-speech component across Eastern Wilayas (highlighting Far-East concentration).

Wilaya (Subfolder)	Duration (Minutes)
Tébessa	31.86
Annaba	30.90
Guelma	29.92
Skikda	20.30
Souk Ahras	19.79
El Tarf	17.86
Khenchela	12.04
Oum El Bouaghi	11.35
Constantine	9.39
Total Algerian East Subset	183.41 mins (3.05 h)
TUNIFRA Cross-Border Subset	116.59 mins (1.95 h)
Total Augmentation Corpus	300.00 mins (5.00 h)

clips ranging from 20 seconds to over 2 minutes, in stark contrast to the short 3- to 9-second segments used during training) was strategically curated from two distinct social media ecosystems:

- **TikTok Streams:** Selected to capture extreme real-world acoustic degradation, characterized by rapid speech rates, dense code-switching (Algerian Darja mixed with French/Standard Arabic), overlapping multi-speaker dialogues, and omnipresent background music or trendy audio filters.
- **YouTube Field Broadcasts:** Sourced from popular Algerian street-journalism and television programs. These recordings capture unconstrained, spontaneous public expressions across a vast spectrum of social, cultural, and political topics, inherently embedded with real-world urban ambient noise.

To ensure a comprehensive geo-linguistic stress test, the clips were meticulously balanced across Algeria’s four geographical macro-regions. For each region, we deliberately sampled native speakers from at least 5 of the largest and most dialectally representative wilayas. This includes major hubs such as *Oran*, *Tlemcen*, and *Mascara* representing the West; *Tébessa* and *Annaba* representing the East; the capital *Algiers* anchoring the Center; and *Adrar* and *Ouargla* representing the South, among others.

By exposing the models to this highly diverse geographical span and its accompanying chaotic acoustic environments, this dataset stands as a strict, zero-shot validation framework that proves the true robustness and real-world applicability of our fine-tuned systems.

2.2 Experimental Setup and Preprocessing

All models were implemented via the PyTorch and Hugging Face ecosystems and fine-tuned on a single NVIDIA Tesla T4 GPU (16GB VRAM), utilizing Mixed Precision (FP16) and Gradient Checkpointing to accommodate hardware constraints.

Prior to training, audio files were resampled to 16 kHz mono. Transcripts underwent a deterministic orthographic normalization protocol to mitigate the dialect’s lack of standardized spelling. This included removing diacritics and Latin characters, unifying Alef and Hamza variants to their bare forms, and standardizing Maghrebi-specific consonants (e.g., mapping phoneme variants like /g/ to Arabic Gaf “g” or Qaf, and /p/ to Baa “b”). To prevent out-of-memory (OOM) errors, audio segments exceeding 15 seconds were excluded (accounting for less than 0.3% of the total data).

2.3 Acoustic Models and Fine-Tuning Strategy

Due to architectural differences, we adopted a model-specific optimization strategy. All models utilized Early Stopping based on validation Word Error Rate (WER).

2.3.1 Wav2Vec 2.0 (XLS-R-53)

Initialized from an Arabic-specific checkpoint [Babu et al. \(2022\)](#); [Baevski et al. \(2020\)](#), this Connectionist Temporal Classification (CTC)-based model underwent full fine-tuning. To strictly align with the tokenizer’s constraints, specialized dialectal characters were temporarily re-mapped. We applied SpecAugment, a learning rate (LR) of 1×10^{-4} , and an effective batch size of 64.

2.3.2 MMS-1B

We utilized the 1-billion parameter Massively Multilingual Speech (MMS) model [Pratap et al. \(2023\)](#), initialized with the Arabic adapter. To maximize adaptation while saving memory, we fully fine-tuned the transformer layers (using the 8-bit AdamW optimizer) while keeping the convolutional feature encoder frozen. The LR was set to 5×10^{-5} with an effective batch size of 32.

2.3.3 Whisper Small

For the Sequence-to-Sequence (Seq2Seq) paradigm, Whisper Small [Radford et al. \(2023\)](#) was fully fine-tuned. The tokenizer was forced to output Arabic transcription tokens. A conservative LR of 1×10^{-5} was applied to prevent catastrophic forgetting of its pre-trained multilingual representations. During validation, sequences were actively generated up to 225 tokens.

2.4 Language Model Integration

To mitigate acoustic ambiguity, a 5-gram language model (LM) was constructed using KenLM [Heafield \(2011\)](#). To prevent data leakage, the LM was trained exclusively on a 3.4 million-word composite corpus combining the ASR training transcripts and crawled social media text, completely isolating the validation and test sets.

For the CTC models (MMS and XLS-R), the LM was integrated via Shallow Fusion [Kannan et al. \(2018\)](#) directly into the beam search decoder utilizing the `pyctcdecode` library [Technologies \(2022\)](#). Decoding hyperparameters (e.g., LM weight α and insertion penalty β) were systematically optimized via grid search.

2.5 Evaluation Metrics

Model performance was benchmarked using the standard Word Error Rate (WER). Additionally, Character Error Rate (CER) was utilized as a vital metric for acoustic fidelity, as it provides a realistic assessment without strictly penalizing phonetically valid orthographic variations inherent to the Algerian dialect.

3 Results

To rigorously assess the generalization and robustness of our models in unconstrained environments, we evaluated their performance on 'in-the-wild' data. For this purpose, we curated a diverse test set sourced directly from Algerian social media platforms (TikTok and YouTube). This evaluation aims to benchmark the real-world capabilities of our locally fine-tuned model against its zero-shot counterpart and other prominent ASR baselines, particularly when confronted with spontaneous speech, background noise, and unseen acoustic conditions.

Table 2 Comprehensive evaluation of ASR systems on real-world Algerian dialect data: Fine-tuning impact vs. SOTA zero-shot models and external baselines.

Model	Setup	WER (%)	CER (%)	WER Reduction (%)
<i>Models fine-tuned in this study (800 min)</i>				
MMS-1B	Zero-Shot	100.58	89.27	–
	Fine-Tuned (Greedy)	66.93	25.87	33.46
	Fine-Tuned (+ KenLM)	61.58	30.98	38.77
Whisper Small	Zero-Shot	110.32	75.31	–
	Fine-Tuned (Greedy)	67.66	29.19	38.67
XLS-R-53	Zero-Shot	94.08	49.12	–
	Fine-Tuned (Greedy)	73.65	27.74	21.72
	Fine-Tuned (+ KenLM)	64.44	29.72	31.51
<i>SOTA Zero-Shot Baselines</i>				
Whisper Large v3	Zero-Shot	90.25	62.74	–
Seamless M4T v2 Large	Zero-Shot	85.45	67.45	–
Whisper Large v3 Turbo	Zero-Shot	94.84	64.05	–
NVIDIA NeMo F.C	Zero-Shot	70.43	35.38	–
<i>External Fine-Tuned Baseline (Darija)</i>				
Whisper Turbo (Darija FT)	External FT	60.63	22.48	–

Analysis and Evaluation of ASR Systems for the Algerian Dialect

Table 2 presents a comprehensive evaluation of ASR systems on real-world Algerian dialect data. By comparing the models fine-tuned in this study with state-of-the-art (SOTA) zero-shot and regional baselines, we can deduce the following key findings:

- **1. The Shortcomings of Giant Zero-Shot Models:** Massive language models failed to adapt to the Algerian dialect in its natural setting. Models such as Whisper Large v3 and Seamless M4T v2 Large recorded high Word Error Rates (WER) of 90.25% and 85.45%, respectively. This underperformance is attributed to the dialect’s complex characteristics, including heavy code-switching and lexical overlap with French and Berber—challenges that general models cannot overcome without domain-specific training.
- **2. Data Quality and Fine-Tuning Outperform Model Size:** The Whisper Small model strongly validates our methodological approach. Despite its compact architectural size, it achieved a 67.66% WER after fine-tuning, vastly outperforming its massive zero-shot counterpart (90.25% WER). This demonstrates that domain adaptation using limited (800 minutes) but highly reliable and high-quality data is far more efficient and effective than blindly relying on out-of-domain models with billions of parameters.
- **3. Radical Reduction in Character Error Rate (CER):** The CER dropped dramatically post-training. For instance, the CER of the MMS-1B model plummeted from 89.27% to 25.87%, while the other fine-tuned models stabilized around 28%. This metric proves that the acoustic models successfully learned to capture Algerian phonemes with high precision. The persistently higher WER is primarily due to orthographic variability and the lack of standardized writing rules, which complicates assembling characters into words that perfectly match the reference text.
- **4. Structural Robustness of the NVIDIA NeMo Model:** The NeMo FastConformer model exhibited exceptional robustness in the zero-shot setting, achieving a 70.43% WER. This suggests a potentially greater diversity in its pre-training data or the high efficiency of its architecture in capturing the acoustic features of unfamiliar languages. Nevertheless, our fine-tuned models maintained a noticeable advantage, reaffirming that fine-tuning remains an inevitable step for real-world practical applications.
- **5. Performance Comparison with Regional (Darija) Baselines:** When comparing our results with the Whisper Turbo model—externally trained on massive Moroccan Darija datasets (60.63% WER)—our proposed model (MMS-1B + KenLM) closely approached this performance ceiling, achieving a 61.58% WER using only 800 minutes of data. This achievement demonstrates that seamlessly integrating acoustic fine-tuning with Language Modeling (LM) is the optimal strategy to bridge the data gap in low-resource dialects.

3.1 Regional comparison across ASR models on Algerian dialect (real-world data)

Table 3 Regional comparison across ASR models on Algerian dialect (real-world data).

Model	Setup	Centre	West	South	East
<i>Fine-Tuned Models (This Work)</i>					
MMS-1B	Fine-Tuned + KenLM	62.50	61.66	60.34	61.90
Whisper Small	Fine-Tuned	71.10	72.79	63.39	61.75
XLS-R-53	Fine-Tuned + KenLM	65.26	64.46	63.03	65.14
<i>Baselines (Before Fine-Tuning)</i>					
MMS-1B	Zero-Shot	100.41	100.80	100.44	100.59
Whisper Small	Zero-Shot	128.40	111.22	98.24	103.78
XLS-R-53	Zero-Shot	94.95	95.53	93.88	91.39
<i>SOTA / External Models</i>					
Whisper Large v3	Zero-Shot	90.54	105.70	81.03	78.85
Seamless M4T v2 Large	Zero-Shot	97.90	92.34	88.17	85.45
Whisper Large v3 Turbo	Zero-Shot	94.07	103.76	86.77	92.25
NVIDIA NeMo FastConformer	Zero-Shot	74.74	73.06	65.11	70.43
Whisper Turbo (Darija FT)	External FT	67.33	62.04	58.34	54.29

Regional Performance Analysis and the Impact of Linguistic Variation

The regional analysis (Table 3) reveals fundamental disparities that reflect the nature of the training data and decoding architectures. Through quantitative and qualitative evaluation of error rates across the four regions, we deduce the following key findings:

- **1. Superior Spatial Robustness and Generalization of (MMS-1B + KenLM):** The exceptional stability of the (MMS-1B + KenLM) model emerges as the most prominent finding, recording the lowest standard deviation (std = 0.91) with a remarkably narrow range (Range = 2.16). This phenomenon is further corroborated by (XLS-R-53 + KenLM), which recorded (std = 1.02). Compared to the Whisper Small model without language model support (std = 5.50), it is evident that KenLM acts as a linguistic “homogenizer” that narrows the geographic bias gap, granting the system high spatial robustness against dialectal variations.
- **2. Geographic Bias and Training Data Impact in Whisper Small:** This model exhibited a clear geographic bias that perfectly aligns with its 800-minute training corpus. The figures reveal a **9.37-point** gap between the average performance in the South and East (**62.57%**) and the Center and West (**71.94%**). This disparity is logically explained by the composition of our training data, which was predominantly sourced from the Centre-East and the South (such as Bordj

Bou Arreridj, Biskra, and their surrounding areas). Linguistically, the Centre-East dialect aligns closely with Eastern phonetic structures, and both these regions largely lack the heavy French code-switching characteristic of the true Centre (the Algiers region). Consequently, the model excelled in the East and South, but struggled in the Center and West due to the high density of French loanwords in those regions, which presented a direct challenge to a model not originally trained to process them.

- **3. The Linguistic Paradox of the Moroccan Darija Model (Darija FT):** Despite the geographic expectation that this external model would excel in the West due to border proximity, it paradoxically achieved its best performance in the East (**54.29%**), followed by the South (**58.34%**), and declined in the West (**62.04%**). This can be attributed to two probable factors: First, linguistically, the East and South retain purer Arabic morphological structures with rare French interference. Second, from a data perspective, there is a strong possibility that the external model’s training data contained Algerian texts (particularly Eastern), which is highly plausible given the author’s non-disclosure of the exact datasets utilized for training.
- **4. Divergent Collapse of Zero-Shot Models:** Zero-shot models displayed contrasting behaviors prior to fine-tuning. (Whisper Small Zero-Shot) recorded the highest variance (Range = 30.16), completely collapsing in the Center (**128.40%**) while remaining relatively acceptable in the South (**98.24%**). Additionally, a statistical anomaly emerged in (Whisper Large v3) with a massive **26.85-point** gap between its best region and the West (**105.70%**). Conversely, (MMS-1B Zero-shot) showed a consistent and total failure across all regions with a narrow range (Range = 0.39) and error rates near **100%**. This indicates it was a complete “blank slate” regarding the Algerian dialect, making its radical improvement post-fine-tuning a doubly significant success for our adopted approach.

3.2 Qualitative Error Analysis

3.2.1 Structural Error Profile and Model Personalities

The granular error distributions in Figure 1 reveal distinct structural “error personalities” beyond absolute counts. First, insertion patterns highlight architectural differences between CTC and autoregressive frameworks. Whisper Small exhibits a high propensity for hallucination, registering 753 insertions (peaking at 5% in the West) due to its global sentence-level attention, as illustrated in Figure 1. Conversely, MMS-1B and XLS-R-53 are structurally constrained by CTC frame-level alignments, preventing token emission without acoustic evidence. While KenLM further suppresses insertions (193 in MMS-1B), this resistance is fundamentally rooted in the CTC architecture.

Second, an eco-linguistic footprint appears via deletions in the Central region (*alg_centre*). Incorporating KenLM introduces contextual penalties under intense code-switching. As demonstrated in Figure 1, MMS-1B and XLS-R-53 suffer from anomalous deletion spikes in the Central zone of 30% and 24% respectively—significantly outstripping other regions (where XLS-R-53 ranges between 16–19%). Because the

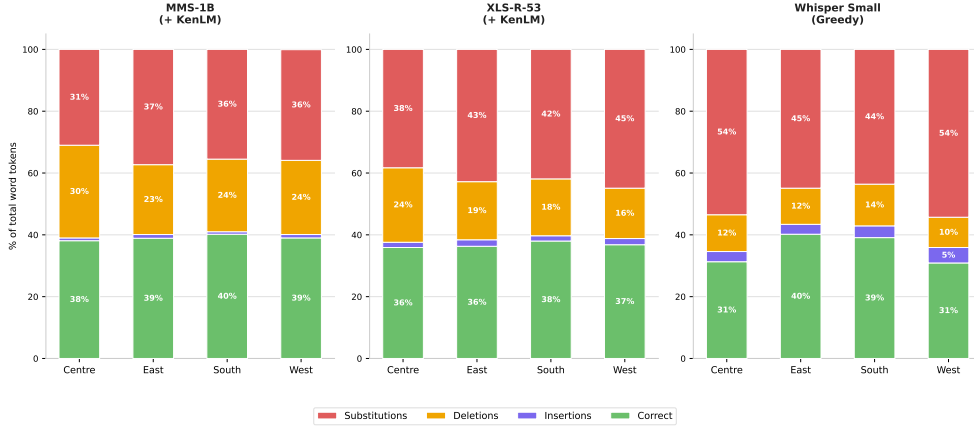


Fig. 1 Word-level Error Breakdown by Region and Model.

Algiers dialect heavily blends French loanwords, KenLM flags these terms as Out-of-Vocabulary (OOV) relative to its Arabic-centric training corpora, systematically omitting foreign tokens to maximize Arabic syntactic coherence.

Third, global attention frameworks prove highly sensitive to dialectal drift. In the West (*alg.west*), Whisper Small exhibits a catastrophic 54% substitution rate alongside its highest insertions, a trend clearly visible in Figure 1. Whisper’s macro-level sentence attention suffers from global contextual drift under Western phonetic and lexical variations, causing arbitrary decoder guessing. In contrast, MMS-1B maintains a stable 24% deletion rate across the regions, proving that phone-based CTC representations preserve local acoustic integrity against dialectal shifts better than autoregressive context guessing.

3.2.2 The Character Error Rate (CER) Paradox under Language Model Rescoring

A critical evaluation discrepancy emerges under language model rescoring: while KenLM decoding substantially reduces the Word Error Rate (WER), it paradoxically degrades the Character Error Rate (CER) due to a systematic *substitution-to-deletion shift*. Incorporating KenLM into MMS-1B reduces word substitutions by 2,130 errors but triggers a surge of 1,155 deletions; a mirroring trend occurs in XLS-R-53 where substitutions drop by 2,187 while deletions rise by 443.

Under greedy decoding, classifiers emit phonetically close but orthographically flawed substitutes. In highly inflected Algerian Arabic, these words share substantial character-level overlap (matching roots, prefixes, or suffixes) with the reference, incurring minimal CER penalties. However, when KenLM overrides these hypotheses due to contextual incongruence, it converts them into complete word omissions. Because a deleted word penalizes its *entire character length* in the CER numerator without partial matching credit, this structural conversion severely inflates the character-level error metric while optimizing word-level lexical consistency.

3.3 Analysis of Multimodal LLMs Performance on Algerian ASR

Table 4 Overall and regional performance (WER% / CER%) of zero-shot Multimodal LLMs (Google Gemini and OpenAI GPT-4o) on Algerian social media audio.

Model	Metric	Overall	Centre	West	South	East
<i>Google Gemini Models</i>						
Gemini 2.5 Flash	WER	60.48	70.76	46.17	44.55	87.17
	CER	30.36	37.25	17.40	15.49	57.35
Gemini 2.5 Pro	WER	32.41	32.93	33.02	32.05	31.43
	CER	10.47	11.38	10.26	10.58	9.74
Gemini 3.1 Pro	WER	25.75	55.11	20.02	17.08	12.55
	CER	11.89	32.01	7.17	5.86	4.72
<i>OpenAI GPT-4o Audio Model</i>						
GPT-4o Transcribe	WER	91.62	96.90	120.99	55.25	85.82
	CER	61.28	74.60	79.82	24.09	63.51

1. Overall Performance:

Google’s Gemini models significantly outperformed the commercial, paid API model, GPT-4o Transcribe. Gemini 3.1 Pro achieved the state-of-the-art result with the lowest Word Error Rate (WER) at 25.75%. In stark contrast, the paid GPT-4o Transcribe model failed drastically, recording a staggering 91.62% overall WER, rendering it unsuitable for zero-shot verbatim transcription.

2. Regional Disparities:

Performance varied heavily depending on the regional dialect. Gemini 3.1 Pro excelled in the East (12.55% WER) and South (17.08% WER), where dialects are structurally closer to Modern Standard Arabic (MSA). Conversely, the Centre region challenged all models—due to heavy French code-switching—pushing Gemini 3.1 Pro’s error to 55.11%. Strikingly, the paid GPT-4o Transcribe model completely collapsed in the fast-paced Western dialect, exceeding 120% WER.

3. Alignment Errors Breakdown:

- **Substitutions (MSA Bias):** Gemini models predominantly suffered from substitutions. They tend to “over-correct” dialectal terms into MSA equivalents to preserve meaning, which violates strict verbatim rules.
- **Insertions (Linguistic Hallucinations):** The catastrophic failure of the paid GPT-4o Transcribe (> 100% WER) was driven by massive insertions. When the model loses the dialect context, it hallucinates entire MSA sentences to fill the acoustic gaps.

- **Deletions:** Lighter models (like Gemini 2.5 Flash) frequently dropped filler words and struggled to keep up with fast speech rates.

3.4 Comparative Analysis: Fine-Tuned Acoustic Models vs. Zero-Shot Multimodal LLMs

Comparing our fine-tuned acoustic models (trained on just 800 minutes of Algerian speech) against massive zero-shot Multimodal LLMs reveals significant insights into the current landscape of dialectal ASR.

- **The Efficiency of Local Fine-Tuning vs. Commercial APIs:** A striking finding is the competitiveness of small, domain-specific models against massive commercial systems. Our fine-tuned **MMS-1B + KenLM** (trained on just 800 minutes) achieved an overall WER of 61.58%, successfully matching the performance of **Gemini 2.5 Flash** (60.48% WER) and drastically outperforming a dedicated commercial STT API like **GPT-4o Transcribe** (91.62% WER). This demonstrates that targeted acoustic fine-tuning can effectively enforce strict verbatim dialectal transcription, circumventing the over-normalization tendencies and linguistic hallucinations of general-purpose LLMs.
- **Data Scale vs. Model Architecture:** **Gemini 3.1 Pro** achieved the overall state-of-the-art result (25.75% WER), significantly outperforming our fine-tuned models. However, this superiority is not solely an inherent architectural advantage, but rather a reflection of the astronomical disparity in training data volume. While our models were strictly constrained to 800 minutes, Gemini 3.1 Pro leverages millions of hours of multimodal data, allowing its contextual reasoning to predict missing or highly co-articulated dialectal words. The fact that Gemini 2.5 Flash could not surpass our MMS-1B model proves that zero-shot superiority is not a universal trait of LLMs, but rather exclusive to the largest and most advanced iterations.
- **Regional Stability vs. Contextual Volatility:** Our fine-tuned models exhibited remarkable regional stability. For instance, MMS-1B + KenLM scored consistently across all testing regions (Centre 62.50%, West 61.66%, South 60.34%, East 61.90%). Since our 800-minute training corpus was predominantly sourced from the Centre-East and the South, the model’s resilience in the underrepresented Western dialect and the true Centre dialect is particularly noteworthy. This suggests that acoustic fine-tuning successfully captured the shared phonetic foundation of Algerian Darja rather than merely memorizing regional vocabulary. Furthermore, the KenLM language model acted as a robust homogenizer, smoothing out cross-regional lexical variance. In contrast, LLMs displayed high contextual volatility, fluctuating wildly depending on the region (e.g., Gemini 3.1 Pro dropping from 12.55% in the East to 55.11% in the Centre, and GPT-4o exceeding 120% in the West).

3.5 Ablation Study: Regional Specialization vs. Cross-Dialectal Generalization in MMS-1B

Motivated by the linguistic distinctiveness of the Eastern Algerian dialect—characterized by phonological proximity to Tunisian Arabic and minimal

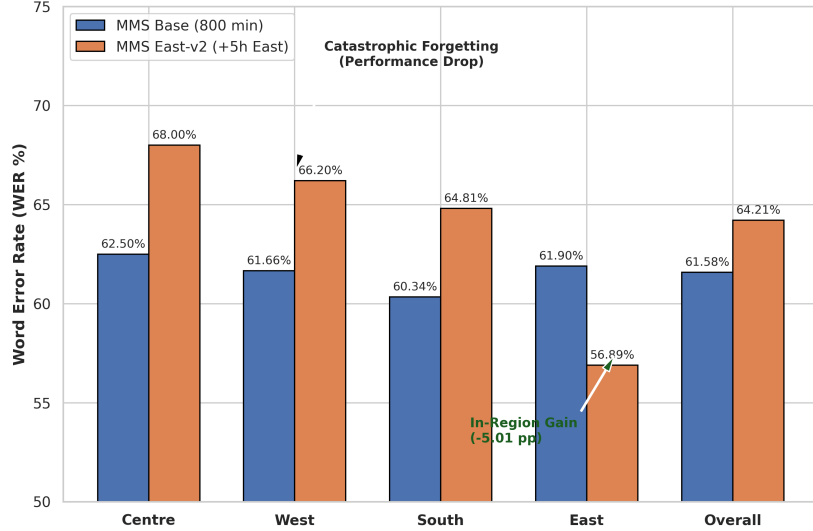


Fig. 2 Regional WER comparison of MMS-1B before and after targeted Eastern dialect augmentation. The significant performance gain in the target region (East) is heavily offset by a uniform regression across all other unseen regions, demonstrating a clear case of catastrophic forgetting.

French code-switching—we conducted a targeted fine-tuning experiment to investigate whether deep regional specialization improves overall dialectal robustness or induces performance regression in unseen regions.

A detailed description of this Eastern regional augmentation corpus is provided in Section 3.1. We augmented the MMS-1B training pipeline with an additional 5 hours of Eastern Algerian speech (East-v2), applied on top of the 800-minute multi-regional base model.

As illustrated in Figure 2, this targeted adaptation yielded a notable improvement in the East region, where the Word Error Rate (WER) dropped significantly (61.90% to 56.89%, $\Delta = -5.01$ pp). This confirms that dialect-specific acoustic exposure provides measurable in-region gains.

However, this localized improvement came at a systematic and severe cost across all other unrepresented regions. Performance degraded uniformly in the Centre (+5.50 pp), West (+4.54 pp), and South (+4.47 pp) regions, leading to a net overall WER increase of 2.63 points for the model.

This contradictory behavioral pattern constitutes a textbook manifestation of *catastrophic forgetting* Kirkpatrick et al. (2017); McCloskey and Cohen (1989). The neural network overrode previously acquired cross-dialectal representations when exposed to a narrower regional data distribution. The finding carries a critical practical implication for low-resource dialectal ASR: **geographic diversity in the training corpus is far more valuable than regional depth**. The original 800-minute corpus, despite its modest volume, captured a broader shared phonetic foundation of Algerian Darja that single-region intensification failed to preserve.

4 Conclusion

This work demonstrates that in the era of large-scale foundation models, targeted dialectal fine-tuning remains not merely competitive—but essential. By constructing a full ASR pipeline for Algerian Darja, from curated data collection to real-world regional deployment, we provide compelling evidence that 800 minutes of well-curated in-domain speech can deliver what billions of parameters cannot: genuine acoustic alignment with an under-resourced dialect in unconstrained real-world conditions.

The inadequacy of scale alone is perhaps most starkly illustrated by the failure of massive zero-shot systems. Models of enormous capacity—Whisper Large v3 (90.25% WER), Seamless M4T v2 Large (85.45%), and Whisper Large v3 Turbo (94.84%)—all collapsed on spontaneous Algerian speech, recording error rates that render them unsuitable for practical deployment. Our fine-tuned models, by contrast, achieved consistent improvements of up to 38.77% WER reduction, with MMS-1B + KenLM (61.58% WER) surpassing every zero-shot acoustic baseline—including the structurally robust NVIDIA NeMo FastConformer (70.43%)—using a fraction of their training resources.

Beyond zero-shot baselines, the broader significance of fine-tuning is further underscored by external reference points. The Whisper Turbo model externally fine-tuned on large-scale Moroccan Darija data achieved 60.63% WER—nearly identical to our MMS-1B result built on just 800 minutes of Algerian speech. This near-parity confirms that dialect-specific fine-tuning consistently outperforms zero-shot systems of any scale, and that even a linguistically adjacent dialect benefits enormously from targeted acoustic adaptation.

In the Multimodal LLM space, our fine-tuned MMS-1B matched Gemini 2.5 Flash (60.48% WER)—a frontier multimodal system representing an entirely different order of training magnitude—while the dedicated GPT-4o Transcribe API, despite its commercial positioning as a state-of-the-art STT solution, recorded a catastrophic 91.62% WER. Nonetheless, intellectual honesty demands acknowledging the absolute frontier. Gemini 2.5 Pro (32.41% WER) and Gemini 3.1 Pro (25.75% WER) did surpass our fine-tuned systems—but the conditions under which they did so are revealing. Outperforming a model trained on 800 minutes of Algerian speech required nothing less than the most advanced frontier multimodal architectures available, trained on estimated multi-trillion token corpora accumulated over years of industrial-scale investment. That such systems are necessary to exceed our baseline is, in itself, a testament to the efficiency of targeted in-domain fine-tuning. Furthermore, their practical deployment for Algerian dialect ASR remains constrained by closed API access, significant inference costs, data privacy concerns, and a complete absence of offline capability. Taken together, these results establish a clear empirical hierarchy: fine-tuning on modest in-domain data substantially outperforms both massive zero-shot acoustic systems and standard commercial APIs, yielding ground only to the most massive, resource-prohibitive frontier models.

Regional evaluation across four Algerian macro-regions identified a universal bottleneck: French code-switching in Centre and West dialects consistently degrades performance across all tested architectures—fine-tuned, zero-shot, and LLM alike. The

KenLM language model proved to be a powerful cross-regional homogenizer for CTC-based architectures, conferring exceptional geographic stability (MMS-1B std = 0.91) that autoregressive models could not match.

The targeted ablation study on Eastern dialect augmentation produced a textbook instance of catastrophic forgetting, establishing a critical design principle: geographic diversity in training data is more valuable than regional depth. The 5.01 pp gain in the target region was offset by uniform degradation across all remaining regions, confirming that the original 800-minute multi-regional corpus captured a broader shared phonetic foundation that single-region intensification could not preserve.

These findings carry direct implications for the broader challenge of building ASR systems for under-resourced languages and dialects. Scale is not a substitute for domain specificity; and for communities whose linguistic identity is absent from the training data of frontier models, locally built, carefully curated systems remain the most viable path to genuine digital inclusion.

Future work will target the persistent code-switching gap through dedicated multilingual training strategies for Centre and West dialects, and will investigate continual learning techniques to enable regional adaptation without catastrophic forgetting.

Statements and Declarations

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

The authors declare no competing interests.

Compliance with Ethical Standards

Research Involving Human Participants

This study involved the collection of voice recordings from human volunteer participants for the purpose of constructing a speech corpus for Algerian Dialectal Arabic. All procedures performed were in accordance with the ethical standards of the institution at which the study was conducted.

Informed Consent

Informed consent was obtained from all individual participants included in the study prior to data collection. Participants were informed of the purpose of the recordings, the intended research use of their data, and the conditions under which their data would (and would not) be shared, in accordance with the restrictions described in the Data Availability statement.

Data Availability

To support transparency and reproducibility, the evaluation benchmark, a dataset sample, and the fine-tuned model checkpoints developed in this study have been made publicly available on Hugging Face under the user profile <https://huggingface.co/khier12>.

Specifically, the complete 2-hour "in-the-wild" social media evaluation benchmark is hosted at https://huggingface.co/datasets/khier12/In_the_wild_Evaluation_Dataset. Additionally, a 5% representative sample of our primary volunteer training corpus is available at <https://huggingface.co/datasets/khier12/algerian-train-sample-5pct>.

However, the main training repositories—specifically the complete 800-minute baseline fine-tuning corpus and the 5-hour Eastern Algerian augmentation corpus—cannot be publicly distributed. These larger sets consist of raw, unscripted human voice recordings that carry sensitive biometric features. The informed consent obtained from the volunteer participants strictly limits the use of their raw audio data to internal research purposes and explicitly prohibits public dissemination to safeguard their anonymity.

Furthermore, the fully fine-tuned acoustic models can be accessed via the following repositories:

- https://huggingface.co/khier12/800min_mms-1b_Full_FT_Algerian_Dialect
- https://huggingface.co/khier12/800min_whisper_small_FT_Algerian_Dialect
- https://huggingface.co/khier12/800min_wav2vec_xls-r53_FT_Algerian_Dialect
- <https://huggingface.co/khier12/mms-1b-algerian-east-v2>

Code Availability

The source code and scripts for model fine-tuning and evaluation are currently available from the corresponding author upon reasonable request.

Author Contributions

Both authors contributed to the study conception and design. Data collection, model fine-tuning, evaluation, and analysis were performed by Khier Charif. The first draft of the manuscript was written by Khier Charif, and Lyazid Sabri supervised the work and critically revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgments

The authors would like to express their sincere gratitude to the volunteer participants who contributed their voice recordings and transcriptions, without whom the construction of the primary Algerian dialect corpus would not have been possible.

References

- Alhaksan, A., Al-Turaiki, I., Al-Twaires, N. (2025). Arabic dialect identification: A survey. *Proc. 31st int. conf. comput. linguist. (coling)* (pp. 4578–4595).

- Babu, A., Wang, C., Tjandra, A., Kushman, K., Gabriel, F., Evensen, R., ... Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of interspeech* (pp. 2278–2282). ISCA.
- Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, ,
- Berrimi, M. (2024). *Deep models for generating and understanding textual Arabic data* (PhD Thesis, Ferhat Abbas University Sétif 1). Retrieved from <http://dspace.univ-setif.dz:8888/jspui/bitstream/123456789/4261/1/Final-Thesis-Mohamed-Berrimi.pdf>
- Bougares, F. (2022). *TUNIFRA: A speech corpus for Tunisian Arabic*. <https://huggingface.co/datasets/fbougares/TUNIFRA>. (Accessed: 2026-06-09)
- Elleuch, H., Saidi, Y., Mdhaftar, S., Estève, Y., Bougares, F. (2025). Elyadata & lia at nadi 2025: Asr and adi subtasks. *Proc. 3rd arabic nat. lang. process. conf. (wanlp)* (pp. 762–766).
- Haboussi, S., Oukas, N., Zerrouki, T., Djettou, H. (2025). Arabic speech recognition using neural networks: concepts, literature review and challenges. *Journal of Umm Al-Qura University for Applied Sciences*, , <https://doi.org/10.1007/s43994-025-00213-w>
- Heafield, K. (2011). Kenlm: Faster and smaller language model queries. *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197).
- Kannan, A., Wu, Y., Nguyen, P., Sainath, T.N., Chen, Z., Prabhavalkar, R. (2018). An analysis of incorporating an external language model into a sequence-to-sequence model. *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526,
- Lachemat, H.E.-O., Abbas, A., Oukas, N., El Kheir, Y., Haboussi, S., Chowdhury, S.A. (2024). Cafe: A novel code-switching dataset for algerian dialect, french, and english. *Proc. ACM Meas. Anal. Comput. Syst.*, 37(4), 1–24,
- McCloskey, M., & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*,

- Nasri, M. (2024). *Whisper fine-tuned for Algerian Dialect*. Hugging Face Repository. Retrieved from <https://huggingface.co/MohammedNasri/whisper-algerian-dialect> (Accessed: Feb. 2026)
- Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S. (2024). End-to-end speech recognition: A survey. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32, 325–351,
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... others (2023). Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.09688*, ,
- Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning (icml)* (pp. 28492–28518).
- Salhab, M., Sait, S., Abusheikh, M., Abusheikh, H. (2025). Munsit at nadi 2025 shared task 2: Pushing the boundaries of multidialectal arabic asr with weakly supervised pretraining and continual supervised fine-tuning. *Proc. 3rd arabic nat. lang. process. conf. (wanlp)* (pp. 734–739).
- Tabrej, M.S., Deb, K.J., Hakim, M.A., Goswami, S., Nayeem, M. (2025). Automatic speech recognition in the modern era: Architectures, training, and evaluation. *arXiv preprint arXiv:2510.12827*, ,
- Talafha, B., Kadaoui, K., Magdy, S.M., Habiboullah, M., Chafei, C.M., et al. (2024). Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*, ,
- Technologies, K. (2022). *pyctcdecode: Ctc decoding with language models*. <https://github.com/kensho-technologies/pyctcdecode>. GitHub.