fihrIA-ASR: State-of-the-Art Speech Recognition and Semantic Evaluation for Moroccan Darija

Anas Belfadil[1,2,3] **Ismail Berrada**[1,4]

[1]fihrIA [2]Barcelona Supercomputing Center (BSC) [3]Universitat Politècnica de Catalunya (UPC)
[4]Mohammed VI Polytechnic University (UM6P)

Abstract

Speech-language models now rival human transcription accuracy on standardized, high-resource languages, yet they remain brittle on the unstandardized, code-switched vernaculars spoken by billions of people daily. Moroccan Darija—spoken by more than 35 million people who routinely interleave Arabic, French, and English without a codified orthography—epitomizes this divide. Progress is impeded by two coupled bottlenecks: general-purpose multilingual tokenizers severely fragment dialectal morphology, inflating decoding latency, while surface edit-distance metrics such as Word Error Rate (WER) conflate benign spelling and script variation with acoustic recognition failure. Here we address modeling and evaluation jointly by pairing a dialect-adapted, open-weight speech recognition family (**fihrIA-ASR**; 0.6B and 1.7B parameters) with a human-validated, span-level **Semantic Error Rate (SER)** framework. Trained on 1.51 million decontaminated clips ($\approx$ 10,000 hours) under a consistent bilingual script convention, our 89,000-token Darija tokenizer cuts subword fertility by 42% (2.29 $\rightarrow$ 1.32 tokens per word) and accelerates server inference by $1.65\times$ at unchanged accuracy. Across 2,006 native-speaker annotations, large language model (LLM) SER correlates strongly with human judgment (Spearman $\rho = 0.71$), reduces calibration error nearly seven-fold relative to WER (4.8 versus 33.2 percentage points), and outperforms 11 competing metrics. On held-out and external benchmarks, fihrIA-ASR 1.7B outperforms larger open-weight systems such as mistralai/Voxtral-Mini-4B-Realtime-Arabic (6.09% mean SER; 6.24% and 8.71% CER on DarijaCS Prism and Probe) while running $7\times$ faster, and our 4-bit 0.6B counterpart transcribes speech at $6\times$ real time on a laptop CPU—providing a transferable blueprint for equitable, low-latency speech technology across unstandardized languages.

Introduction

Most of the world's spoken communication occurs in vernaculars that lack a codified orthography and routinely blend multiple languages within a single utterance. Yet modern automatic speech recognition (ASR) has been engineered almost exclusively for high-resource languages whose standardized writing systems provide a nearly one-to-one mapping from acoustics to text. Moroccan Darija—the everyday spoken vernacular of more than 35 million people across Morocco and its diaspora—sits squarely at this fault line[1–3]. In natural conversation, Moroccan Arabic morphology intertwines fluidly with French and English vocabulary, and identical spoken words admit multiple valid written realizations: borrowed terms alternate between Latin and phonetic Arabic scripts ("potassium" versus لبوتاسيوم), prepositions and conjunctions attach to or detach from adjacent nouns (فستانفورد versus في ستانفورد), and numerals alternate between digits and multi-word phrases ("40" versus أربعين).

This linguistic elasticity creates two coupled bottlenecks that stall dialectal ASR. First, general-purpose multilingual speech decoders[4] are starved of diverse, decontaminated dialectal audio[1,5] and allocate little vocabulary capacity to Arabic-script vernacular morphology. Ordinary Darija words fragment into long subword chains (2.29 tokens per whitespace word in the Qwen3-ASR base vocabulary), inflating autoregressive decoding latency on servers and precluding real-time deployment on edge hardware. Second, standard evaluation metrics break down. Word and Character Error Rates (WER and CER) compute normalized string edit distance against a single reference transcript, penalizing legitimate variation in script choice, clitic spacing, or vowel spelling as heavily as a completely misrecognized word[6–8]. Prior remedies—spanning social-text variant dictionaries[9], script transliteration[10], LLM-generated variant lattices[11], and contextual embedding scores[12–15]—partially soften isolated word mismatches but struggle when multi-word clitic splits or numeral expansions desynchronize token alignments across scripts[16–18].

Here we show that dialectal ASR modeling and evaluation are best solved as a single, co-designed problem: **training benefits from a consistent script convention and a dialect-matched vocabulary that concentrate probability mass and compress token sequences, whereas evaluation requires a semantic metric invariant to surface conventions**. We curate 1.51 million strictly video-decontaminated Darija clips ($\approx$ 10,000 hours) transcribed under an explicit bilingual script convention, replace the multilingual vocabulary with an 89,000-token Darija byte-level BPE tokenizer, and fine-tune open-weight encoder–decoder models at 0.6B and 1.7B parameter scales (**fihrIA-ASR**). In parallel, we formalize a span-level **Semantic Error Rate (SER)** protocol[19] and validate LLM judges against 2,006 native-speaker annotations. Controlled ablations confirm that dialect-adapted tokenization accelerates server decoding by $1.65\times$–$1.67\times$ without sacrificing accuracy, while span-level SER reduces human-calibration error nearly seven-fold relative to WER and outperforms 11 alternative metrics. Together, these advances deliver state-of-the-art open-weight efficiency—running $7.1\times$ faster than Voxtral Mini 4B[20] on a single H100 GPU and $5.98\times$ faster than real time in 721 MB on a consumer laptop CPU—while establishing a rigorous evaluation standard for unstandardized speech.

Results

Dialect-adapted tokenization accelerates decoding without sacrificing accuracy

To overcome data scarcity without risking evaluation leakage, we assembled 36,657 hours of public Moroccan YouTube audio across 460 channels spanning everyday conversation, humor, culture, politics, and business, with specialist scientific and engineering domains in the long tail (**Fig. 1a**). After voice-activity segmentation and strict video-level decontamination—purging all 1,848 clips originating from any source video represented in our held-out evaluation pool—we obtained 1,508,646 supervised clips ($\approx$ 10,000 hours) transcribed with Gemini under an explicit bilingual convention: Moroccan Darija and Arabic words appear in Arabic script, whereas spoken French and English spans retain standard Latin orthography (Methods). Although external human transcripts often transliterate loanwords into Arabic script, enforcing a single script convention during training prevents the decoder from dividing probability mass across orthographically disjoint token sequences for identical acoustic inputs.

In encoder–decoder speech architectures such as Qwen3-ASR[4], the audio encoder processes a waveform in a single forward pass, whereas the language-model decoder generates text autoregressively; end-to-end latency therefore scales linearly with output token length. Because Qwen3-ASR's base 151,936-token vocabulary is tailored to high-resource languages, it fragments Arabic-script Darija into 2.29 subwords per whitespace word across our corpus (**Fig. 1b**). To eliminate this bottleneck, we trained an 89,000-token byte-level byte-

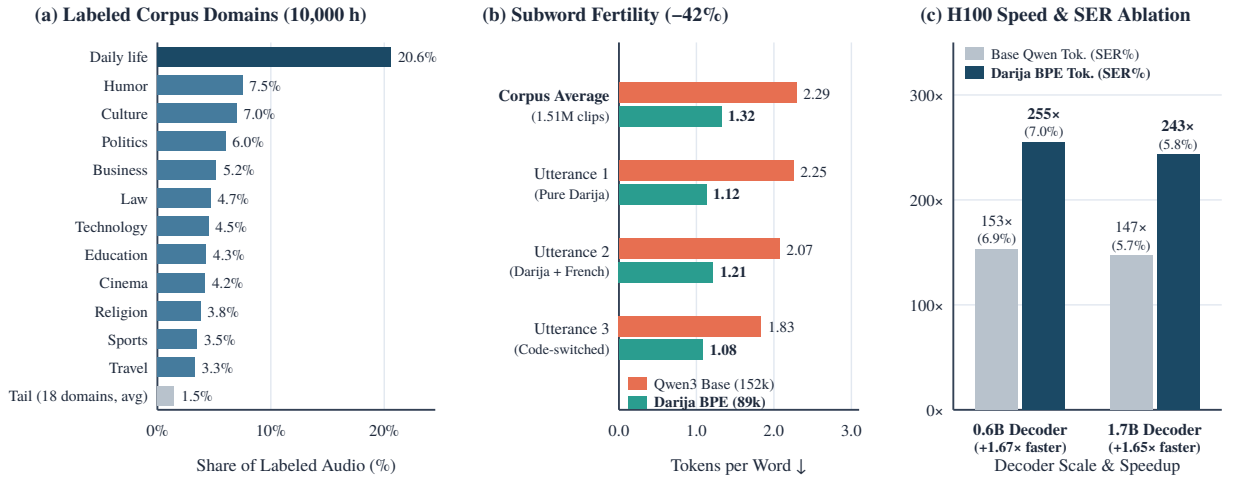


Figure 1: Dialect-adapted BPE tokenization compresses Darija sequences by 42% and accelerates decoding without accuracy loss. a, Topical distribution of the 10,000-hour labeled training pool across 460 Moroccan YouTube channels (1,508,646 decontaminated clips); conversational domains dominate while specialist engineering domains sit in the long tail. **b**, Subword fertility (tokens per whitespace word) across the full corpus and three representative utterances, comparing the 151,936-token Qwen3-ASR base vocabulary against our 89,000-token Darija BPE vocabulary. **c**, Controlled H100 batch-32 ablation at 0.6B and 1.7B parameter scales showing $1.67\times$ and $1.65\times$ real-time throughput gains with SER (parentheses) preserved within 0.1 percentage points.

pair encoding (BPE) tokenizer on our Darija training transcripts supplemented with French and English text. The adapted vocabulary reduces subword fertility by 42% (from 2.29 to 1.32 tokens per word), compressing representative Darija and code-switched utterances from 18 to 9 tokens and from 29 to 17 tokens, and halving elided French forms such as "l'économie" and "l'objectif" from 4 tokens to 2. We transferred pretrained embeddings for all 33,762 shared tokens, mean-initialized the 55,238 new dialectal tokens, and stabilized adaptation via a two-stage schedule that warms up token embeddings before fine-tuning the full decoder with a frozen audio encoder (Methods).

To isolate the causal effect of vocabulary replacement from training data, we trained matched control models at both 0.6B and 1.7B scales using identical supervision, augmentation, and optimization schedules while retaining the base 151,936-token Qwen vocabulary. Across both scales, replacing the tokenizer leaves recognition fidelity virtually identical (ΔWER ≤ 0.3 percentage points [pp]; ΔSER $= 0.1$ pp) while accelerating batch-32 H100 inference by $1.67\times$ at 0.6B ($153\times \rightarrow 255\times$ real time) and $1.65\times$ at 1.7B ($147\times \rightarrow 243\times$ real time; **Fig. 1c**)—approaching the theoretical $1/(1-0.42) = 1.72\times$ autoregressive speedup ceiling.

Span-level semantic evaluation resolves orthographic confounding

Surface edit distance cannot reliably separate orthographic convention from acoustic recognition quality in unstandardized speech. In our human-annotated validation cohort, one utterance pair with zero native-speaker semantic errors incurred a WER of 82.4% purely from clitic spacing and loanword transliteration, whereas another pair with 60% genuinely misrecognized words received a nearly identical WER of 81.0%. To decouple meaning from surface spelling, we formalized span-level **Semantic Error Rate (SER)**, which aligns reference (R) and hypothesis (H) word spans and partitions surface discrepancies into meaning-preserving *acceptable variations* (zero penalty: dialectal spelling, clitic splits, cross-script loanwords, punctuation, diacritics, and equivalent numeral forms) and meaning-altering *semantic errors* (deletions, insertions,

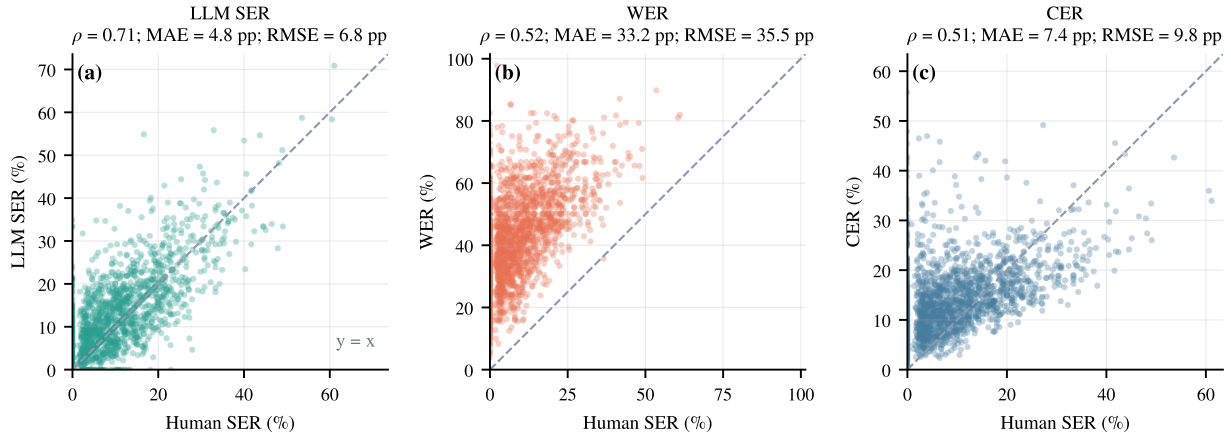


Figure 2: Span-level LLM SER tracks native-speaker judgments and eliminates orthographic inflation across $n = 2{,}006$ matched pairs. Each panel plots an automatic metric against human SER (%), with dashed diagonals ($y = x$) indicating ideal calibration. **a**, LLM SER (GPT-5.6 Luna) aligns closely with native-speaker judgments ($\rho = 0.71$, MAE $= 4.8$ pp). **b**, WER exhibits severe vertical inflation across the entire error range ($\rho = 0.52$, MAE $= 33.2$ pp). **c**, CER compresses scale relative to WER but remains weakly correlated with semantic errors ($\rho = 0.51$, MAE $= 7.4$ pp).

or substitutions). Utterance-level SER is defined symmetrically on $[0, 100\%]$ as

$$\mathrm{SER}(R, H) = 100 \times \frac{n_{\mathrm{err},R} + n_{\mathrm{err},H}}{W_R + W_H},$$

where W_R and W_H denote whitespace word counts and $n_{\mathrm{err},R}, n_{\mathrm{err},H}$ denote words inside aligned error spans. Because SER is bounded per utterance, an isolated repetition loop cannot drive the corpus average above 100%; we therefore report SER alongside median and micro WER/CER so that semantic fidelity and pathological decoding loops are captured jointly.

To validate automated LLM judges (prompted to emit structured JSON span alignments before computing SER; Methods) against native-speaker perception, seven native Moroccan Darija speakers annotated 2,006 reference–hypothesis pairs spanning 507 unique utterances across five ASR systems (Gemini, an early 1.7B development checkpoint fihrIA-ASR v0.5, MoulSot[21], Gemma-4-E4B, and Gemma-4-E2B; 79.1% of the 2,535 target grid). At the utterance level, our primary judge (GPT-5.6 Luna) achieves a Spearman rank correlation of $\rho = 0.71$ (Pearson $r = 0.76$) and reduces mean absolute calibration error nearly seven-fold relative to WER, from 33.2 pp down to 4.8 pp (+1.9 pp bias versus +32.9 pp for WER; **Fig. 2** and **Extended Data Table ED.1**). Evaluating a second independent judge from a distinct model family (Gemini 3.8 Flash) reproduces this agreement ($\rho = 0.72$, $r = 0.77$; cross-judge $\rho = 0.88$; **Extended Data Table ED.2**). Across all $\binom{5}{2} = 10$ pairwise system comparisons, LLM SER estimates inter-system performance gaps with an MAE of just **1.09 pp** ($r = 0.95$)—more than three times more accurate than WER (3.51 pp) and CER (3.68 pp)—and recovers the exact five-system human ordering in 100% of 2,000 audio-cluster bootstrap resamples, whereas CER swaps MoulSot and Gemma-4-E4B ($\rho = 0.70$; **Extended Data Fig. ED.1**).

The sharpest diagnostic test occurs on the 497 pairs (24.8%) where native speakers marked zero semantic errors (Human SER $= 0\%$). On this error-free subset, WER averages **34.0%** (with 82% of flawless transcripts receiving WER $> 20\%$), whereas LLM SER averages **4.2%** (**Extended Data Table ED.1** and

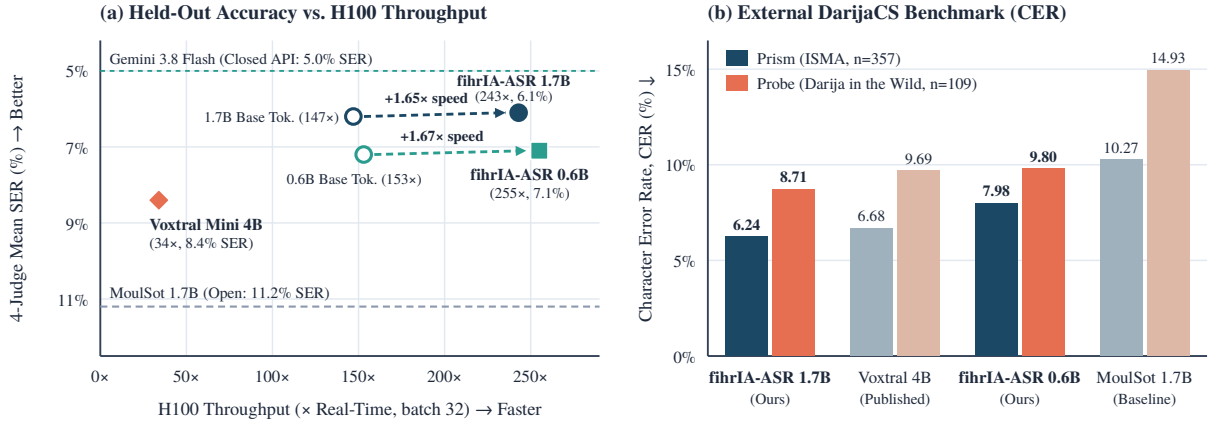


Figure 3: fihrIA-ASR establishes a new accuracy–throughput Pareto frontier on held-out and external Darija benchmarks. a, Accuracy–throughput frontier on the 508-clip held-out benchmark (four-judge mean SER versus NVIDIA H100 batch-32 throughput; lower SER and higher throughput are better). Dashed arrows show the $1.65\times$–$1.67\times$ decoding speedup from replacing the base Qwen3-ASR tokenizer with our Darija BPE vocabulary at fixed model scale. **b**, Character Error Rate (CER) on the external DarijaCS benchmark across Prism (ISMA, $n = 357$) and Probe (Darija in the Wild, $n = 109$), comparing our 16-bit models against Voxtral Mini 4B's published scores and our local evaluations of MoulSot (Methods).

Extended Data Fig. ED.2). Qualitative case studies (**Extended Data Section 1**) illustrate why: LLM SER pardons prepositional clitic splitting (WER $= 82.4\%$, SER $= 3.8\%$), digit-versus-word numeral expansion (WER $= 78.6\%$, SER $= 0.0\%$), and Latin-script French code-switching (WER $= 75.6\%$, SER $= 11.4\%$), while still penalizing genuine lexical or morphological shifts. Across a comprehensive benchmark of **13 automatic metrics** spanning LLM-aware judges, rule/dictionary normalizers, contextual embeddings, and surface edit distances (**Extended Data Fig. ED.3** and **Extended Data Table ED.3**), LLM SER outperforms every alternative on both rank correlation ($\rho = 0.71$ versus 0.60 for runner-up SN-WER [10]) and calibration error (MAE $= 4.8$ pp versus 7.4 pp for CER and 9.1–34.3 pp for all others). Pre-alignment normalizers (SN-WER, Norm-WER, LLM-WER) retain high calibration error (22.8–25.2 pp) because multi-word clitic splits and numeral expansions desynchronize Levenshtein alignments, whereas aggressive variant-lattice relaxation (OIWER-WER [11], WERd-proxy [9]) lowers MAE to 9.4 pp at the cost of collapsing rank correlation to $\rho = 0.35$.

Open-weight accuracy, multi-judge consensus, and external generalization

We next benchmarked fihrIA-ASR (0.6B and 1.7B) against four open-weight baselines (Voxtral Mini 4B Realtime Arabic [20], MoulSot 1.7B [21], Gemma-4-E4B, and Gemma-4-E2B) and three closed commercial APIs (Gemini 3.8 Flash, GPT-Audio-Mini, and Inkling-Small) on our 508-clip decontaminated held-out test set (≈ 2.05 hours). To guard against single-provider judge bias, every transcript was scored independently by four LLM judges from distinct model families—GPT-5.6 Luna, Gemini 3.8 Flash, GLM-5.3, and Claude Sonnet 5—and ordered by unweighted mean SER (**Fig. 3a**, **Fig. 4a**, and **Extended Data Table ED.4**).

Across the four-judge mean, **fihrIA-ASR 1.7B achieves the lowest SER among all open-weight systems (6.09%)**, followed by **fihrIA-ASR 0.6B (7.11%)**, Voxtral Mini 4B (8.35%), and MoulSot 1.7B (11.15%), trailing only the closed teacher model Gemini 3.8 Flash (4.95%). Notably, our 0.6B model—despite using roughly one-seventh of Voxtral's parameters and one-third of MoulSot's—outperforms Voxtral by 1.24 pp and MoulSot by 4.04 pp. Multi-judge disaggregation (**Fig. 4a**) shows complete cross-provider consensus: **all**

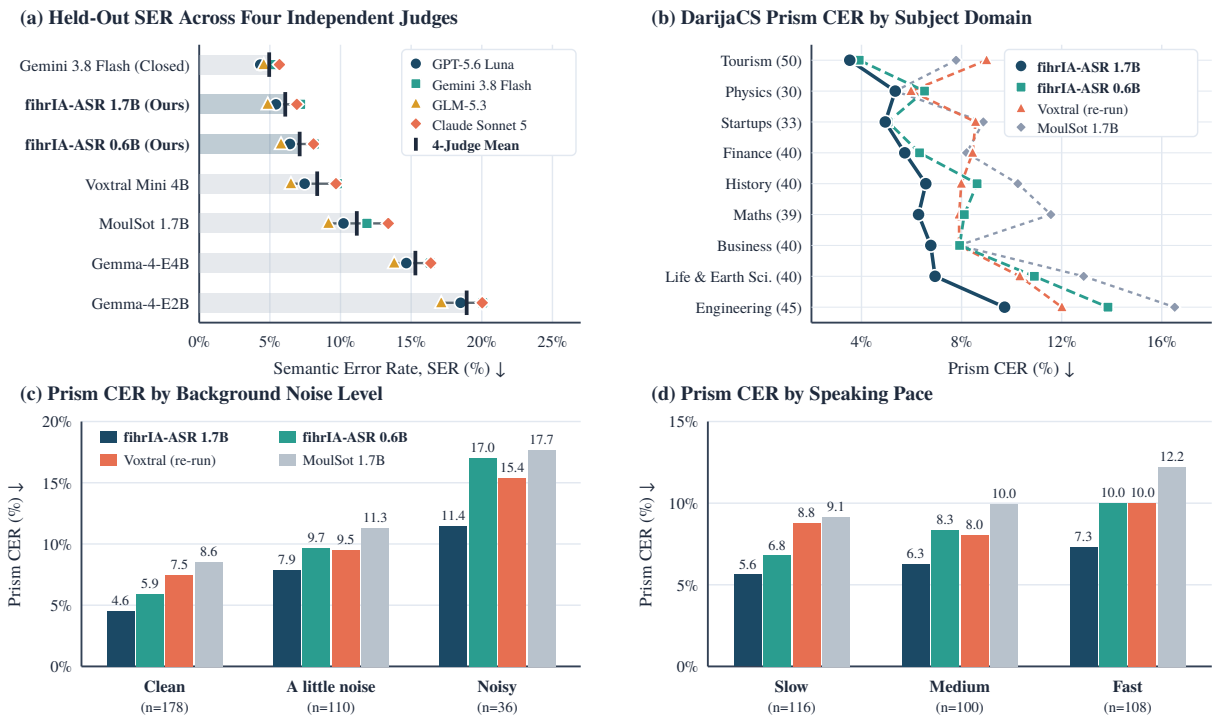


Figure 4: Multi-judge held-out profile and external DarijaCS subgroup robustness. a, Held-out SER (%) across four independent LLM judges (colored markers) and their unweighted mean (shaded bars and vertical ticks). All four judges unanimously rank fihrIA-ASR 1.7B first and fihrIA-ASR 0.6B second among open-weight models, ahead of Voxtral Mini 4B. **b**, DarijaCS Prism CER across all nine subject domains ($n = 357$), where fihrIA-ASR 1.7B achieves the lowest error in every domain. **c**, Prism CER stratified by background noise level ($n = 324$). **d**, Prism CER stratified by speaking pace ($n = 324$).

four judges unanimously rank fihrIA-ASR 1.7B first and fihrIA-ASR 0.6B second among open-weight models (5.43% GPT-5.6 Luna, 7.17% Gemini 3.8 Flash, 4.85% GLM-5.3, and 6.91% Claude Sonnet 5 for 1.7B; 6.42%, 8.14%, 5.79%, and 8.09% for 0.6B), reducing SER by 40%–48% relative to MoulSot at 1.7B and outperforming Voxtral Mini 4B at both scales under every evaluator. Alongside this unanimous accuracy lead at 1.7B, fihrIA-ASR 1.7B delivers a 7.1× **server throughput advantage** (243× versus 34× real time on a single H100 GPU; **Fig. 3a**). Comparing clip-median against corpus-micro WER (**Extended Data Table ED.4**) also demonstrates diagnostic complementarity: MoulSot's median WER (42.3%) masks a single-clip autoregressive loop repeating "gu" roughly 8,000 times that drives its micro WER to 91.1% (micro CER 40.2%), confirming why bounded SER and micro WER should be reported together.

To verify external generalization under an independent protocol, we evaluated open-weight systems on Morocco's public **DarijaCS** benchmark[22] across its **Prism** (ISMA; 357 clips, ≈ 90 min) and **Probe** (Darija in the Wild; 109 clips, ≈ 15 min) partitions using Voxtral's greedy decoding and Arabic text normalizer (**Fig. 3b** and **Extended Data Table ED.5**). Against Voxtral Mini 4B's published numbers, **fihrIA-ASR 1.7B achieves lower CER on both partitions** (6.24% versus 6.68% on Prism; 8.71% versus 9.69% on Probe), while our compact 0.6B model trails published Voxtral by 1.30 pp on Prism (7.98%) and nearly matches it on Probe (9.80%). Moreover, our local Hugging Face 16-bit replication of Voxtral suffered inflated surface CER (8.94% Prism, 12.11% Probe) from subtle wrapper and streaming-chunk differences, yet scoring those exact replicated transcripts with SER restored the published Prism ordering (7.95% Prism, 6.78% Probe)—

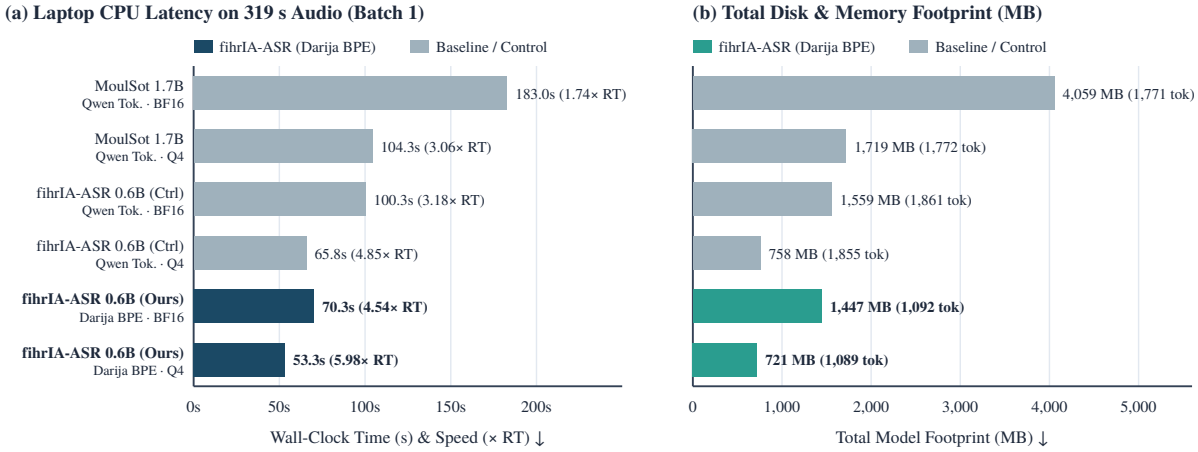


Figure 5: Tokenizer adaptation and 4-bit quantization enable $5.98\times$ real-time transcription in 721 MB on a laptop CPU. Six-step ablation in `llama.cpp` (Intel Core Ultra 7 258V, 8 threads, batch size 1, 319 s of audio). **a**, End-to-end wall-clock latency and real-time throughput across configurations; combining the 0.6B scale, Darija BPE tokenizer ($1{,}861 \rightarrow 1{,}092$ generated tokens), and 4-bit quantization reduces total runtime $3.4\times$ (183.0 s $\rightarrow$ 53.3 s) and lifts throughput from $1.74\times$ to $5.98\times$ real time. **b**, Total disk and memory footprint across configurations, dropping $5.6\times$ from $4{,}059$ MB to 721 MB.

an incidental demonstration of SER's invariance to preprocessing shifts. Across Prism subgroups (**Fig. 4b–d** and **Extended Data Tables ED.6–ED.8**), three structural patterns emerge: (1) fihrIA-ASR 1.7B achieves the lowest CER and WER in all nine subject domains, ranging from Tourism (3.53% CER) to Engineering (9.72% CER)—mirroring the scarcity of engineering video in Moroccan media (**Fig. 1a**); (2) acoustic background noise drives a $2.5\times$ increase in 1.7B CER from clean (4.56%) to noisy clips (11.45%), whereas fast speech adds only 1.71 pp over slow speech; and (3) quantizing the 1.7B and 0.6B decoders from 16-bit to 4-bit precision shifts CER by less than 0.6 pp across both partitions ($6.24\% \rightarrow 6.54\%$ for 1.7B and $7.98\% \rightarrow 7.98\%$ for 0.6B on Prism), confirming near-lossless compression.

Sub-gigabyte, real-time transcription on consumer CPUs

To evaluate offline and on-device deployability, we exported the models to GGUF format and profiled a six-step ablation in `llama.cpp` on an Intel Core Ultra 7 258V laptop CPU (8 threads, batch size 1, 24 clips totaling 319 s of audio; **Fig. 5** and **Extended Data Table ED.9**). Transitioning from 16-bit MoulSot 1.7B ($1.74\times$ real time, 4,059 MB) to 4-bit fihrIA-ASR 0.6B shrinks total disk and memory footprint 5.6-fold to **721 MB** and lifts throughput to $5.98\times$ **real time**.

Comparing the six configurations reveals three multiplicative efficiency gains on CPU: downsizing from 1.7B to 0.6B cuts 16-bit wall-clock latency nearly in half (183.0 s $\rightarrow$ 100.3 s; $3.18\times$ real time); replacing the base Qwen vocabulary with our Darija BPE tokenizer cuts generated tokens by 41.3% ($1{,}861 \rightarrow 1{,}092$) and reduces 16-bit 0.6B wall-clock time to 70.3 s ($4.54\times$ real time) without quantization; and 4-bit quantization further lowers total runtime to **53.3 s** ($5.98\times$ real time, $3.4\times$ faster end-to-end than 16-bit MoulSot).

Discussion

Our results demonstrate that advancing speech recognition for unstandardized, code-switched vernaculars requires co-designing representation and evaluation rather than scaling general-purpose multilingual models

alone. When a spoken language lacks prescriptive spelling and fluidly interleaves multiple scripts, training under a consistent bilingual convention with a dialect-matched vocabulary concentrates decoder probability mass and eliminates over 40% of autoregressive decoding steps. Equally important, decoupling semantic fidelity from surface orthography via span-level SER removes a pervasive measurement distortion that has long obscured genuine acoustic progress in dialectal ASR. Because SER is bounded on $[0, 100\%]$, it complements rather than replaces surface metrics: corpus-micro WER remains indispensable for exposing rare autoregressive repetition loops (such as MoulSot's 8,000-syllable loop), while surface CER and WER remain appropriate when downstream products enforce a strict house orthography. Reporting span-level SER alongside median and micro surface error rates offers a transferable standard for Arabic dialects, Indic code-switched speech, African vernaculars, and other unstandardized languages.

Several limitations bound our conclusions and define clear priorities for future work. First, because our 1.51-million-clip training pool was transcribed by Gemini, fihrIA-ASR inherits stylistic habits and occasional non-speech hallucinations from its teacher, which we mitigate via lightweight non-speech stopping guards across both 0.6B and 1.7B models (**Extended Data Table ED.8**). Furthermore, while our tokenizer ablation is causally controlled, the bilingual script convention itself was not ablated against an Arabic-only training control and remains a principled design choice rather than an isolated causal factor. Second, our human validation study achieved 79.1% grid completion (2,006 / 2,535 pairs) with a single native speaker per pair and evaluated an early 1.7B development checkpoint (v0.5) without Voxtral Mini 4B; collecting multi-rater annotations on a shared overlap subset is an important next step to establish an exact human–human agreement ceiling and audit non-response randomness. Third, while all four LLM judges unanimously rank fihrIA-ASR 1.7B first and fihrIA-ASR 0.6B second among open-weight models, absolute SER levels vary by 2.0–3.3 pp across judge families, underscoring the value of multi-family judge ensembles whenever comparing closely matched systems.

Methods

Web-scale corpus curation and video-level decontamination

We constructed a large-scale Moroccan Darija audio catalogue from public Moroccan YouTube channels across three iterative channel-discovery passes, recording channel-level topical metadata at acquisition. The raw catalogue comprised 213,089 videos across 460 channels, totaling 36,657 hours of audio spanning news broadcasts, regional media, talk shows, comedy, culinary programs, legal and financial commentary, education, and daily-life vlogs. Voice activity detection and segmentation partitioned raw audio into 11,060,528 speech segments (33,613 hours) stored as 16-kHz mono 16-bit PCM; an initial 30-second window ceiling was followed by a merging pass that joined adjacent short segments from the same video to preserve syntactic continuity. From this segmented pool, approximately 10,000 hours were transcribed using Gemini.

To guarantee zero overlap between training and evaluation, volunteers manually transcribed excerpts from a held-out selection of YouTube videos to form our 508-clip benchmark test set, and we enforced strict **video-level decontamination**: every clip originating from any source video represented in the evaluation pool—not merely the annotated timestamp range—was purged from the training catalogue. This removed 1,848 clips across 855 source videos, eliminating shared speakers, room acoustics, channel jingles, and recurring topical phrasing. Filtering out clips outside $[3.0\text{ s}, 30.0\text{ s}]$ and removing 7,441 transcripts exceeding 620 characters (the upper 0.5% tail, predominantly reflecting repetitive decoding loops) yielded the final supervised training corpus of **1,508,646 clips**.

Bilingual script convention

To prevent unconstrained cross-script variance during training, our Gemini transcription prompt enforced two rules: (1) Moroccan Darija and Arabic lexical items are written in Arabic script, and (2) spoken French and English words or phrases are written in Latin script using standard orthography (e.g., هادي هي "la discipline"). Secondary dialectal choices—such as clitic attachment (فداركم versus في داركم), numeral formatting, and short-vowel spelling—were left to natural contextual distribution.

Darija BPE tokenizer and two-stage fine-tuning

We trained an 89,000-token byte-level BPE tokenizer on our 1.51M Darija training transcripts supplemented with French and English text to preserve compact subword coverage during code-switching. Vocabulary surgery initialized the 33,762 tokens shared with Qwen3-ASR's base 151,936-token vocabulary[4] to their pretrained input and output embedding vectors, while the 55,238 newly introduced Darija tokens were initialized to the mean of all pretrained embeddings.

Because directly fine-tuning all decoder weights when 62% of the vocabulary shares a single mean-initialized vector can destabilize early gradients, we used a two-stage adaptation schedule across four NVIDIA H100 GPUs (learning rate 10^{-4}): in **Stage 1 (embedding warm-up)**, all transformer blocks and the audio encoder were frozen while only the input and output token embedding matrices were optimized for 2,000 text-only steps; in **Stage 2 (acoustic fine-tuning)**, the audio encoder remained frozen to preserve pretrained acoustic representations while the full language-model decoder was fine-tuned on all 1,508,646 audio–transcript pairs for up to 5 epochs (batch size 20 clips per GPU) with on-the-fly additive noise and music augmentation. Checkpoints minimizing validation loss were selected at step 56,000 (0.6B; 1,447 MB in 16-bit precision) and step 50,000 (1.7B; 3,830 MB in 16-bit precision). Controlled baseline models at both scales were trained identically using the base 151,936-token Qwen vocabulary.

Span-level SER formalization and LLM judge protocol

Both human annotators and automated LLM judges followed an identical two-category span taxonomy: *Acceptable Variations* (zero penalty: dialectal spelling variants, prepositional/conjunction clitic splitting, punctuation, vowel diacritics, cross-script loanword rendering, and equivalent numeral representations) and *Semantic Errors* (penalized: any deletion, insertion, or substitution of content or function words that alters, omits, or fabricates meaning). For automated scoring, each LLM judge—blinded to ASR system identity—was prompted with Darija-specific annotation rules requiring structured JSON output listing all aligned variation and error spans prior to computing $\mathrm{SER}(R, H) = 100 \times (n_{\mathrm{err},R} + n_{\mathrm{err},H})/(W_R + W_H) \in [0, 100]$.

Native-speaker validation study and 13-metric benchmark

Seven native Moroccan Darija speakers annotated reference–hypothesis pairs via a custom web interface displaying both strings side by side, highlighting and linking word spans corresponding to semantic errors while leaving acceptable variations unmarked. The cohort spanned 507 unique utterances across five ASR systems (Gemini, fihrIA-ASR v0.5, MoulSot, Gemma-4-E4B, and Gemma-4-E2B). Of $507 \times 5 = 2{,}535$ target pairs, 2,006 pairs (79.1%; ≈ 401 per system, each rated by one native speaker) received both completed human annotation and valid primary-judge output. System-level stability was assessed using 2,000 audio-cluster bootstrap resamples. On the same 2,006 pairs, we implemented and evaluated 12 competing metrics from the literature spanning LLM-aware judges (SN-WER[10], LLM-WER, OIWER-WER[11]), rule/dictionary normalizers (Norm-WER, S-WER proxy[13], WERd-proxy[9]), contextual embeddings (SeMaScore[15], BERTScore[12], AraDiaWER-proxy[14], SemDist), and surface edit distances (WER, CER).

Benchmark evaluation and hardware profiling

On the 508-clip held-out benchmark (≈ 2.05 hours), nine ASR systems were scored independently by four LLM judges via OpenRouter (GPT-5.6 Luna [`openai/gpt-5.6-luna`], Gemini 3.8 Flash [`google/gemini-3.8-flash`], GLM-5.3 [`z-ai/glm-5.3`], and Claude Sonnet 5 [`anthropic/claude-sonnet-5`]) alongside clip-median and corpus-micro WER/CER. On the external DarijaCS benchmark[22] (Prism $n = 357$, Probe $n = 109$), all models were decoded greedily at temperature 0 and scored with Voxtral's Arabic text normalizer plus our primary SER judge (GPT-5.6 Luna). For DarijaCS, our fihrIA-ASR evaluations included two lightweight inference-stability guards (a non-speech stopping threshold and a repetition-loop breaker); MoulSot was evaluated in 4-bit precision, while our local Voxtral replication used 16-bit weights in Hugging Face Transformers with greedy decoding and the default 480-ms streaming delay. Server GPU throughput was measured on a single NVIDIA H100 GPU at batch size 32 over the 508 held-out clips. Edge CPU latency and memory footprints were measured in `llama.cpp` (GGUF format) on an Intel Core Ultra 7 258V laptop CPU (8 threads, batch size 1) over 24 clips totaling 319 s of audio.

Data and code availability

The **fihrIA-ASR model weights** (0.6B and 1.7B) and the **508-clip held-out benchmark** (evaluation clips and expert human reference transcripts, subject to underlying public source media terms) are released with this work. The raw 36,657-hour video catalogue, machine-generated training transcripts, training-video metadata, individual annotator records, and internal analysis scripts remain private. The external DarijaCS benchmark[22] is governed by its original release terms.

References

1. Boumehdi, A. & Yousfi, A. MDVC corpus: Empowering Moroccan Darija speech recognition. *Indones. J. Electr. Eng. Comput. Sci.* **34**, 290–301 (2024). https://doi.org/10.11591/ijeecs.v34.i1.pp290-301.
2. Zarnoufi, R., Jaafar, H., Bachri, W. & Abik, M. MANorm: A normalization dictionary for Moroccan Arabic dialect written in Latin script. In *Proc. Workshop on Arabic Natural Language Processing (WANLP)* (2020). https://aclanthology.org/2020.wanlp-1.14/.
3. Sedrati, A., Afifi, M. & Benkhadra, R. Standardizing Darija: Collaborative approaches in the Moroccan Darija Wikipedia. *Linguistica Posnaniensis* **67** (2025). https://doi.org/10.14746/linpo.2025.67.1.4.
4. Qwen Team. Qwen3-ASR technical report. *Preprint at https://arxiv.org/abs/2601.21337* (2026).
5. Ali, A. *et al.* Casablanca: A multilingual, multidialectal Arabic speech corpus. In *Proc. EMNLP* (2024). https://aclanthology.org/2024.emnlp-main.1211/.
6. Ali, A., Khalifa, S. & Habash, N. Towards variability-resistant dialectal speech evaluation. In *Proc. Interspeech* (2019). https://www.isca-archive.org/interspeech_2019/ali19_interspeech.html.
7. Hamed, I. *et al.* Evaluating code-switched and dialectal Arabic ASR with human acceptability. In *Proc. IEEE Spoken Language Technology Workshop (SLT)* (2023). https://arxiv.org/abs/2211.16319.
8. Messaoudi, O. *et al.* Automatic speech recognition for Tunisian Arabic dialect with code-switching. In *Proc. IEEE ICASSP* (2024). https://arxiv.org/abs/2309.11327.
9. Ali, A., Nakov, P., Bell, P. & Renals, S. WERd: Using social text spelling variants for evaluating dialectal speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2017). https://arxiv.org/abs/1709.07484.
10. Pattnayak, P. SN-WER: Script-normalized WER for multi-script Indic ASR evaluation. In *Proc. MeLLM* (2026). https://aclanthology.org/2026.mellm-1.24.pdf.
11. Bhogale, K. S., Javed, T., John, G. S. *et al.* Towards orthographically-informed evaluation of speech recognition systems for Indian languages. In *Proc. IEEE ICASSP* (2026). https://arxiv.org/abs/2603.00941.

12. Zhang, T. *et al.* BERTScore: Evaluating text generation with BERT. In *Proc. International Conference on Learning Representations (ICLR)* (2020). https://arxiv.org/abs/1904.09675.
13. Roy, S. Semantic-WER: A unified metric for the evaluation of ASR transcripts for end usability. *Preprint at https://arxiv.org/abs/2106.02016* (2021).
14. Elmadany, A. *et al.* AraDiaWER: A dialect-aware metric for evaluating Arabic ASR systems. In *Proc. Field Matters Workshop* (2023). https://aclanthology.org/2023.fieldmatters-1.8/.
15. Sasindran, S. *et al.* SeMaScore: A new evaluation metric for automatic speech recognition tasks. In *Proc. Interspeech* (2024). https://arxiv.org/abs/2401.07506.
16. Bañeras-Roux, T. *et al.* Generative vs. encoder large language models for ASR evaluation: A comparative study. *Preprint at https://doi.org/10.48550/arXiv.2608.25574* (2026).
17. Hugging Face. The Open ASR Leaderboard adds its first Global South language. Hugging Face Blog (2026). https://huggingface.co/blog/open-asr-leaderboard-global-south.
18. Sarvam AI. Evaluating Indian-language ASR: Beyond WER to LLM and semantic metrics. Engineering blog post (2026). https://www.sarvam.ai/blogs/evaluating-indian-language-asr.
19. Li, J. *et al.* Interactive ASR: Towards human-like interaction and semantic coherence evaluation for agentic speech recognition. *Preprint at https://arxiv.org/abs/2604.09121* (2026).
20. Mistral AI. Voxtral Mini 4B Realtime Arabic. Hugging Face model card (2026). https://huggingface.co/mistralai/Voxtral-Mini-4B-Realtime-Arabic.
21. AtlasIA. MoulSot v0.3: Automatic speech recognition for Moroccan Darija. Hugging Face model card (2026). https://huggingface.co/atlasia/moulsot.v0.3.
22. Ministry of Artificial Intelligence, Digital Transformation and Administration Reform. DarijaCS: A public benchmark for Moroccan Darija speech recognition. Hugging Face dataset (2026). https://huggingface.co/datasets/Morocco-MITNRA-Labs/DarijaCS-asr-eval. Released 10 November 2026; Prism (ISMA) and Probe (Darija in the Wild) partitions.

Extended Data: Validation, Case Studies, and Full Benchmark Tables

Extended Data Section 1: Human Validation & Qualitative Zero-Error Case Studies

Extended Data Tables ED.1–ED.3 and Extended Data Figures ED.1–ED.3 report the complete native-speaker validation statistics, cross-judge consistency (GPT-5.6 Luna versus Gemini 3.8 Flash), pairwise system-gap calibration, zero-error orthographic failure distributions, and 13-metric benchmark across $n = 2{,}006$ matched pairs.

Table ED.1: Validation of automatic metrics against native-speaker SER across the $n = 2{,}006$ human-study cohort (507 utterances $\times$ 5 systems; lower error is better). System-level Human SER is macro-averaged across utterances, while automatic metrics are micro-pooled over the matched cohort. Bottom rows report pair-level agreement, calibration error on the $n = 497$ human-zero subset, and pairwise inter-system gap MAE across all $\binom{5}{2} = 10$ system pairs.

System / Summary Statistic	Pairs (n)	Human SER	LLM SER	WER	CER
Gemini	400	**3.29**	**4.06**	**31.62**	**8.50**
fihrIA-ASR v0.5 (1.7B dev.)	401	5.41	7.32	37.26	13.67
MoulSot 1.7B [21]	401	8.85	11.86	44.73	15.74
Gemma-4-E4B	401	12.75	14.22	46.91	14.66
Gemma-4-E2B	403	14.59	17.01	48.87	15.48
Utterance Spearman ρ / Pearson r ($\uparrow$)	2,006	—	**0.71 / 0.76**	0.52 / 0.54	0.51 / 0.48
Utterance MAE / RMSE vs. Human (pp $\downarrow$)	2,006	—	**4.8 / 6.8**	33.2 / 35.5	7.4 / 9.8
Mean score when Human SER $= 0\%$ ($n = 497$, $\downarrow$)	497	0.00	**4.22**	34.04	10.35
Pairwise system-gap MAE across 10 pairs (pp $\downarrow$)	10	—	**1.09**	3.51	3.68

Table ED.2: Cross-judge consistency comparing two independent LLM SER judges (GPT-5.6 Luna and Gemini 3.8 Flash) against native-speaker annotations across the $n = 2{,}006$ validation cohort.

Evaluation Quantity	GPT-5.6 Luna	Gemini 3.8 Flash
Utterance-level Spearman ρ vs. Human SER	0.711	0.725
Utterance-level Pearson r vs. Human SER	0.765	0.771
Utterance-level MAE vs. Human SER (pp)	**4.75**	5.45
Utterance-level RMSE vs. Human SER (pp)	**6.82**	7.68
Mean score elevation above Human SER (pp)	**+1.92**	+3.85
Mean score on Human-Zero subset ($n = 497$, %)	**4.22**	5.35
Exact 5-system rank match (2,000 bootstrap resamples)	100%	100%
Cross-judge utterance Spearman ρ / Pearson r	0.877 / 0.892	

Below we reproduce the complete Arabic and Latin transcripts for the three representative human-zero pairs (WER $> 75\%$) shown in Extended Data Fig. ED.2c–e:

Case 1: Prepositional clitic attachment and proper-noun transliteration (Gemma-4-E4B).

- *Reference:* وحنا محتاجين مرا متعلمة، ولكن التعلم ماشي معناه هو الاختلاط. اليوم راه يمكن ليك تقرا فداركم فستانفورد، وكامبريدج، وإم آي تي، وأوبن كورس وير
- *Hypothesis:* و حنا محتاجين مرا متعلمة ولكن التعلم ماشي المعنى هو الاختلاط اليوم راه يمكن ليك تقرا في داركم في ستانفورد و كامبريدج و ام آي تي و اوبن كورس

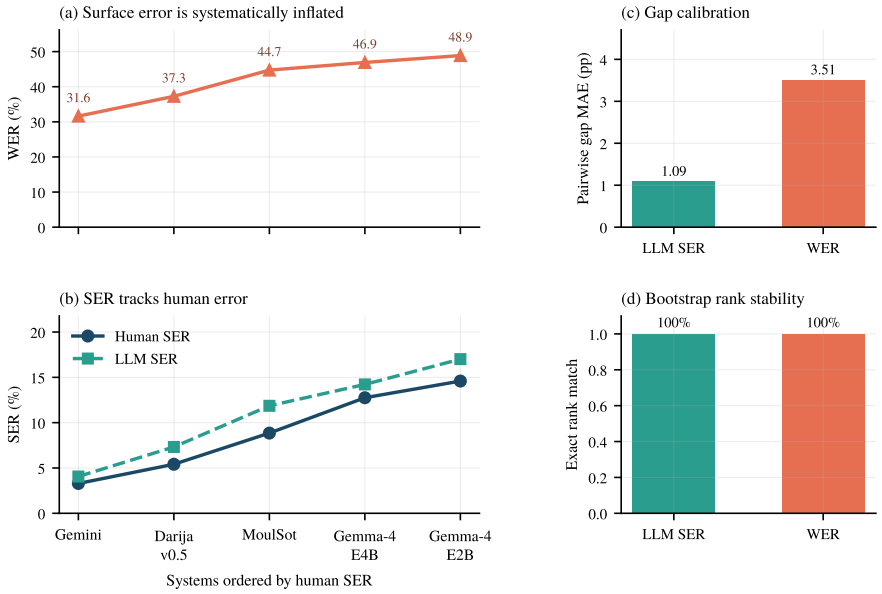


Figure ED.1: System ordering and pairwise gap calibration on the human-study cohort ($n = 2{,}006$). a, Model-level WER is uniformly inflated by > 30 pp above human SER. **b**, LLM SER closely tracks both the trajectory and absolute scale of native-speaker error rates. **c**, Mean absolute error across all 10 pairwise inter-system gaps (1.09 pp for LLM SER versus 3.51 pp for WER). **d**, Exact five-system ranking recovery probability across 2,000 audio-cluster bootstrap resamples.

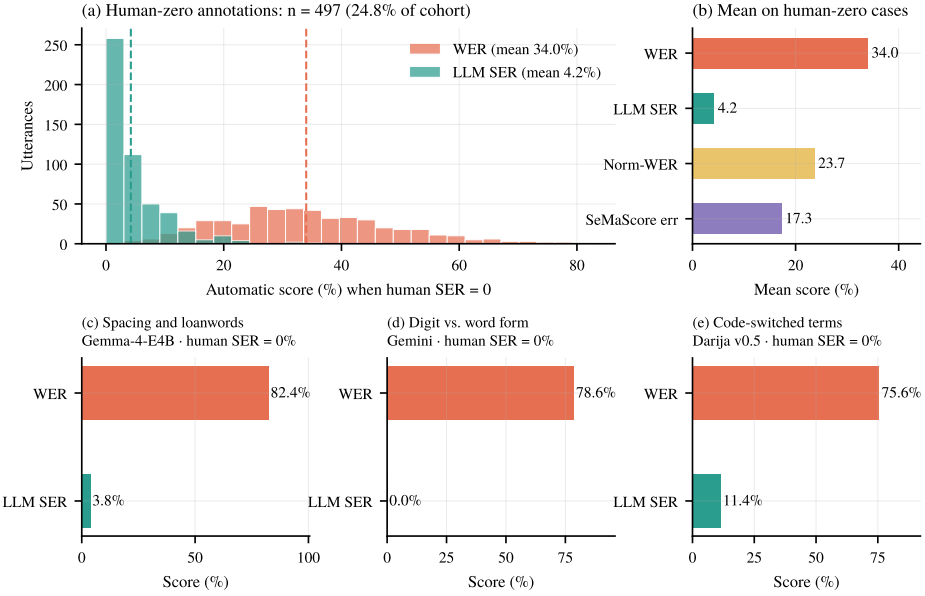


Figure ED.2: Orthographic variation without semantic error ($n = 497$ human-zero transcripts). a, Distribution of automatic WER and LLM SER conditioned on Human SER $= 0\%$; vertical dashed lines indicate subset means (34.0% WER versus 4.2% LLM SER). **b**, Subset mean scores comparing WER, LLM SER, Norm-WER, and SeMaScore error. **c–e**, Three representative utterances where native speakers marked zero semantic errors yet WER exceeded 75% due to clitic spacing and loanword script, digit-versus-word numerals, and French code-switching.

- *Scores and mechanism:* Human SER $= 0.0\%$, LLM SER $= 3.8\%$, WER $= 82.4\%$. Separating prefixed clitics (وحنا, فداركم, فستانفورد ↔ و حنا, في داركم, في ستانفورد) triggers insertions plus substitutions on 14 of 17 words under WER, whereas LLM SER accepts all clitic splits and hamza variants, penalizing only the minor omission of the final particle وير (“ware”).

Case 2: Digit representation versus multi-word spelled-out numerals (Gemini).

- *Reference:* وماشي غير الطوموبيلات الجديدة لي فيها ارتفاع كبير، حتى الطوموبيلات القديمة أرقام قياسية من ناحية التعاملات، كنشوفو 775000 عملية تحويل الملك
- *Hypothesis:* و ماشي غير طوموبيلات الجديدة اللي فيها ارتفاع كبير، حتى طوموبيلات القديمة أرقام قياسية من ناحية التعاملات، كنشوفو سبعمية و خمسة و سبعين ألف
- *Scores and mechanism:* Human SER $= 0.0\%$, LLM SER $= 0.0\%$, WER $= 78.6\%$. Expanding “775000” into six Arabic words (سبعمية و خمسة و سبعين ألف) alongside minor spelling variants inflates WER to 78.6%, whereas both native speakers and LLM SER assign 0.0% error.

Case 3: Cross-script French code-switching (fihrIA-ASR v0.5).

- *Reference:* يعني هاد الطراجي لي غنضرب دابا، كون مشيت فالكار غادي نذير ثمانية ولا تسع سوايع، دابا غنديرو فساعة ونص. لي لاحظت الإخوان أن اليابانيين فالمو ...
- *Hypothesis:* اللي غنضرب دابا كون مشيت ف الكار كان غادي نديرو تمانية ولا تسعة د السوايع ”trajet“ يعني هاد حنا دابا غنديروه ساعة ونص اللي لاحظتي أخويا أن الي ...
- *Scores and mechanism:* Human SER $= 0.0\%$, LLM SER $= 11.4\%$, WER $= 75.6\%$. LLM SER pardons “trajet” ↔ الطراجي, clitic splitting (ف الكار), and numeral spelling, while flagging the morphological person shift (نديرو vs. نذير, لاحظتي أخويا vs. لاحظت الإخوان) to yield 11.4% SER rather than 75.6% WER.

Table ED.3: Quantitative comparison of 13 automatic metrics against native-speaker SER across $n = 2{,}006$ pairs. Metrics labeled as proxies implement the cited methodology for Moroccan Darija.

Metric	Family	Spearman $\rho \uparrow$	Pearson $r \uparrow$	MAE (pp) $\downarrow$	RMSE (pp) $\downarrow$	Zero-Error Mean (%) $\downarrow$
LLM SER (GPT-5.6)	**LLM-aware**	**0.71**	**0.76**	**4.8**	**6.8**	**4.2**
SN-WER [10]	LLM-aware	0.60	0.63	23.7	26.5	22.8
S-WER proxy [13]	Normalization	0.55	0.57	34.3	36.4	34.2
LLM-WER	LLM-aware	0.54	0.57	25.2	27.8	26.2
Norm-WER	Normalization	0.53	0.56	22.8	25.9	23.7
WER	Exact-match	0.52	0.54	33.2	35.5	34.0
CER	Exact-match	0.51	0.48	7.4	9.8	10.4
SeMaScore err [15]	Embedding	0.51	0.55	14.5	17.3	17.3
BERTScore err [12]	Embedding	0.46	0.48	10.2	12.2	13.7
OIWER-WER [11]	LLM-aware	0.35	0.38	9.4	12.5	11.3
WERd-proxy [9]	Normalization	0.35	0.38	9.4	12.5	11.3
AraDiaWER-proxy [14]	Embedding	0.35	0.29	9.1	13.0	11.0
SemDist err	Embedding	0.21	0.16	13.4	20.4	14.4

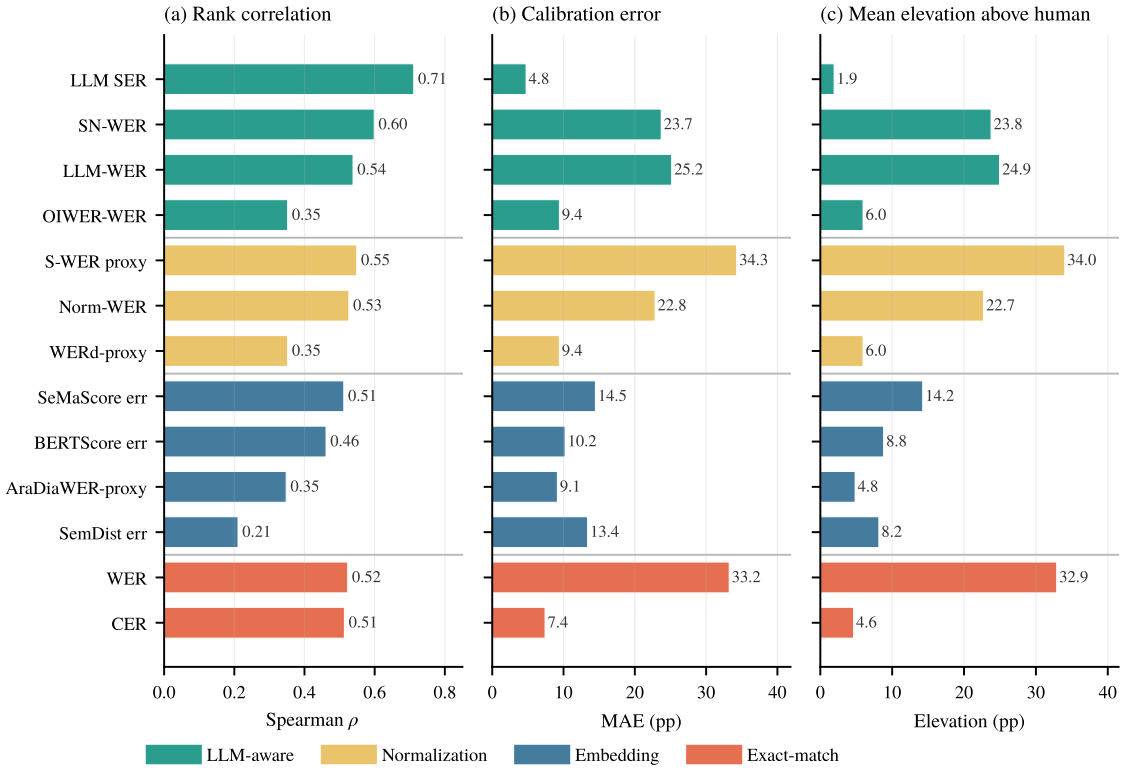


Figure ED.3: Benchmark of 13 automatic evaluation metrics against native-speaker SER ($n = 2{,}006$)**.** Metrics are grouped by family (LLM-aware, Normalization, Embedding, and Exact-match) and ordered by rank correlation within each family. **a**, Utterance-level Spearman rank correlation ρ (higher is better). **b**, Calibration mean absolute error in percentage points (lower is better). **c**, Mean score elevation above human SER in percentage points (closer to zero is better).

Extended Data Section 2: Held-Out, DarijaCS, and Hardware Profiling Tables

Table ED.4: Complete evaluation on the 508-clip held-out Darija benchmark, ordered by Mean SER (%; lower is better). Both clip-level median and corpus-pooled micro WER/CER are reported alongside individual SER scores for all four independent LLM judges. Bold denotes the best open-weight result in each column.

		Surface Error Rates (%)				Semantic Error Rate, SER (%) by Judge				
System	Access	Med. WER	Micro WER	Med. CER	Micro CER	GPT-5.6	Gemini	GLM-5.3	Claude 5	**Mean SER**
Gemini 3.8 Flash	Closed	26.9	28.4	6.8	8.5	4.32	5.24	4.57	5.68	4.95
fihrIA-ASR 1.7B	**Open**	**31.6**	**31.7**	10.6	12.6	**5.43**	**7.17**	**4.85**	**6.91**	**6.09**
fihrIA-ASR 0.6B	**Open**	32.3	32.6	**10.5**	12.6	6.42	8.14	5.79	8.09	7.11
Voxtral Mini 4B [20]	Open	36.5	35.8	10.6	**11.7**	7.46	9.77	6.49	9.68	8.35
MoulSot 1.7B [21]	Open	42.3	91.1	13.0	40.2	10.21	11.87	9.15	13.38	11.15
Gemma-4-E4B	Open	43.8	44.7	12.8	13.7	14.66	16.35	13.80	16.39	15.30
Gemma-4-E2B	Open	46.2	46.8	14.0	14.8	18.50	20.06	17.13	20.03	18.93
GPT-Audio-Mini	Closed	51.8	53.0	17.4	19.7	22.40	25.24	21.06	24.20	23.23
Inkling-Small	Closed	66.7	71.9	25.3	34.5	36.14	37.64	33.21	39.36	36.59

Table ED.5: Complete DarijaCS results [22] including 16-bit and 4-bit quantized fihrIA-ASR configurations across the Prism ($n = 357$) and Probe ($n = 109$) partitions (%; lower is better). CER and WER follow Voxtral's Arabic normalization protocol; SER is scored by GPT-5.6 Luna.

	Prism Partition (ISMA, $n = 357$)			Probe Partition (In the Wild, $n = 109$)		
System	CER (%) ↓	WER (%) ↓	SER (%) ↓	CER (%) ↓	WER (%) ↓	SER (%) ↓
fihrIA-ASR 1.7B (16-bit)	**6.24**	**13.02**	**5.84**	**8.71**	**17.21**	**5.22**
fihrIA-ASR 1.7B (4-bit)	6.54	13.67	6.19	9.23	18.23	5.77
Voxtral Mini 4B (published [20])	6.68	—	—	9.69	—	—
fihrIA-ASR 0.6B (16-bit)	7.98	16.38	8.81	9.80	19.36	6.53
fihrIA-ASR 0.6B (4-bit)	7.98	16.43	8.68	9.84	19.57	7.04
Voxtral Mini 4B (our 16-bit re-run)	8.94	17.31	7.95	12.11	21.73	6.78
MoulSot 1.7B [21]	10.27	22.22	12.26	14.93	26.30	9.40

Table ED.6: DarijaCS Prism results stratified by subject domain ($n = 357$). Each cell reports corpus-pooled CER / WER (%; lower is better). Bold indicates the best CER and WER in each row.

Subject Domain	Clips (n)	**fihrIA-ASR 1.7B**	**fihrIA-ASR 0.6B**	Voxtral 4B (re-run)	MoulSot 1.7B
Tourism	50	**3.53 / 9.19**	3.91 / 10.27	9.00 / 16.43	7.78 / 17.80
Physics	30	**5.35 / 8.72**	6.52 / 9.79	5.98 / 10.64	**5.35** / 11.49
Startups	33	**4.95 / 10.56**	5.06 / 11.44	8.56 / 16.20	8.87 / 19.54
Finance	40	**5.73 / 11.75**	6.32 / 13.80	8.44 / 15.78	8.20 / 17.34
History	40	**6.57 / 14.44**	8.62 / 18.46	7.99 / 18.37	10.25 / 22.49
Maths	39	**6.28 / 13.82**	8.11 / 16.36	7.90 / 15.36	11.57 / 23.45
Business	40	**6.77 / 11.81**	7.92 / 14.16	7.90 / 15.42	7.88 / 17.40
Life and earth sciences	40	**6.94 / 14.82**	10.91 / 21.49	10.33 / 19.84	12.88 / 28.08
Engineering	45	**9.72 / 20.37**	13.85 / 29.05	12.01 / 23.79	16.52 / 36.73

Table ED.7: DarijaCS Prism results stratified by background noise level and speaking pace ($n = 324$; the 33 startup clips carry distinct metadata tags and are omitted). Each cell reports CER / WER (%; lower is better).

Dimension	Subgroup	Clips (n)	**fihrIA-ASR 1.7B**	**fihrIA-ASR 0.6B**	Voxtral 4B (re-run)	MoulSot 1.7B
Background Noise	Clean	178	**4.56 / 10.90**	5.94 / 13.58	7.45 / 14.88	8.55 / 19.32
	A little noise	110	**7.88 / 14.94**	9.65 / 18.67	9.50 / 19.08	11.31 / 24.03
	Noisy	36	**11.45 / 21.37**	17.03 / 30.48	15.42 / 25.92	17.68 / 35.36
Speaking Pace	Slow	116	**5.63 / 12.36**	6.80 / 14.89	8.79 / 17.00	9.14 / 20.47
	Medium	100	**6.26 / 12.71**	8.34 / 16.93	8.03 / 16.16	9.97 / 21.97
	Fast	108	**7.34 / 14.97**	10.00 / 19.44	10.01 / 19.05	12.21 / 25.29

Table ED.8: DarijaCS Probe results stratified by clip difficulty label (reporting all label groups with $n \geq 5$ clips). CS denotes code-switching. Each cell reports CER / WER (%; lower is better).

Probe Difficulty Category	Clips (n)	**fihrIA-ASR 1.7B**	**fihrIA-ASR 0.6B**	Voxtral 4B (re-run)	MoulSot 1.7B
No difficulty label	23	**5.31 / 10.62**	5.87 / 12.93	7.75 / 15.24	8.23 / 16.63
Ordinary CS	38	**7.61 / 15.91**	8.60 / 18.27	13.82 / 22.99	14.81 / 25.98
Tricky CS	11	**4.05** / 14.37	4.51 / **13.77**	**4.05 / 13.77**	9.02 / 28.14
Technical CS	9	**13.97 / 18.75**	15.25 / 22.50	19.56 / 27.50	26.78 / 35.00
Ordinary + Technical CS	5	**6.21 / 13.29**	7.66 / 13.99	8.45 / 16.08	16.91 / 18.18
Silence	5	**20.46 / 27.78**	21.33 / 29.17	24.78 / 29.17	27.67 / 36.11

Table ED.9: Complete server GPU throughput (1× NVIDIA H100, batch size 32, 508 clips totaling 2.05 hours) and laptop CPU edge deployment profiling (`llama.cpp`, Intel Core Ultra 7 258V, 8 threads, batch size 1, 24 clips totaling 319 s).

		Server GPU (1× H100, Batch 32)			Laptop CPU (llama.cpp, Batch 1)		
System Configuration	Tokenizer	Wall (s)	Clips/s	× RT ↑	Wall (s)	Size (MB) ↓	× RT ↑
Voxtral Mini 4B (16-bit)	Voxtral	217.3	2.3	34×	—	—	—
MoulSot 1.7B (16-bit)	Qwen Base (152k)	—	—	—	183.0	4,059	1.74×
MoulSot 1.7B (4-bit)	Qwen Base (152k)	—	—	—	104.3	1,719	3.06×
fihrIA-ASR 1.7B (Control, 16-bit)	Qwen Base (152k)	50.1	10.1	147×	—	4,059	—
fihrIA-ASR 1.7B (Ours, 16-bit)	**Darija BPE (89k)**	**30.3**	**16.8**	**243**×	—	3,830	—
fihrIA-ASR 0.6B (Control, 16-bit)	Qwen Base (152k)	48.1	10.6	153×	100.3	1,559	3.18×
fihrIA-ASR 0.6B (Control, 4-bit)	Qwen Base (152k)	—	—	—	65.8	758	4.85×
fihrIA-ASR 0.6B (Ours, 16-bit)	**Darija BPE (89k)**	**28.9**	**17.6**	**255**×	70.3	1,447	4.54×
fihrIA-ASR 0.6B (Ours, 4-bit)	**Darija BPE (89k)**	—	—	—	**53.3**	**721**	**5.98**×