Probing Geographic Performance Gaps in Moroccan ASR with ALAMA: Annotated Local Audio of Moroccan Arabic

Avery C. Kanel[1,2] **Christian Schuler**[1,2]
Bouazza Laracha[3,*] Imrane Lbouhli[4,*] Yassine Chaouri[5,*]
Yusser Al Ghussin[1,2] **Timo Baumann**[6]

[1]Saarland University, [2]German Research Center for Artificial Intelligence (DFKI),
[3]Ibn Zohr University, [4]Mohammed V University, [5]Hassan II University of Casablanca
[6]OTH Regensburg

Abstract

Moroccan Arabic (Darija) is routinely treated as a single dialect in NLP research, yet it exhibits substantial regional variation across its geography. This pilot study questions the monolithic framing by collecting a regionally annotated speech corpus, 60 minutes of broadcast audio from Casablanca, Tangier, and Oujda, transcribed and verified by native speakers, and evaluating 17 state-of-the-art ASR systems against it. Our results show architecture-dependent performance gaps across regions: systems exhibit significant gaps in WER points between the best- and worst-served regions, and fine-tuned "Moroccan Arabic" checkpoints exhibit inconsistent regional profiles that are hard to reconcile with balanced training data. These findings suggest that national-level dialect labels obscure regionally uneven model coverage with real performance consequences. We release the dataset[1] and support the adoption of more fine-grained provenance labeling as standard metadata for Arabic dialect resources.

1 Introduction

Moroccan Arabic (Darija) is spoken by nearly all of Morocco's 37 million inhabitants as a first or second language (Heath, 2020). Many dialectical datasets are labeled simply as 'Moroccan Arabic', foregoing further geographic specification, glossing over substantial regional variation, phonological, lexical, and contact-driven, that characterizes the language (Heath, 2013; Aguadé, 2008). Darija also exists within a state of diglossia with Standard Arabic and features heavy code-switching into French and Tamazight languages as well as occasional use of English and Spanish, with patterns that differ systematically by region and social context (Skiredj et al., 2024; Hermas, 2025).

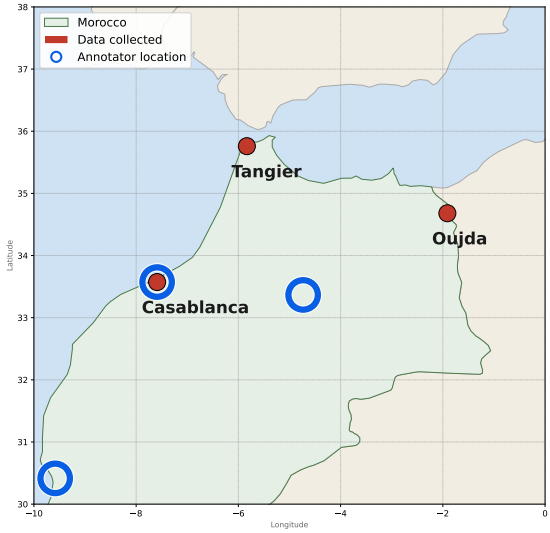


Figure 1: Data sources and human annotator locations.

Recent years have seen substantial progress in Darija ASR, with reported improvements of fine-tuned systems (Bounhar and Majjodi, 2025; Atlasia, 2026; Boumehdi and Yousfi, 2024). It is, however, an open question whether these advances equally benefit speakers across all of Morocco's regions. ASR models inherit the distributional features of their training data (Koenecke et al., 2020); if Moroccan Arabic corpora are, in fact, regionally concentrated as they are in other North African countries (Mohamed Eida and Habash, 2025), without disclosing this at the national-label level, speakers from less-represented areas face systematic performance penalties with no way to diagnose this from published benchmarks or model cards (Keleg and Magdy, 2023; Zaidani et al., 2024).

In this pilot study we present ALAMA, a regionally stratified Darija dataset spanning Casablanca, Tangier, and Oujda (Figure 1), with transcriptions verified by native speakers and metadata covering accent, register, and loan-word use, among others.

[1] https://github.com/Ackanel/alama_audio_dataset

*: Equal contribution (Human annotators).

Table 1: Salient samples to highlight phonological and morphological differences across the Northern, Central and Eastern regions of Morocco, illustrated through closely paired locations Tangier and Larache, Casablanca and Temara, and Oujda and Berkane, respectively (Boumans, 2005; Heath, 2013; Falchetta and Guerrero, 2023).

Linguistic Feature	**"Northern"**	**"Central"**	**"Eastern"**
Historical Dialect Group	Pre-Hilalian	Hilalian	Hilalian
Genitive Exponent	*d-* / *dyal*	*dyal*	*dyal* / *ntaʕ*
Affrication of /t/	Alveo-dental affricate [t͡s]	Alveo-palatal affricate [t͡ʃ]	Dental plosive [t̪]
Reflex of etymological *q	[q]	[g] and [q]	[g] and [q]

The remainder of this work presents our findings by evaluating state-of-the-art ASR-systems against ALAMA.

2 Related Work

Arabic NLP has long focused on distinguishing broad dialect groups from Modern Standard Arabic (MSA), while fine-grained intra-dialectal variation remains underexplored. The ADI17 dataset (Shon et al., 2020) advanced dialect identification to 17 country-level Arabic varieties, and work on micro-dialect identification (Abdul-Mageed et al., 2020) further narrowed the scope. The ALDi framework (Keleg et al., 2023, 2024) shows that utterances within a single dialect label vary along a dialectness continuum correlated with inter-annotator disagreement, a finding directly relevant to our annotation quality. Despite these advances, regional variation within country, including Moroccan Arabic, remains largely unaddressed.

Broadcast speech has anchored Arabic ASR research since MGB-2 (Ali et al., 2019a), extended by the QASR corpus (Mubarak et al., 2021) and MGB-5 (Ali et al., 2019b), which provided an early Moroccan regional reference point. Talafha et al. (2023) studied zero-, few-, and full-shot Whisper adaptation across Arabic dialects as a methodological predecessor to our evaluation. The most direct prior work is the Casablanca dataset (Talafha et al., 2024), which grounded the NADI 2025 shared task (Talafha et al., 2025) and its top system (Salhab et al., 2025); contemporaneous benchmarks Sullivan et al. (2026); Abdullah et al. (2025) similarly evaluate multi-system Arabic ASR but at national granularity. Our work complements these by making regional provenance within Moroccan Arabic a primary experimental variable. Hutiri and Ding (2022) operationalize fairness through worst-group and max-minus-min gap metrics, a framework that we apply at the regional level. Bootstrap methods for Word-Error Rate (WER) confidence intervals (Bisani and Ney, 2004; Liu et al., 2021) are discussed further in our evaluation strategy.

Moroccan Arabic exhibits substantial inter-regional variation as summarized in Table 1, shaped by urbanization, pre- and post-Hilalian mixing, and Berber contact influence (Heath, 2020; Ramos et al., 2024; Francisco, 2021; D'Anna, 2024), documented in linguistic studies of Casablanca (Aguadé, 2003; Hachimi, 2018, 2012), Tangier (Montes, 2015; Grigore and Bițună, 2016), Oujda (Fernández, 2017), and cross-regionally (Boumans, 2005). The ARCADE dataset (Nacar et al., 2026), which contributes to our corpus, illustrates how regional asymmetries arise even in carefully constructed resources, containing forty audio samples from Casablanca, fifteen from Tangier, and none from the Oriental region where Oujda is located. This motivates the balanced sampling pursued here.

3 Data

ALAMA consists of 60 minutes of regional broadcast speech drawn evenly from Casablanca, Tangier, and Oujda with approximately 20 minutes each. Audio was drawn from two sources: first, our manually gathered recordings from local radio channels, specifically call-in programs, in-studio interviews, and occasional sports broadcasts and radio dramas. Second, the ARCADE dataset (Nacar et al., 2026), which provided 9 minutes for Casablanca and 3 minutes for Tangier. To limit speaker-level bias, no single speaker contributes more than 3.4% of the to-

Table 2: Per-region word counts of human-curated transcriptions. **Radio region**: location of the broadcast; **Gold strict**: 3/3 agree; **Gold majority**: 2/3 agree.

Region	**Radio region**	**Gold strict**	**Gold majority**
Casablanca	2,402	1,912	4,318
Tangier	2,498	605	719
Oujda	2,483	0	1,208
Total	7,383	2,517	6,245

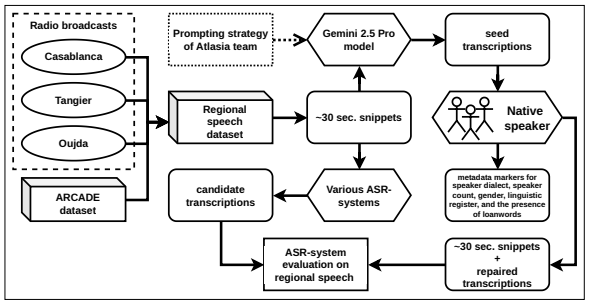


Figure 2: ASR experiment pipeline.

tal dataset. All segments were trimmed to between 27 and 30 seconds for Whisper-compatibility and no audio enhancement or denoising was applied, preserving raw broadcast conditions. Table 2 shows how human-annotated detected speaker accents differ from recording source labels.

Transcriptions were first produced by Gemini 2.5 Pro (Google, 2025) following the same prompting strategy used by Atlasia (2026), an open-source research group based in Morocco that developed one of the models examined in this paper, MoulSot. During this process we observed that identical interviews were occasionally broadcast across multiple regional channels, meaning the broadcast city did not always reflect the speaker's geographic origin. The geographic metadata was designed specifically to track these discrepancies. All transcriptions were subsequently verified and corrected by three native speakers of Moroccan Arabic (Table 3), producing the gold-label references used throughout our evaluation.

4 Methodology

Figure 2 provides an overview of the experimental pipeline and Table 5 in Appendix A shows information about the included ASR systems. We evaluate each of the 17 ASR systems (see Table 4 for details) against all three native-speaker reference transcriptions and report both Word-Error Rate (WER) and Character-Error Rate (CER) in two variants. **Raw** scores strip only chat-template artifacts present in generative model outputs; **Normalized** scores additionally apply the Open Universal Arabic ASR Leaderboard normalization (punctuation removal, diacritic stripping, hamza and madda normalization, Eastern-to-Western numeral conversion). We also report **Oracle WER**, the per-utterance minimum WER across the references, which provides a lower bound exposing the contribution of inter-annotator disagreement to observed error rates.

Scoring was performed using NIST SCTK `sclite` for word-level error decomposition (substitutions, deletions, insertions) and a `jiwer`-based port of the Leaderboard normalization pipeline, both run on all reference-hypothesis pairs. Per-region WER is computed separately for Casablanca, Tangier, and Oujda. To operationalize regional fairness we follow Hutiri and Ding (2022) and report the max-minus-min WER gap across regions as a disparity proxy.

We flag the `Gemini 2.5` system for gold-label circularity throughout: reference transcriptions were seeded by Gemini 2.5 Pro before native-speaker correction, and this system also relies on the Gemini model family for inference.

Table 3: Native-speaker annotator demographics.

Region	Sex	Age	Education
Souss-Massa	male	25	bachelor's student
Beni Méllal-Khenifra	male	25	master's student
Casablanca-Settat	male	20	bachelor's degree

5 Results and Analysis

Table 4 reports overall WER and CER and per-region breakdown for all 17 systems. Normalized WER spans a wide range: MoulSot v0.3 achieves the best non-circularity-flagged result at 35.5 %, while other systems exceed 90 %.

The per-region picture is the more suggestive one. 12 of 17 systems perform worst on the Oujda dialect which is a significant deviation from chance (χ^2 goodness-of-fit, df=2, p=.023), with gaps ranging from 0.87 to 17.8 WER points relative to the best-served region. Crucially, the direction of this disadvantage is not uniform: several systems fine-tuned on data labeled only as "Moroccan Arabic" perform *better* on Oujda than on Casablanca or Tangier (e.g., WSmall-D and WLarge-M fine-tunes), an inconsistency hard to reconcile with the assumption of homogeneous regional coverage in their training data.

The gold-circularity diagnostic makes this argument sharper. Gemini 2.5 Pro, which seeded our reference transcriptions, scores near-zero WER on Casablanca and Tangier (2.9 % and 4.4 %) but 23.1 % on Oujda (Table 4). Error decomposition (last column group in Table 4) shows this Oujda gap is driven almost entirely by insertions (7.0 %) rather than substitutions or deletions, indicating that the native-speaker post-editing effort required to correct the Gemini-seeded transcriptions differed

Table 4: ASR evaluation results on the 60-min ALAMA dataset (all values in %). **Normalized (LB)** shows leaderboard-normalized WER/CER and Oracle WER (per-utterance minimum across three references); **Per-Region WER** shows normalized WER per city, with **Gap** as the max-minus-min regional disparity; **sclite Decomposition** shows word-level Substitution/Deletion/Insertion breakdown. Green/yellow/red: column best/second-best/worst. Orange-tinted row is gold-circularity flagged because the reference transcripts were seeded by this system before native-speaker correction, making its WER an artifact rather than a comparable baseline.

	Normalized (LB)			Per-Region WER				sclite Decomposition		
Model	**WER**	**CER**	**Oracle**	**Casa.**	**Tang.**	**Oujda**	**Gap**	**Sub**	**Del**	**Ins**
DEDICATED ASR *(single-purpose end-to-end ASR)*										
WLarge-DL	45.5	24.9	41.0	43.2	41.2	51.6	10.5	32.9	12.8	3.5
WLarge	51.8	26.9	47.6	52.1	47.1	55.9	8.8	36.8	16.0	3.1
WLarge-M	56.9	21.8	53.2	57.3	57.2	55.6	1.7	42.9	8.1	9.3
WLargeCER-M	57.4	21.9	53.7	58.0	57.1	56.6	1.4	43.6	8.1	8.9
WMedium-D	53.4	26.9	49.9	45.5	52.6	61.6	16.1	30.2	8.9	18.3
WMedium	62.5	33.4	59.0	60.7	60.4	65.9	5.5	42.5	19.6	2.8
WSmall	72.9	36.3	70.0	71.0	72.3	74.9	3.9	52.6	18.7	3.4
WSmall-D	81.3	59.2	77.9	81.4	87.3	74.8	12.5	40.0	29.0	13.7
WTurbo	55.4	29.3	51.3	53.1	50.7	62.0	11.3	36.5	19.3	2.9
WTiny	91.4	56.3	90.3	91.6	90.5	92.1	1.6	58.2	31.6	2.0
W2V2-MD	55.6	21.4	51.7	58.4	56.5	51.5	6.9	45.2	10.1	3.2
W2V2-M	69.1	25.9	67.3	68.9	69.5	68.6	0.9	52.8	8.1	10.4
MoulSot	35.5	16.0	31.4	33.4	35.2	37.4	4.0	27.2	10.3	2.4
Qwen3ASR	54.2	23.1	49.6	51.6	51.4	59.0	7.6	43.1	8.9	4.9
MULTIMODAL LLM / MULTITASK SPEECH *(transcription as one of many tasks)*										
Gemini	9.7	5.7	0.2	2.9	4.4	23.1	20.2	2.7	0.1	7.0
Gemma4	42.1	20.1	38.1	38.8	38.6	48.4	9.8	32.2	11.2	2.2
Omni-7	50.1	30.1	45.8	47.1	47.3	55.3	8.2	28.2	23.6	1.2
Seamless	84.8	58.7	81.5	73.6	91.4	89.2	17.8	42.9	36.0	8.9

systematically by region. Taken together, these results suggest that regional origin is a measurable, architecture-dependent factor in ASR performance, and that training-data provenance shapes regional generalization in ways not visible from model labels alone.

6 Discussion

Our results provide evidence for ASR performance differences between regions of Morocco. A potential cause of which could be treating Moroccan Arabic as a single homogenous target in dataset construction, obscuring regional variation. The inconsistent, sometimes reversed regional skew across "Moroccan Arabic"-labeled checkpoints is hard to explain under the assumption of regionally balanced training data; a more parsimonious explanation is that these corpora are regionally concentrated, introducing undisclosed biases invisible at the national level.

This concern extends beyond the specific systems we evaluate. Labeling speech data as "Moroccan Arabic" while training predominantly on one region creates models whose blind spots are invisible to downstream users, and we argue that region-level provenance should become standard metadata for datasets and model cards not only for Arabic dialects.

Several caveats apply. Our three-region, broadcast-based setup cannot fully disentangle dialectal variation from acoustic recording conditions; isolating these effects would require larger per-region samples and more controlled environments. Because we lack provenance information for the fine-tuning corpora we evaluate, our account of their regional skew remains an inference rather than a verified claim. The gold-circularity in our annotation pipeline means WER figures are measured against a non-fully-independent standard, though the Gemini 2.5 Pro diagnostic corroborates region-dependent transcription difficulty. Additionally our gold-strict labels count for Oujda is n = 0. This could be due to annotator's unfamiliarity with Oujda shibboleths and subsuming the Oujda accent into the Casablanca category as none of the annotators come from Oujda. It is also worth considering that the general pattern of ASR models performing worse on Oujda data is a trickle-down effect of Gemini's worse performance as it was the basis for the annotator's transcriptions. This pilot motivates, rather than conclusively demonstrates, the need for regionally-stratified Darija ASR research.

7 Conclusion

ALAMA provides evidance that regional origin is a potential, architecture-dependent factor in Darija ASR, one that is invisible from national-level labels alone. The inconsistent profiles of "Moroccan Arabic"-labeled fine-tunes further suggest their training corpora are regionally skewed – a hypothesis that this pilot motivates but cannot conclusively verify at scale. We suggest the adoption of region-level provenance as metadata for Arabic dialect datasets and model cards when feasible, and to scale regionally-stratified Darija evaluation beyond this pilot.

For the future, we hope for dialect awareness *within* models rather than only in their description.

Acknowledgments

Christian Schuler gratefully acknowledges financial support from GraduSaar, University of Saarland.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020. Toward Micro-Dialect Identification in Diaglossic and Code-Switched Environments.

Badr M. Abdullah, Yusser Al Ghussin, Zena Al-Khalili, Ömer Tarik Özyilmaz, Matias Valdenegro-Toro, Simon Ostermann, and Dietrich Klakow. 2025. Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing under Data Constraints. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 745–751, Suzhou, China. Association for Computational Linguistics.

Jordi Aguadé. 2003. Notes on the Arabic dialect of Casablanca (Morocco). In *Proceedings of the Fifth Conference of AIDA (Association Internationale de Dialectologie Arabe). Ed. by I. Ferrando & JJ. Sánchez Sandoval, Cádiz, Servicio de Publicaciones de la Universidad de Cádiz*, pages 301–308.

Jordi Aguadé. 2008. Morocco [dialectological survey]. In Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich, and Andrzej Zaborski, editors, *Encyclopedia of Arabic Language and Linguistics*, volume 3, pages 287–297. Brill, Leiden–Boston.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019a. The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019b. The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033.

Atlasia. 2026. MoulSot: A curated moroccan darija speech dataset and fine-tuned ASR model.

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409.

Louis Boumans. 2005. Lexical variation in Moroccan Arabic: The genitive exponent in experimental data from four cities. *EDNA, Estudios de dialectología norteafricana y andalusí*, 9:125–137.

Ahmed Boumehdi and Abdellah Yousfi. 2024. MDVC corpus: Empowering Moroccan Darija speech recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(1):290–301.

Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings. *Hugging Face Blog*.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation.

Luca D'Anna. 2024. A dialect type of Eastern first-layer Maghrebi Arabic. *Kervan. International Journal of African and Asian Studies*, 28(2).

Jacopo Falchetta and Jairo Guerrero. 2023. Koineization trends in spoken arabic across three moroccan towns. *Kervan – International Journal of Afro-Asiatic Studies*, 27(1).

Montserrat Benítez Fernández. 2017. Variation linguistique dans le parler arabe d'Oujda: Nord-est marocain vs. communauté marocaine dans la diaspora. *Dialectologia*, pages 1–18.

Felipe Benjamin Francisco. 2021. The Southern Moroccan Dialects and the Hilāli Category. *Languages*, 6(4):192.

Google. 2025. Gemini 2.5 Pro.

George Grigore and Gabriel Bițună, editors. 2016. *Arabic Varieties: Far and Wide: Proceedings of the 11th International Conference of AIDA, Bucharest, 2015*. editura universității din București, București.

Atiqa Hachimi. 2012. The urban and the urbane: Identities, language ideologies, and arabic dialects in morocco. *Language in Society*, 41(3):321–341.

Atiqa Hachimi. 2018. Arabic Dialect Contact and Change in Casablanca: The Role of Simplification and Salience in the Adoption of a Morphosyntactic Variable. *Anthropological Linguistics*, 60(1):60–93.

Jeffrey Heath. 2013. *Jewish and Muslim Dialects of Moroccan Arabic*. Routledge, London.

Jeffrey Heath. 2020. Moroccan arabic. *Arabic and contact-induced change*, 1:213.

Abdelkader Hermas. 2025. Managing Morocco's linguistic market: A linguistic citizenship proposal. *European Journal of Language Policy*, 17(2):237–254.

Wiebke Toussaint Hutiri and Aaron Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 230–247.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic Level of Dialectness of Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the Level of Dialectness Predicts Interannotator Agreement in Multi-dialect Arabic Datasets.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions.

Mai Mohamed Eida and Nizar Habash. 2025. Beyond Cairo: Sa'idi Egyptian Arabic Literary Corpus Construction and Analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 292–304, Albuquerque, USA. Association for Computational Linguistics.

Valerie Susana Montes. 2015. *Urban Mediterranean Dialects of Arabic : Tangier and Tunis*. Master of Arts, The University of Texas at Austin.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera Speech Resource A Large Scale Annotated Arabic Speech Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Omer Nacar, Serry Sibaee, Adel Ammar, Yasser Alhabashi, Nadia Samer Sibai, Yara Farouk Ahmed, Ahmed Saud Alqusaiyer, Sulieman Mahmoud AlMahmoud, Abdulrhman Mamdoh Mukhaniq, Lubaba Raed, Sulaiman Mohammed Alatwah, Waad Nasser Alqahtani, Yousif Abdulmajeed Alnasser, Mohamed Aziz Khadraoui, and Wadii Boulila. 2026. ARCADE: A City-Scale Corpus for Fine-Grained Arabic Dialect Tagging.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Carmen Berlinches Ramos, Jairo Guerrero, and Montserrat Benítez Fernández. 2024. *AIDA Granada: A Pomegranate of Arabic Varieties*, volume 21. Prensas de la Universidad de Zaragoza.

Mahmoud Salhab, Shameed Sait, Mohammad Abusheikh, and Hasan Abusheikh. 2025. Munsit at NADI 2025 Shared Task 2: Pushing the Boundaries of Multidialectal Arabic ASR with Weakly Supervised Pretraining and Continual Supervised Fine-tuning.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-ASR Technical Report.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. ADI17: A Fine-Grained Arabic Dialect Identification Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248.

Abderrahman Skiredj, Ferdaous Azhari, Ismail Berrada, and Saad Ezzini. 2024. DarijaBanking: A New Resource for Overcoming Language Barriers in Banking Intent Detection for Moroccan Arabic Speakers. ArXiv:2405.16482 [cs.CL].

Peter Sullivan, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2026. Arab Voices: Mapping Standard and Dialectal Arabic Speech Technology.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, Sara Shatnawi, Yasir Ech-chammakhy, Amal Makouar, Yousra Berrachedi, Mustafa Jarrar, Shady Shehata, Ismail Berrada, and Muhammad Abdul-Mageed. 2024. Casablanca: Data and Models for Multidialectal Arabic Speech Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarrar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

Hajar Zaidani, Abderrahim Maizate, Mohammed Ouzzif, and Rim Koulali. 2024. Building a Corpus for the Underexplored Moroccan Dialect (CFMD) Through Audio Segmentations. *Revue d'Intelligence Artificielle*, 38(3):857–866.

A ASR Systems

Table 5 provides additional information about all ASR-systems currently included in the experiments of this pilot study. Our aim is to cover a broad spectrum of model sizes for base- as well as fine-tuned models, to allow comparisons across modalities.

Label	Name	Type	Size	Reference
DEDICATED ASR *(single-purpose end-to-end ASR)*				
MoulSot	MoulSot v0.3	*Fine-tune*	–	(Atlasia, 2026)
Omni7B	Omni ASR 7B	*Base model*	7 B	(team et al., 2025)
Qwen3ASR	Qwen3-ASR	*Base model*	1.7 B	(Shi et al., 2026)
W2V2-M	W2V2 Moroccan	*Fine-tune*	300 M	`hf:`/othrif/wav2vec2-large-xlsr-moroccan
W2V2-D	W2V2 Darija	*Fine-tune*	300 M	(Boumehdi and Yousfi, 2024)
WLarge	Whisper Large	*Base model*	1.55 B	(Radford et al., 2022)
WLarge-DL	Whisper Darija LoRA	*Fine-tune*	1.55 B	`hf:`/Hafsa0/whisper-large-v3-darija-lora
WLarge-M	Whisper Morocco FT	*Fine-tune*	1.55 B	(Abdullah et al., 2025)
WMedium	Whisper Medium	*Base model*	769 M	(Radford et al., 2022)
WMedium-D	Whisper Medium Darija	*Fine-tune*	769 M	`hf:`/ychafiqui/whisper-medium-darija
WSmall	Whisper Small	*Base model*	244 M	(Radford et al., 2022)
WSmall-D	Whisper Small Darija	*Fine-tune*	244 M	`hf:`/KandirResearch/Whisper-Small-Darija
WTurbo	Whisper Turbo	*Base model*	809 M	`hf:`/openai/whisper-large-v3-turbo
WTiny	Whisper Tiny	*Base model*	39 M	(Radford et al., 2022)
MULTIMODAL LLM / MULTITASK SPEECH *(transcription as one of many tasks)*				
Gemini	Gemini 2.5	*Proprietary*	–	(Google, 2025)
Gemma4	Gemma4 E4B Instruct	*Base model*	4 B (effective)	`hf:`/google/gemma-4-E4B-it
Qwen3Omni	Qwen3-Omni	*Base model*	30 B (3 B active)	`hf:`/Qwen/Qwen3-Omni-30B-A3B-Instruct
Seamless	Seamless-M4T	*Base model*	2.3 B	(Communication et al., 2023)

Table 5: ASR systems evaluated in this study, grouped by purpose. *Type*: whether the system is a general-purpose base model or a Darija-specific fine-tune. *Size*: approximate parameter count (– = undisclosed or proprietary). *Reference*: citation or path of the model on `huggingface.co`.