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, [*]Equal contribution

Tuesday 29th September, 2026

The Team



Avery C. Kanel



Christian Schuler



Bouazza Laracha



Imrane Lbouhli



Yassine Chaouri



Yusser Al Ghussin



Prof. Dr. Timo Baumann

Question of Interest

Are regional varieties of Moroccan Arabic
served equally well by ASR systems?

We compare word error rates (WER) of 17 ASR systems
on local radio audio from Casablanca, Tangier and Oujda.

Automatic Speech Recognition (ASR) Systems

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)

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`.

Why Regional Variation Matters

1 Regional variation

Moroccan Arabic varies substantially by region in its phonology, lexicon, and contact with other languages.

(Boumans, 2005; Aguadé, 2008; Heath, 2013; Falchetta and Guerrero, 2023)

2 Skewed corpora

Corpora may be regionally concentrated, as they are in other countries, including Egypt.

(Eida and Habash, 2025)

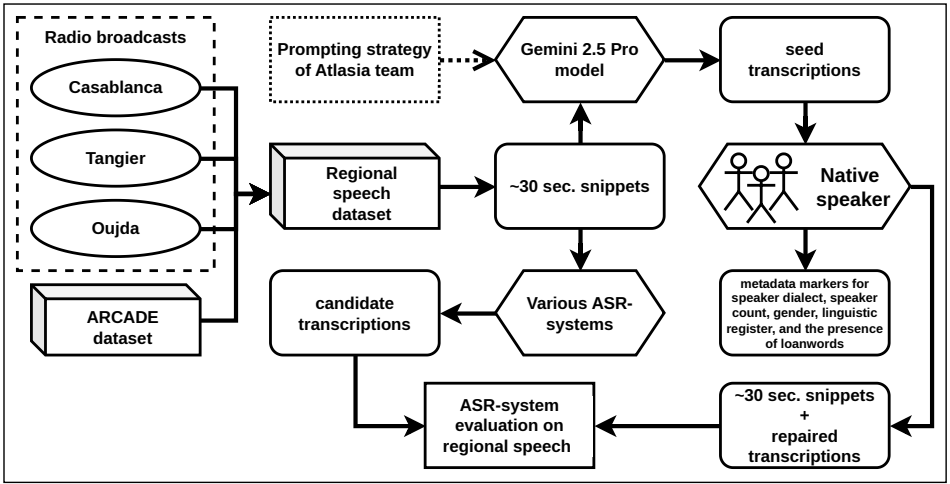
3 Performance gaps

ASR systems perform worse for speaker groups under-represented in their training data.

(Koenecke et al., 2020)

The problem: Moroccan speech corpora are labeled only at the national level, so such under-representation can neither be measured in training data nor detected in evaluation.

Our Approach in Short



ASR experiment pipeline.

From Radio to Machine to Human

1 Search

Searched for Moroccan Arabic audio datasets tagged at the city level: found ARCADE.

2 Record

6 hour-long radio sessions from local stations in Casablanca, Tangier and Oujda, morning to evening.

3 Trim

Audio manually trimmed to 27–30 s clips, favoring single-speaker segments.

4 Transcribe

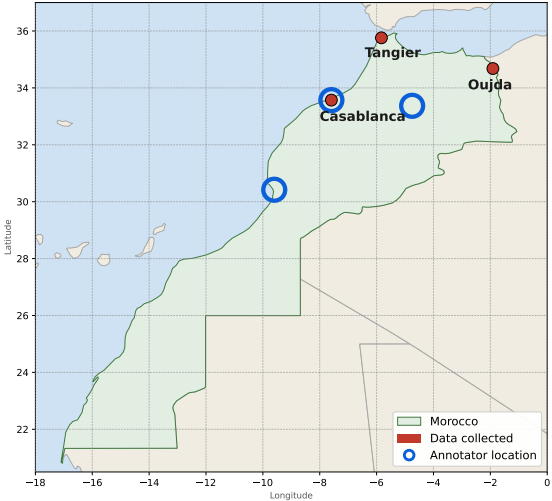
Gemini 2.5 Pro produces a machine transcription of each clip.

5 Annotate

Annotators correct the machine transcripts and add metadata tags.

Metadata tags: Speaker accent · Register · Loan-word language use · Genders present · Speaker count

Where does it comes from?



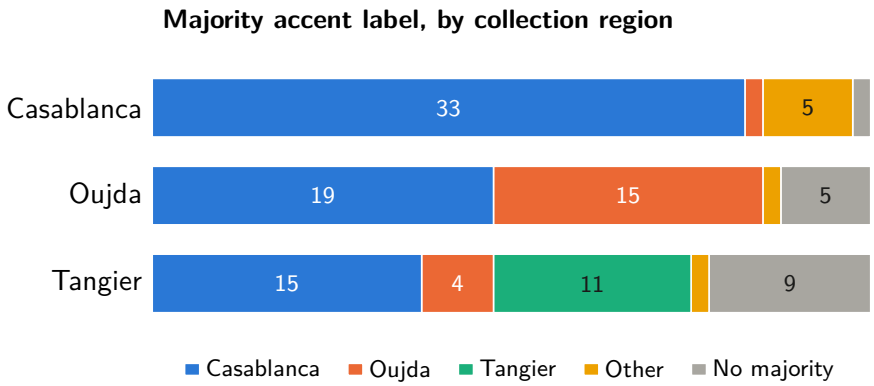
Data sources and human annotator locations.

Regional Comparison

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

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

What the Annotators tagged: Speaker Accent



Machine-Transcription System

	Normalized (LB)			Per-Region WER				sclite Decomposition		
Model	**WER**	**CER**	**Oracle**	**Casa.**	**Tang.**	**Oujda**	**Gap**	**Sub**	**Del**	**Ins**
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

ASR system scores.

Worst Observed Performance

	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)*										
WTiny	91.4	56.3	90.3	91.6	90.5	92.1	1.6	58.2	31.6	2.0
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

ASR system scores.

Best Observed Performance

	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)*										
WTiny	91.4	56.3	90.3	91.6	90.5	92.1	1.6	58.2	31.6	2.0
MoulSot	35.5	16.0	31.4	33.4	35.2	37.4	4.0	27.2	10.3	2.4
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

ASR system scores.

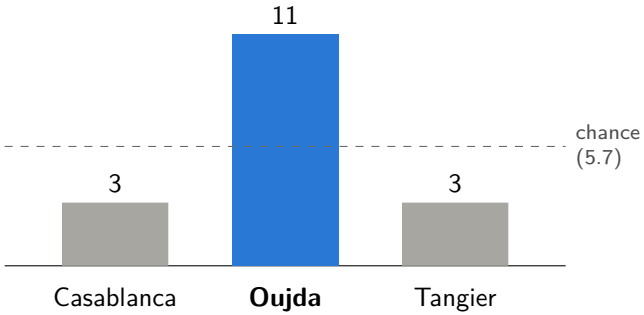
All Observed Performances

	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

All of the investigated ASR system scores.

Most systems do worst on Oujda

11 of 17 systems have their highest WER on Oujda, far more than the ≈6 expected by chance (χ^2 goodness-of-fit, df = 2, $p = .023$).



Systems with their highest WER in each region

Is Oujda just harder Audio?

	Measure	Casablanca	**Oujda**	Tangier	Oujda worse?
Noise	WADA-SNR (dB) ↑	15.0	**21.2**	10.8	no, cleaner
	Energy-percentile SNR (dB) ↑	23.6	**30.2**	20.6	no, cleaner
	DNSMOS background (1–5) ↑	2.86	2.67	2.80	no
Speaking rate	Characters / s of speech	9.8	9.9	11.1	no
	Words / s of speech	2.08	2.30	2.28	≈ Tangier
Bandwidth	99% spectral roll-off (kHz) ↑	5.02	5.15	5.45	no

Outlook & Future Work

International Conference LAnguage TEchnologies for Low-resource Languages

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

Avery C. Kanel[1,*] Christian Schuler[1,*]
Bouacha Lamiche[2,*] Jasmin Ulrichs[3,*] Yassine Chaouni[2,*]
Yasser Al Ghussin[1,*] Timo Baumann[1]

[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 assumption of uniformity by introducing an annotated speech corpus, 431 minutes, of broadcast audio from Casablanca, Tangier, Tetouan, and Oujda, spanning 27 hours of 40 speakers, and evaluating 17 state-of-the-art ASR systems against it. Our results show that systems trained on Moroccan speech overwhelmingly exhibit significant geographic gaps in WER points between the best- and worst-served regions, and fine-tuned "Moroccan Arabic" checkpoints exhibit inconsistent behavior, sometimes degrading performance on balanced training data. These findings suggest that national-level dialect labels obscure substantial variation and underscore the need for performance consequences. We release the dataset[1] and support the adoption of more fine-grained provenance labeling as standard practice for Arabic dialect resources.



Figure 3: Data sources and human annotator locations.

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 syntactic divides, that characterize the dialect (Caubet, 2006; Boumalk, 2013). This also exists within a state of deglossia 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 (Bentahila et al., 2021; Herman, 2023).

Recent years have seen substantial progress in Darija NLP, from large-scale corpora (Outchakoucht and Es-Samaali, 2021; Shaitarova et al., 2024; Boumehdi and Terrab, 2024) to benchmarks for speech recognition (Bouhandi et al., 2024; Ait-Mokhtar et al., 2025; Issa et al., 2025) and sentiment analysis (Outchakoucht and Es-Samaali, 2021; Issa et al., 2025). However, an open question whether these advances equally benefit speakers across all of Morocco's regions. Most Darija datasets are collected from users of their training data (Kerrouche et al., 2025). If Moroccan Arabic corpora are, in fact, regionally concentrated as they are in other North African countries (Mohamed Faka and Habash, 2025), without disclosing this at the national label level, speakers from less-represented areas face systematic performance degradation. This is not a new problem, it stems published benchmarks or model cards (Kirkedal and Maegaard, 2025; Zaidan et al., 2024).

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



30th Annual Conference of the Foundation for Endangered Languages

Surveying Perspectives and Perceived Needs of Moroccan Arabic and Tamazight Speakers on Language Technologies

Avery C. Kanel[1,*] Christian Schuler[1,*] Yasser Al Ghussin[1,*]
Yassine Chaouni[2] Rafi Al Attass[3] Simon Ostermann[1,4] Josef van Genabith[1,4]

[1]Saarland University [2]German Research Center for Artificial Intelligence (DFKI)
[3]Hassan II University of Casablanca [4]Independent Researcher

Abstract

Morocco is a country home to around 34 million speakers of Moroccan Arabic (Darija) and Tamazight varieties, as well as Standard Tamazight varieties (Central Atlas Tamazight, Tarifit, and Tachelhit). Language technologies are not equally available to these groups, with most technology being digital in form. However, outside of Europe, very few countries have conducted NLP-community-wide surveys for unofficial and minority languages. Our study bridges this gap by surveying 98 participants. We survey speakers of these languages (56.2% and ask both about their attitudes towards language technologies in general and which ones they would find useful. We found the majority of respondents were willing to contribute data to support these technologies, especially chatbots. We also scored the community was split on the topic of spell and grammar checkers and standardized orthographies. Finally, for speakers of written Tamazight speakers, all respondents reported Darija as the language they felt closest to.

1 Introduction

Morocco is home to a dense and layered ecology of language varieties, Darija and Tamazight being two that are under pressure and changing from diglossia (Fathi, 2021) and disappearance of dialectal diversity (Kan, 2020; Bouadaa, 2025). Recent sociolinguistic work has begun to document this narrowing from marginalization (wa) and institutionalization (Idhssaine, 2022), while also noting that implementation remains uneven in education, media, or media and public communication (Belkadi et al., 2024).

Darija, spoken by nearly all of Morocco's 37 million inhabitants as a first or second language (Heath, 2020), appears related by speakers to count. Yet its rich regional diversity (phonology (Caubet, 2006), morphosyntax (Boumalk, 2013; Idhssaine, 2025), is being leveled by urbanization-driven homogenization (Violakaita and Guerraoui, 2022; Hachimi, 2012), the process by which a new simplified dialect emerges in a city where speakers of different, mutually unintelligible dialects mix, while levelers and participants dismantle structure accepted (Trudgill, 1986). Darija is also in a state of extreme diglossia with pervasive code-switching (Heath et al., 2021; Herman, 2023). Tension thus operates at two levels at once, with the endangerment of the Tamazight languages, and the quieter disappearance of regional Darija varieties, the latter of which languages resources routinely flatten into a single "Moroccan Arabic" label (Heath, 2017).

We present a community-centered survey investing speakers of Darija and Tamazight varieties to share how they relate to their language, what they want and do not want from language technology and what role they think technology can play. Overall, the survey asks the following without prompting that standardization or increased digitization are self-evidently desirable

1. **Language Use:** Which languages the respondents know and in what settings they use them.
2. **Language Beliefs:** Respondent attitudes on variation in and standardization of their languages.
3. **Language Technologies:** Priorities of respondents for future language technologies and their disposition towards the field in general.

Findings will be reported back to the participating communities via the same channels used to advertise this survey to begin with, and will directly shape what we build and study next.





References I

Badr M. Abdullah, Yusser Al Ghussin, Zena Al-Khalili, Ömer Tarik Özyilmaz, Matias Valdenegro-Toro, Simon Ostermann, and Dietrich Klakow. Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing under Data Constraints. In Kareem Darwish, Ahmed Ali, Ibrahim Abu Farha, Samia Touileb, Imed Zitouni, Ahmed Abdelali, Sharefah Al-Ghamdi, Sakhar Alkhereyf, Wajdi Zaghouani, Salam Khalifa, Badr AlKhamissi, Rawan Almatham, Injy Hamed, Zaid Alyafeai, Areeb Alowisheq, Go Inoue, Khalil Mrini, and Waad Alshammari, editors, *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 745–751, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-356-2. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.102. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.102/.

Jordi Aguadé. 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, 2008. ISBN 978-90-04-14475-0. URL https://www.academia.edu/9064836/Morocco_dialectological_survey_.

Atlasia. MoulSot: A curated moroccan darija speech dataset and fine-tuned ASR model, 2026. URL https://huggingface.co/atlasia/moulsot.v0.3.

Louis Boumans. Lexical variation in Moroccan Arabic: The genitive exponent in experimental data from four cities. *EDNA, Estudios de dialectología norteafricana y andalusí*, 9:125–137, 2005. URL https://www.academia.edu/24548293/Lexical_variation_in_Moroccan_Arabic_The_genitive_exponent_in_experimental_data_from_four_cities.

Ahmed Boumehdi and Abdellah Yousfi. MDVC corpus: Empowering Moroccan Darija speech recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(1):290–301, April 2024. ISSN 2502-4760. doi: 10.11591/ijeecs.v34.i1.pp290-301. URL https://ijeecs.iaescore.com/index.php/IJEECS/article/view/35709.

References II

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. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation, October 2023. URL http://arxiv.org/abs/2308.11596.

Mai Mohamed Eida and Nizar Habash. Beyond Cairo: Sa'idi Egyptian Arabic Literary Corpus Construction and Analysis. In Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar, editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 292–304, Albuquerque, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-234-3. doi: 10.18653/v1/2025.nlp4dh-1.26. URL https://aclanthology.org/2025.nlp4dh-1.26/.

Jacopo Falchetta and Jairo Guerrero. Koineization trends in spoken arabic across three moroccan towns. *Kervan – International Journal of Afro-Asiatic Studies*, 27(1), 2023. URL https://ojs.unito.it/index.php/kervan/article/view/7714.

Google. Gemini 2.5 Pro, 2025. URL https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro.

Jeffrey Heath. *Jewish and Muslim Dialects of Moroccan Arabic*. Routledge, London, January 2013. ISBN 978-0-203-03736-2. doi: 10.4324/9780203037362.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, December 2022. URL http://arxiv.org/abs/2212.04356.

References III

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. Qwen3-ASR Technical Report, 2026. URL https://arxiv.org/abs/2601.21337.

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. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages, 2025. URL https://arxiv.org/abs/2511.09690.

ALAMA QR code



`github.com/Ackanel/alama_audio_dataset`

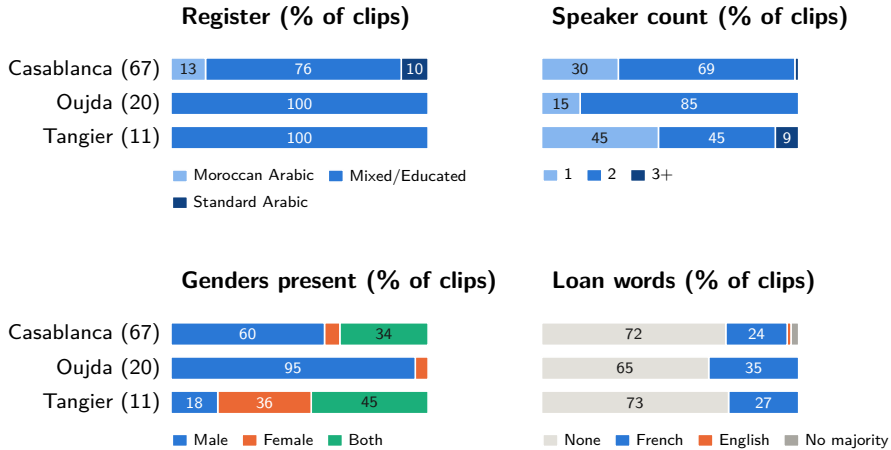
Machine-Transcription Prompt

```
prompt = """\
You are an audio transcriber. Your job is to transcribe the input audios
to text exactly as it was said by the user.

Rules:
* Audio is Moroccan Darija Arabic.
* Do not transliterate: if words from another language are uttered
  (e.g. French, English), transcribe them in their original script.
* If the audio is not clear, respond with "<filename.wav>:" (empty).
"""
```

Prompt used for the Gemini machine pre-transcriptions, taken from the MoulSot pipeline (Atlasia, 2026).

What the Annotators tagged, by Perceived Accent



What the Annotators tagged by broadcast location

