Audar-ASR-V1: A Multilingual, Arabic-First Generative Speech Recognition Foundation Model

Audar AI Team*

Flash · Turbo | Code | Website

Abstract

Audar-ASR-V1 is an Arabic-first generative speech recognition model and the new state of the art on the Open Universal Arabic ASR Leaderboard. Built on a permissively-licensed open-weight audio-LLM foundation and adapted in-house through a four-stage curriculum on 300,000+ hours of labeled audio, Audar-ASR-V1 attains 24.78% average WER on full test sets across six standard Arabic benchmarks — first among the 35 systems ranked on the leaderboard, ahead of the strongest proprietary API system and 3.5pp ahead of the best open-weight model at a third of its size, with the lowest WER on dialect-heavy SADA and on MGB-2 broadcast. We release the model weights in two tiers — Audar-ASR-V1-Flash (0.6B) under the AudarAI Open License and Audar-ASR-V1-Turbo (2B) under the AudarAI Community License — together with the six-dataset evaluation harness so the community can reproduce these results and build on them.

Highlights

- **State-of-the-art Arabic ASR.** 24.78% average WER on full test sets — first of the 35 systems ranked on the leaderboard, ahead of the strongest proprietary API and 3.5pp ahead of the best open-weight 7B model at a third of the size (Figure 1); Audar-ASR-V1-Flash (0.6B decoder) beats its same-class baseline by 8.9pp.
- **Open weights.** Audar-ASR-V1-Flash (0.6B) ships under the AudarAI Open License and Audar-ASR-V1-Turbo (2B) under the AudarAI Community License, together with the six-dataset evaluation harness.

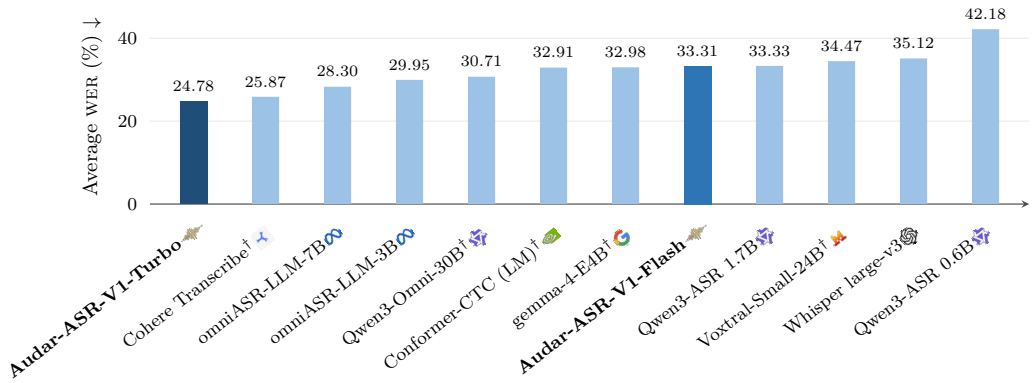

Figure 1: Average WER (%, lower is better) on the six leaderboard test sets (Wang et al., 2025): rank 1 of 35. † = leaderboard-published; others re-run by us.

*Correspondence: research@audarai.com

1 Introduction

Arabic is among the most challenging targets for automatic speech recognition. The language exists in a state of diglossia: Modern Standard Arabic (MSA) is the written and broadcast register, while everyday speech is carried by a wide spectrum of regional dialects that diverge from MSA and from one another in phonology, lexicon, and morphology. Arabic is morphologically rich, with clitics and templatic inflection that inflate the effective vocabulary, and it is conventionally written without short-vowel diacritics, so a single orthographic form maps to many pronunciations and readings. Speakers routinely code-switch into English and French, particularly in technical and urban registers. Together these properties make verbatim transcription, diacritization, and named-entity preservation simultaneously hard, and they have historically kept Arabic word error rates well above those reported for high-resource languages.

Evaluation has also been fragmented, with Arabic speech systems long assessed on disjoint corpora and protocols (Abdelali et al., 2024). The Open Universal Arabic ASR Leaderboard (Wang et al., 2025) consolidated the field in 2025 by scoring systems uniformly across six standard datasets spanning read, broadcast, conversational, and dialectal speech. The dialect-heavy benchmarks in particular (SADA (Alharbi et al., 2024), Casablanca (Talafha et al., 2024)) remain far from solved, with the best reported systems still well above 30% WER. A recent mapping of the Dialectal Arabic speech-technology landscape across 31 corpora and 14 dialects reinforces that data scarcity and dialectal variation remain the dominant obstacles for Arabic ASR, keeping error rates well above those of high-resource languages (Sullivan et al., 2026).

Three modeling lines shape the current ASR landscape. Self-supervised speech encoders — wav2vec 2.0, HuBERT, and WavLM (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) — learn transferable representations that are fine-tuned per task. Large-scale supervised encoder–decoder recognition, exemplified by Whisper and SeamlessM4T (Radford et al., 2023; Seamless Communication et al., 2023), scales multilingual transcription directly. Most recently, audio-conditioned language models such as Qwen2-Audio and Qwen2.5-Omni (Chu et al., 2024; Xu et al., 2025) cast transcription as generative decoding, with instruction-followable audio understanding explored further in SALMONN, Audio Flamingo, and GAMA (Tang et al., 2023; Kong et al., 2024; Ghosh et al., 2024) and adopted by proprietary systems such as GPT-4o and Gemini (OpenAI, 2024; Gemini Team, Google, 2025). Audar-ASR-V1 follows the generative line and pairs it with preference alignment (Ethayarajh et al., 2024; Rafailov et al., 2023).

We introduce Audar-ASR-V1, an Arabic-first generative speech recognition model that recasts transcription as audio-conditioned next-token prediction over a unified text vocabulary — leveraging a language-model decoder rather than aligning under a connectionist temporal classification (CTC) (Graves et al., 2006) or transducer (Graves, 2012) objective. Audar-ASR-V1 is built on a permissively-licensed open-weight audio-LLM foundation in the 0.6B / 2B parameter class (Chu et al., 2024) and is substantially adapted in-house through 300,000+ hours of labeled audio and a four-stage curriculum that ends in preference alignment from native Arabic annotators. The contribution is the adaptation — the data curriculum and the alignment rubric — not the foundation.

On the Open Universal Arabic ASR Leaderboard, Audar-ASR-V1-Turbo achieves 24.78% average WER on full test sets across the six benchmarks — first among the 35 ranked systems, ahead of the strongest proprietary API system, with the lowest WER on dialect-heavy SADA and on MGB-2 broadcast. The 0.6B-decoder Audar-ASR-V1-Flash tier beats its same-class baseline Qwen3-ASR-0.6B by 8.9pp. On an in-house Gulf-Emirati long-form benchmark Audar-ASR-V1 also achieves lower WER than ElevenLabs Scribe v1 and Qwen3-ASR.

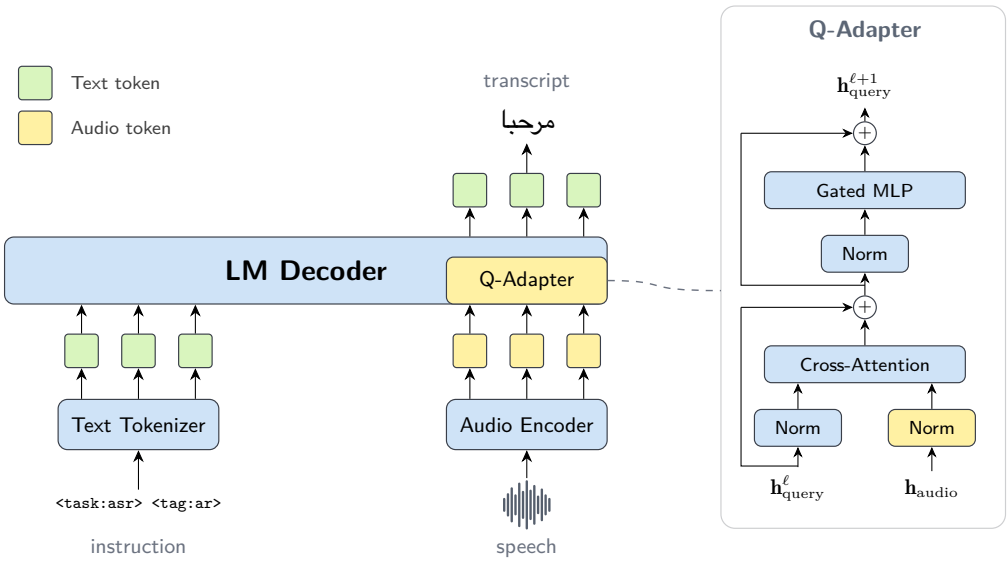

Figure 2: End-to-end system architecture. Audio is encoded by a Conformer encoder, compressed by the Q-Adapter into a fixed-rate 12.5 Hz soft-prompt stream, and decoded by a 0.6B / 2B decoder-only language model that consumes the soft prompts alongside instruction and task-tag text tokens (e.g. `<task:asr>`, `<tag:ar>`).

The remainder of this report describes the encoder–adapter–decoder architecture (Section 2), the four-stage adaptation curriculum and its KTO alignment rubric (Section 3), the full-test-set results that establish the state of the art (Section 4), and how to download the model (Section 5).

2 Method

This section describes the Audar-ASR-V1 architecture; the four-stage training curriculum that adapts a permissively-licensed open-weight foundation into an Arabic-first recognizer is presented in Section 3.

2.1 Architecture

Audar-ASR-V1 is a three-module differentiable pipeline: a Conformer acoustic encoder, a query-based cross-modal adapter, and a generative decoder-only language model in the 0.6B / 2B parameter class; tier names refer to the decoder's parameter count, and each tier pairs its decoder with a matched Conformer encoder (0.19B on Audar-ASR-V1-Flash, 0.32B on Audar-ASR-V1-Turbo). Audio enters as a 16 kHz waveform, is converted to log-mel features, encoded into embeddings at 25 Hz, compressed by the adapter into a 12.5 Hz soft-prompt stream, and decoded into UTF-8 text (Figure 2).

Acoustic frontend and encoder. Raw 16 kHz PCM is converted to log-mel filterbanks with per-utterance normalization to absorb channel variability across studio, telephone, and in-vehicle audio. A Conformer encoder (Gulati et al., 2020) (0.19B parameters on Audar-ASR-V1-Flash, 0.32B on Audar-ASR-V1-Turbo) processes the feature stream, with convolutional subsampling to 25 Hz, so the encoder emits one embedding per 40 ms of audio.

Cross-modal adapter. A learned-query adapter (Q-Adapter) compresses the variable-length 25 Hz embedding stream into a fixed 12.5 Hz soft-prompt stream consumable by the

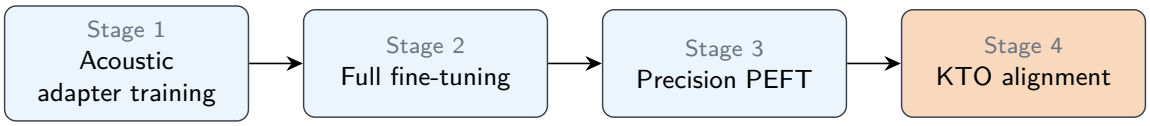

Figure 3: Four-stage training curriculum. Each node on the rail is an artifact (a checkpoint that passes its quality gate); each card describes the stage that produced it.

decoder: learnable query vectors attend over the audio embeddings through a small stack of cross-attention layers, emitting one decoder-dimensional vector per 80 ms of audio. Unlike concatenation or convolutional projectors, the query adapter holds decoder context constant per unit of audio time, independent of speech content.

Generative decoder. The decoder is a decoder-only transformer with grouped-query attention (Ainslie et al., 2023), SwiGLU MLPs (Shazeer, 2020), pre-norm RMSNorm, rotary position encoding (Su et al., 2024), and tied input/output embeddings. Two tiers share the same tokenizer, adapter interface, and prompt protocol: an edge tier, Audar-ASR-V1-Flash (0.6B), and a server tier, Audar-ASR-V1-Turbo (2B), with a context window covering ≈ 30 s of audio plus prompt and transcript. The decoder consumes the adapter's soft-prompt tokens, an optional textual instruction prefix, and a task tag, and emits subword tokens autoregressively — making transcription instructible (language, formatting, biasing, diacritization steered by prompt without retraining).

Vocabulary and tokenization. The byte-pair-encoding tokenizer is Arabic-morphology-aware (affix-rich forms compress to a few tokens rather than many) and diacritization-preserving (diacritics are first-class characters in the BPE alphabet). Reserved control tokens steer language and task at decode time: language tags (`<tag:ar>`, `<tag:en>`, `<tag:ar+en>`) and task tags (`<task:asr>`, `<task:asr-diac>`, `<task:ast>`).

3 Training

Audar-ASR-V1 is trained in four stages, each addressing a failure mode left by the previous one (Figure 3). A checkpoint advances only after passing an automated quality gate covering per-dialect held-out WER, English-regression, and calibration checks.

Stage 1 — Acoustic adapter training. The encoder, adapter, and decoder are initialized from a permissively-licensed open-weight audio-LLM foundation in the 0.6B / 2B class (Chu et al., 2024). With the full model frozen, only the acoustic adapter (Q-Adapter) is trained on labeled Arabic and English audio. The adapter learns to project the encoder's representations into the decoder's embedding space without disturbing the pretrained weights, establishing a stable cross-modal bridge before any backbone parameters are updated.

Stage 2 — Full fine-tuning with adapter. All model parameters are unfrozen and training continues with the adapter in place. Joint optimization of encoder, adapter, and decoder allows the entire system to co-adapt to the Arabic speech domain. This stage produces the multi-task base from which specialization branches.

Stage 3 — Precision PEFT. The adapter weights are frozen and parameter-efficient fine-tuning (LoRA (Hu et al., 2022)) is applied to the decoder using curated, high-quality

data. This stage targets specific weaknesses identified during Stage 2 evaluation — dialectal variants, proper nouns, and code-switching patterns.

Stage 4 — kto alignment. Kahneman-Tversky Optimization (Ethayarajh et al., 2024) is applied on accented dialectal Arabic data, with preference labels from trained native Arabic annotators spanning the Gulf, Levantine, Egyptian, and Maghrebi dialect regions. The preference data is annotated along five axes: verbatim accuracy, diacritic correctness, code-switch handling, named-entity preservation, and output formatting. Annotations are unpaired (single response with a binary desirability label), which KTO consumes directly and which matches the annotation pipeline far better than DPO's paired preferences (Rafailov et al., 2023): annotators frequently rate two transcripts equally bad on different axes, yielding noisy pairs. This stage improves dialect-heavy subsets (SADA, Casablanca) and substantially reduces hallucination on long-form Gulf-Emirati audio.

Training data. The full training corpus comprises 300,000+ hours of labeled audio combining proprietary Arabic speech from internal collections and licensed partnerships, permissively-licensed open-source corpora serving as anchor data, and synthetic augmentation for code-switching and dialect coverage. All open-source datasets are permissively licensed (CC-BY-SA or Apache-2.0); proprietary data is used under internal research licenses with no redistribution rights. Training utterances are de-duplicated against every evaluation set (Section 4.1). A data card will be published alongside the model release.

Training efficiency. Audar-ASR-V1 is built by *adaptation rather than scale*: the entire four-stage curriculum completes in days of wall-clock time on a commodity multi-GPU cluster — a small fraction of the cost of training a speech foundation model from scratch. A checkpoint advances to the next stage only after passing its quality gate, and the final checkpoint is selected on validation-set WER plateaus.

4 Evaluation

We describe the training and evaluation data (Section 4.1), the primary full-test-set comparison, and diagnostic analyses.

4.1 Data

Audar-ASR-V1 is pretrained on 300,000+ hours of labeled audio combining licensed proprietary recordings with permissively-licensed open-source corpora, spanning Arabic (primary), English, and a multilingual tail. The dominant Arabic fraction (MSA, dialectal, and code-switched) carries the specialization, while the English and multilingual shares preserve the foundation's general competence.

Dialect coverage. Arabic coverage spans Gulf, Levantine, Egyptian, and Maghrebi dialect families plus MSA registers (formal and broadcast), with Arabic↔English code-switching as a first-class condition. Stage-3 specialization (Section 3) targets eight dialect groups assembled from these families.

Evaluation integrity. Training utterances are de-duplicated against every evaluation set with an acoustic-fingerprint filter, so no test audio or near-duplicate enters training.

Table 1: Held-out evaluation sets. "Dialects" is the number of distinct dialect labels in the test split.

Dataset	Test (h)	Dialects	Source
CommonVoice 18	12.6	MSA read	Mozilla crowd-sourced
MASC Clean	10.5	7	YouTube (clean)
MASC Noisy	8.9	14	YouTube (noisy)
MGB-2	9.6	5	Al Jazeera broadcast
SADA	10.7	10	Saudi Audio Dataset
Casablanca	8.0	8	multi-country Darija
Gulf-Emirati (Alsanaa)	4.5	Emirati	long-form recordings

Held-out evaluation sets. Table 1 summarizes the held-out sets: the six Open Universal Arabic ASR Leaderboard datasets (Wang et al., 2025) — CommonVoice 18 (Ardila et al., 2020), MASC (Al-Fetyani et al., 2022), MGB-2 (Ali et al., 2016), SADA (Alharbi et al., 2024), and Casablanca (Talafha et al., 2024) — plus the Gulf-Emirati (Alsanaa) long-form set, spanning read, broadcast, conversational, clean, and noisy speech across MSA and the major dialect families.

4.2 Main Results

Our primary evaluation uses full test sets across all six benchmarks under the leaderboard-equivalent normalization protocol (Wang et al., 2025) (Table 2). Audar-ASR-V1-Turbo (2B) achieves 24.78% average WER across the six benchmarks — the lowest of the 35 systems ranked on the Open Universal Arabic ASR Leaderboard (the 34 public entries plus ours): 1.1pp ahead of the strongest proprietary API system, Cohere Transcribe (arabic-07-2026), and 3.5pp ahead of the best open-weight model, omniASR-LLM-7B, at roughly a third of its parameter count.

Table 2 reports our matched evaluation: every system is run by us with the same harness, decoding settings, and normalizer on the full test sets — Meta's Omnilingual ASR family at four scales (Omnilingual ASR team et al., 2025), Qwen3-ASR at 0.6B and 1.7B (Qwen Team, 2025), and Whisper large-v3 (Radford et al., 2023). Audar-ASR-V1-Turbo posts the lowest WER in every column of this evaluation among the open-weight systems; for completeness the table also carries the leaderboard-published row of the strongest proprietary system, Cohere Transcribe (arabic-07-2026). Against the full public leaderboard, Audar-ASR-V1-Turbo ranks first on average WER, first on SADA (by 8.1pp) and MGB-2, second on Casablanca and MASC clean, third on MASC noisy, and fourth on CommonVoice 18 — the read-speech and noisy-read sets are led by proprietary API systems, while Audar-ASR-V1-Turbo leads where Arabic ASR is hardest: spontaneous dialectal speech and broadcast.

Audar-ASR-V1 also ships a compact edge tier, Audar-ASR-V1-Flash (0.6B decoder, 0.78B total), that shares the same architecture, adapter design, and training curriculum with a smaller matched encoder. At 33.31% average WER, Audar-ASR-V1-Flash beats its same-decoder-class baseline Qwen3-ASR-0.6B (42.18%) by 8.87pp and effectively matches Qwen3-ASR-1.7B (33.33%) with roughly a third of the decoder parameters (Table 2).

On the full test sets (Table 2), Audar-ASR-V1-Turbo reaches 29.41% on SADA (Saudi/Gulf) — the lowest of any leaderboard system, 8.1pp ahead of the next best — and 51.58% on Casablanca (Moroccan Darija), within 1.9pp of the leaderboard-best API system; the high absolute Casablanca WER reflects the scarcity of curated Darija data, the hardest Arabic dialect for every leaderboard system, rather than an architectural weakness. Full per-dataset WER/CER for both tiers are given in Table 4.

Table 2: Full-test-set WER (%) across the six Open Universal Arabic ASR Leaderboard benchmarks, sorted by average. Open-weight systems evaluated by us with the same harness and leaderboard-equivalent normalization (see Methodology); runs of leaderboard-listed baselines reproduce their published averages within ±0.12pp. †Cohere Transcribe is a proprietary API system that cannot be run in our harness; its row carries the leaderboard-published numbers (same normalization). Params = total parameters.

System	Params	Avg	CV18	MASC-C	MASC-N	MGB-2	SADA	Casablanca
Audar-ASR-V1-Turbo	2.35 B	**24.78**	8.60	**19.60**	28.35	**11.13**	**29.41**	51.58
Cohere Transcribe (arabic-07-2026)†	—	25.87	**5.82**	**19.60**	**27.07**	15.54	37.47	**49.71**
omniASR-LLM-7B	7 B	28.30	8.97	19.70	29.20	13.96	41.65	56.33
omniASR-LLM-3B	3 B	29.95	9.14	19.89	30.04	14.20	46.10	60.34
omniASR-LLM-1B	1 B	30.08	9.62	19.99	30.55	15.29	44.10	60.90
omniASR-LLM-300M	0.3 B	33.01	12.31	20.67	32.44	16.56	51.39	64.68
Audar-ASR-V1-Flash	0.78 B	33.31	16.02	25.96	35.43	17.11	44.53	60.79
Qwen3-ASR-1.7B	1.7 B	33.33	16.75	24.31	34.29	16.64	43.52	64.49
Whisper large-v3	1.55 B	35.12	15.15	22.52	32.87	15.30	55.54	69.36
Qwen3-ASR-0.6B	0.6 B	42.18	28.19	31.32	42.61	25.46	53.66	71.81

Table 3: Gulf-Emirati Alsanaa long-form benchmark (102 recordings, 272 min). WER/CER in %; lower is better.

Model	WER	CER
Audar-ASR-V1-Turbo	**30.02**	**13.74**
ElevenLabs Scribe v1	35.05	15.26
Qwen3-ASR-1.7B	42.67	16.85

Gulf-Emirati long-form. On an in-house Gulf-Emirati (Alsanaa) set of 102 long-form recordings (272 minutes) from native Emirati speakers, Audar-ASR-V1 reaches 30.02% WER — 5.03pp better than ElevenLabs Scribe v1 and 12.65pp better than Qwen3-ASR-1.7B (Table 3). The margin over both baselines demonstrates robust real-world performance on conversational Gulf-dialect audio.

Methodology. All WER/CER apply identical Arabic normalization to reference and hypothesis (tashkeel and tatweel removal, hamza normalization, teh-marbuta unification, Eastern-to-Western digit conversion, punctuation removal, whitespace collapse, Latin lowercasing). WER is micro-averaged; average WER is the arithmetic mean of per-dataset scores. All evaluations (Tables 2 and 4) use full test sets for every system, run by us with one harness and one normalizer. The upstream leaderboard evaluation script contains a regex bug that silently skips punctuation stripping; all numbers here use the corrected, leaderboard-equivalent normalizer, and our runs of leaderboard-listed baselines reproduce their published averages within ±0.12pp (Whisper large-v3 measures 1.74pp better than its published cell). Every sample returning non-empty output is included; no system produced empty outputs on any set.

4.3 Detailed Full-Test-Set Breakdown

Table 4 provides the detailed per-dataset WER and CER breakdown for both released tiers.

Audar-ASR-V1-Turbo posts sub-3% CER on CommonVoice 18 and sub-6% CER on MASC clean and MGB-2; Audar-ASR-V1-Flash tracks it at a 6–15pp WER offset that widens with dialect density. Both tiers maintain the family's advantage on the dialect-heavy sets (SADA, Casablanca) at full scale.

Table 4: Full-test-set WER (%) and CER (%) for both released tiers across all six leaderboard benchmarks (leaderboard-equivalent normalization). n ranges from 1,045 (Casablanca) to 2,617 (CommonVoice 18).

Tier	Metric	CV18	MASC-C	MASC-N	MGB-2	SADA	Casablanca
Audar-ASR-V1-Turbo	WER (%)	8.60	19.60	28.35	11.13	29.41	51.58
	CER (%)	2.60	5.69	9.95	5.97	13.48	19.24
Audar-ASR-V1-Flash	WER (%)	16.02	25.96	35.43	17.11	44.53	60.79
	CER (%)	5.04	7.84	12.66	7.97	23.63	24.85

4.4 Impact of Adaptation

The adaptation is most visible at matched scale. Audar-ASR-V1-Flash shares its 0.6B decoder class with Qwen3-ASR-0.6B yet posts 33.31% average WER against 42.18% — an 8.87pp (21% relative) reduction attributable to the Arabic-centric curriculum rather than scale — and effectively matches Qwen3-ASR-1.7B with a third of the decoder parameters. Data then compounds with scale: Audar-ASR-V1-Turbo (2.35B total) beats the strongest open-weight model, omniASR-LLM-7B, by 3.5pp average WER at a third of its size — and edges the strongest proprietary API system by 1.1pp — with the gains concentrated where foundations struggle most (SADA 29.41% vs. 41.65% for omniASR-LLM-7B; Casablanca 51.58% vs. 56.33%). The curriculum also buys robustness: on SADA, Audar-ASR-V1 emits 0.0% boilerplate-hallucination transcripts versus 8.3% for Whisper large-v3, with zero empty outputs across all six sets. These gains come from continuous pretraining, multi-task SFT, dialect specialization, and KTO alignment, not from added scale; the largest single-stage gain comes from full fine-tuning on the dialect-heavy benchmarks, with a further reduction in dialect-subset WER and long-form hallucination from the final KTO alignment stage.

Claims and Evidence Summary Table 5 maps each headline quantitative claim in this paper to its supporting evidence and evaluation protocol, providing a single point of reference for readers and reviewers.

Table 5: Mapping of headline claims to supporting evidence.

Claim	Evidence	Evaluation Protocol
Lowest avg. WER (24.78%) of the 35 leaderboard systems	Table 2; leaderboard snapshot 2026-07-07	Full test sets, leaderboard-equivalent normalization
Rank 1 on SADA (by 8.1pp) and MGB-2; rank 2 on Casablanca and MASC clean	Table 2; leaderboard snapshot 2026-07-07	Full test sets, leaderboard-equivalent normalization
Audar-ASR-V1-Flash beats same-class Qwen3-ASR-0.6B by 8.87pp	Table 2	Matched harness, full test sets
Gulf-Emirati WER of 30.02%	Table 3	Internal benchmark, 102 recordings

5 Model Access

Open weights. We release the Audar-ASR-V1 model weights in two tiers: Audar-ASR-V1-Flash (0.6B) under the AudarAI Open License v1.0 (commercial use, redistribution, and

modification permitted with attribution) and Audar-ASR-V1-Turbo (2B) under the AudarAI Community License v1.0 (research and limited commercial use; enterprise deployment under a separate agreement) — together with the evaluation harness for all six Open Universal Arabic ASR Leaderboard datasets. Anyone can reproduce the results in this report, benchmark against Audar-ASR-V1, and fine-tune it for new domains and dialects. Both tiers expose one interface, so an application can move between them without code changes.

Getting started. Model weights are available on Hugging Face at https://huggingface.co/audarai; code and the evaluation harness live at https://github.com/AudarAI/Audar-ASR-V1. See https://www.audarai.com/ for the broader Audar model family.

6 Conclusion

Audar-ASR-V1 advances Arabic speech recognition, achieving 24.78% average WER on full test sets across the six Open Universal Arabic ASR Leaderboard benchmarks — first among the 35 ranked systems, ahead of the strongest proprietary API system, and first on dialect-heavy SADA and MGB-2 broadcast — while the 0.6B-decoder Audar-ASR-V1-Flash tier beats its same-class baseline by 8.9pp. The result comes not from a new architecture but from a four-stage curriculum — acoustic-adapter training, full fine-tuning, precision PEFT, and KTO preference alignment on Arabic-specific axes — applied to a permissively-licensed open-weight foundation; the curriculum is the transferable recipe, and the open evaluation harness across the six leaderboard datasets — together with the evaluation protocol and per-sample metrics for a Gulf-Emirati long-form set whose audio cannot be redistributed under speaker-consent restrictions — is our contribution to the community. We will release the trained models, evaluation scripts, and a detailed data card with the model release. Audar-ASR-V1 is the recognition foundation of a broader Arabic audio program, released alongside its expressive-synthesis counterpart Audar-TTS-V1 (concurrent report, Audar AI 2026).

References

Ahmed Abdelali et al. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024. URL https://arxiv.org/abs/2305.14982.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proc. EMNLP*, 2023. URL https://arxiv.org/abs/2305.13245.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. MASC: Massive Arabic speech corpus. In *IEEE Spoken Language Technology Workshop (SLT)*, 2022. URL https://ieeexplore.ieee.org/document/10022652.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, et al. SADA: Saudi audio dataset for Arabic. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290, 2024. URL https://ieeexplore.ieee.org/document/10446243.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, 2016. URL https://arxiv.org/abs/1609.05625.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020. URL https://arxiv.org/abs/1912.06670.

Audar AI. Audar-TTS-V1: A multilingual, Arabic-first expressive speech synthesis foundation model. Audar AI technical report, concurrent launch, 2026. URL https://www.audarai.com/. Report and code at https://github.com/AudarAI/Audar-TTS-V1; models at https://huggingface.co/audarai.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022. URL https://arxiv.org/abs/2110.13900.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. URL https://arxiv.org/abs/2402.01306.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Sreyan Ghosh et al. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2406.11768.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376, 2006.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020. URL https://arxiv.org/abs/2005.08100.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024. URL https://arxiv.org/abs/2402.01831.

Omnilingual ASR team et al. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025. URL https://arxiv.org/abs/2511.09690.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2212.04356.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peter Sullivan, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. Arab voices: Mapping standard and dialectal Arabic speech technology. *arXiv preprint arXiv:2601.13319*, 2026. URL https://arxiv.org/abs/2601.13319.

Bashar Talafha et al. Casablanca: Data and models for multidialectal Arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2410.04527.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023. URL https://arxiv.org/abs/2310.13289.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. Open universal Arabic ASR leaderboard. In *Proc. Interspeech*, 2025. URL https://arxiv.org/abs/2412.13788. arXiv:2412.13788.

Jin Xu et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. URL https://arxiv.org/abs/2503.20215.