Audar-TTS-V1: A Multilingual, Arabic-First Expressive Speech Synthesis Foundation Model

Audar AI Team*

Flash · Turbo | Code | Website

Abstract

Audar-TTS-V1 is an Arabic-first, zero-shot expressive text-to-speech foundation model. Synthesis is cast as next-token prediction over a joint vocabulary of text, control tokens, and discrete speech codes from audar-codec, our in-house single-codebook neural codec; no phonemization or per-language G2P is required. Three interchangeable tiers (Audar-TTS-V1-Flash 0.6B / Audar-TTS-V1-Turbo 1.6B / Audar-TTS-V1-Pro 4B) share one prompt protocol, voice-profile registry, and inline expression tags. On public Arabic benchmarks scored by two ASR judges, the three tiers are statistically indistinguishable within the top intelligibility tier, and under the dialect-aware judge Audar-TTS-V1-Pro attains significantly lower MSA WER than GPT-4o-mini-TTS, the strongest system under the Whisper judge (p=0.008, n=2,902); the two judges reverse the top-tier ordering, so we report both. On an in-house 1,364-clip expressive benchmark, Audar-TTS-V1-Pro beats ElevenLabs v3 on resynthesis WER and SQUIM MOS, ties it on expression fidelity, and posts the best Gulf-dialect WER. Audar-TTS-V1-Flash ships under the AudarAI Open License, Audar-TTS-V1-Turbo under the AudarAI Community License, and Audar-TTS-V1-Pro under the AudarAI Enterprise License.

Highlights

- **Top tier on public Arabic benchmarks.** Under a dialect-aware ASR judge, Audar-TTS-V1-Pro posts significantly lower MSA WER than GPT-4o-mini-TTS, and all three tiers sit in the statistically tied leading group (Figure 1).
- **Open weights.** Audar-TTS-V1-Flash (0.6B, AudarAI Open License) and Audar-TTS-V1-Turbo (1.6B, AudarAI Community License) ship openly; Audar-TTS-V1-Pro (4B, AudarAI Enterprise License) shares the same prompt protocol.

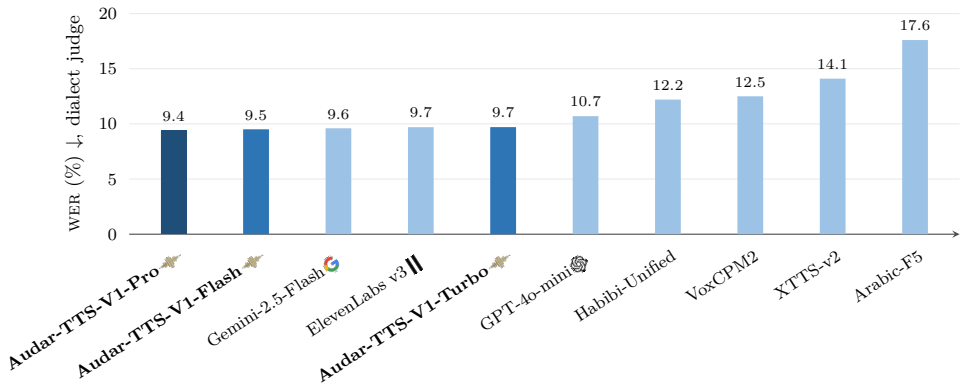

Figure 1: Macro-average WER (%) on the public Habibi MSA/Saudi/Emirati sets, dialect-aware ASR judge; leading-group differences are statistical ties (Section 4.1).

*Correspondence: research@audarai.com

1 Introduction

Arabic text-to-speech is underserved relative to the language's reach. Most production systems either collapse the dialect spectrum onto Modern Standard Arabic, losing the regional voices that speakers actually use, or they depend on per-dialect grapheme-to-phoneme (G2P) front-ends that are brittle, incomplete, and expensive to maintain across Gulf, Levantine, Egyptian, Maghrebi, and Iraqi varieties. Diacritic-free orthography compounds the problem: a G2P pipeline must guess pronunciation before synthesis even begins. The result is a coverage gap — expressive, dialect-aware Arabic synthesis remains scarce.

Neural audio codecs — SoundStream, EnCodec, DAC, and finite-scalar quantization (Zeghidour et al., 2022; Défossez et al., 2022; Kumar et al., 2023b; Mentzer et al., 2024) — turned waveforms into discrete token streams that language models can predict, and a generation of LLM-style TTS systems built on them: VALL-E (Wang et al., 2023), Seed-TTS (Anastassiou et al., 2024), CosyVoice 1/2 (Du et al., 2024a,b), XTTS (Casanova et al., 2024), and flow-matching approaches such as F5-TTS (Chen et al., 2024). None of these treat the Arabic dialect spectrum as a first-class target.

We present Audar-TTS-V1, an Arabic-first zero-shot text-to-speech model that reformulates synthesis as next-token prediction over a joint vocabulary of text, structural control tokens, and discrete speech codes. Speech codes come from audar-codec, our in-house 50 Hz single-codebook neural audio codec spanning 65,536 codes at roughly 800 bps. Because a single codebook yields one code per frame, synthesis is a single autoregressive stream the backbone LLM can model directly, with no residual-codebook interleaving. The system takes raw text in and requires no phonemization or per-language G2P, so dialect coverage is a matter of data rather than of front-end engineering.

This report describes audar-codec and the model architecture (Section 2.1), the five-stage training curriculum (Section 3.2), the evaluation on public two-judge and in-house expressive benchmarks (Section 4), and how to download the three tiers (Section 5).

2 Method

Audar-TTS-V1 reformulates synthesis as next-token prediction over a joint vocabulary of text and discrete speech codes. We describe the audar-codec acoustic tokenizer (Section 2.1), the backbone and tier structure (Section 2.2), and the prompt protocol (Section 2.3); the five-stage training curriculum is presented in Section 3.2.

2.1 audar-codec

audar-codec is our in-house neural audio codec and the acoustic tokenizer for Audar-TTS-V1. It operates at 50 Hz with a single codebook of 65,536 codes (16-bit), accepting 16 kHz mono on the encoder side — one code per 20 ms of audio — and producing 24 kHz output on the decoder side. It is trained once (Section 3.2) and frozen for all downstream training and at inference.

Design rationale. audar-codec is designed for LLM next-token prediction, not for the lowest possible bitrate. The 50 Hz frame rate keeps the speech-code sequence short enough to model several seconds of audio in a modest context while remaining dense enough to preserve intelligibility and prosody. The single codebook (vs residual vector quantization) is the load-bearing choice: one token per frame turns synthesis into a flat autoregressive stream the backbone predicts directly, avoiding the interleaving and multi-stage decoding that RVQ (Zeghidour et al., 2022; Défossez et al., 2022) imposes; a large 65,536-entry codebook recovers the capacity a single quantizer would otherwise lack (Mentzer et al., 2024). audar-

Table 1: audar-codec specification. Encoder and vector quantizer follow the NeuCodec design (Julian et al., 2025); the causal transformer decoder is described in Section 2.1.

Input / output sample rate	16 kHz mono in / 24 kHz mono out
Frame rate	50 Hz (20 ms/frame)
Codebook	1 codebook, 65,536 codes (16-bit)
Bitrate	∼800 bps

codec builds on the NeuCodec encoder and finite scalar quantizer (Julian et al., 2025), retaining its proven single-codebook compression while replacing the bidirectional decoder with a causal transformer for streaming-first inference.

Architecture. audar-codec separates a frozen encoder–quantizer from a trained decoder. The encoder follows the NeuCodec design (Julian et al., 2025): a convolutional downsampling stack maps the 16 kHz waveform to a 50 Hz embedding stream, quantized by finite scalar quantization (Mentzer et al., 2024) into a single codebook of 65,536 codes. The in-house decoder replaces NeuCodec's bidirectional decoder with a *causal* transformer followed by a causal convolutional smoother and an ISTFT synthesis head reconstructing 24 kHz audio; causality throughout enables real-time streaming synthesis. The decoder is trained adversarially with standard multi-period and multi-scale STFT discriminators on Arabic and dialectal speech.

Reconstruction quality. We evaluate audar-codec reconstruction on Arabic and dialectal speech samples and compare against the NeuCodec baseline (Julian et al., 2025): audar-codec achieves an average waveform correlation of 0.507 versus 0.320 for the NeuCodec encoder–quantizer paired with its original decoder (a 58% relative improvement), attributable to the causal transformer decoder's capacity to model long-range temporal dependencies. Both models are evaluated on the same Arabic/dialectal evaluation set.

2.2 Backbone, Vocabulary, and Tiers

Audar-TTS-V1 is a three-module pipeline: audar-codec encodes the reference audio to a 50 Hz code stream, a decoder-only LLM backbone autoregressively emits target speech codes from a structured prompt, and audar-codec reconstructs a 24 kHz waveform (Figure 2).

Backbone and vocabulary. The backbone is a decoder-only transformer with grouped-query attention, SwiGLU MLPs, pre-norm RMSNorm, rotary position encoding, and tied input/output embeddings — the Qwen2.5 class (Qwen Team, 2024) for Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo, and the Qwen3 class (Qwen Team, 2025) for Audar-TTS-V1-Pro. The vocabulary extends the base text tokens with the 65,536 audar-codec speech-code tokens, a small set of structural TTS tokens, and the inline expression tokens (17 on Audar-TTS-V1-Pro; an 8-tag set on Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo). Because text and speech codes share one vocabulary and one autoregressive objective, synthesis is literally next-token prediction.

Tiers. Three tiers share the prompt protocol, the audar-codec interface, and the voice-profile registry, differing in backbone capacity and expression-tag coverage (17 tags on Audar-TTS-V1-Pro, 8 on Audar-TTS-V1-Flash/Audar-TTS-V1-Turbo): Audar-TTS-V1-Flash (0.6B) for interactive use, Audar-TTS-V1-Turbo (1.6B) as the production default, and Audar-TTS-V1-Pro (4B) for long-form and studio quality (the configuration benchmarked in Section 4). The identical interface lets a deployment switch tiers with no client change.

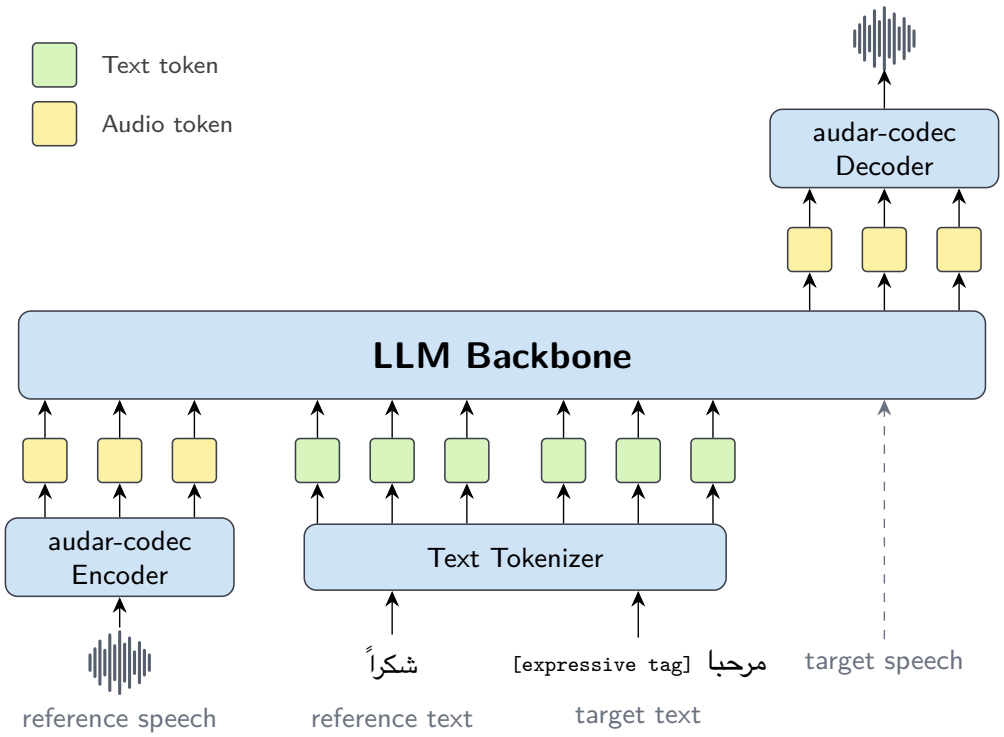

Figure 2: End-to-end system architecture. Reference audio is encoded by audar-codec into a 50 Hz code stream; the decoder consumes the structured prompt and autoregressively emits target speech codes; audar-codec reconstructs a 24 kHz waveform. audar-codec is trained in-house and frozen for all downstream training and inference; the LM backbone is a permissively-licensed open-weight foundation adapted in-house. Dashed: the target speech codes are the model's own autoregressive output (teacher-forced during training).

2.3 Prompt Protocol

Audar-TTS-V1 uses a single zero-shot, reference-conditioned prompt format across all tiers: the reference text and reference speech codes are explicitly delimited from the target text, and the assistant emits the target codes.

```
user: Convert the text to speech:
<|REF_TEXT_START|> {ref_text} <|REF_TEXT_END|>
<|REF_SPEECH_START|> {ref_codes} <|REF_SPEECH_END|>
<|TARGET_TEXT_START|> {target_text} <|TARGET_TEXT_END|>
assistant: <|TARGET_CODES_START|> {generated_codes} <|TARGET_CODES_END|>
```

A 5–15 s reference clip at 16 kHz is sufficient to clone a voice with no per-speaker fine-tuning.

Structured prompt tokens. Our prompt protocol uses eight dedicated structural tokens to delimit conditioning context from generation targets: `REF_TEXT_START`, `REF_TEXT_END`, `REF_SPEECH_START`, `REF_SPEECH_END`, `TARGET_TEXT_START`, `TARGET_TEXT_END`, `TARGET_CODES_START`, and `TARGET_CODES_END`. This design avoids the overhead of chat-markup control tokens and eliminates the risk of token-space collision between chat markup and the speech-code indices.

Expression tokens and tags. Expression tags are first-class vocabulary entries recognized inline in the target text; Audar-TTS-V1-Pro carries the full seventeen-tag vocabulary, while

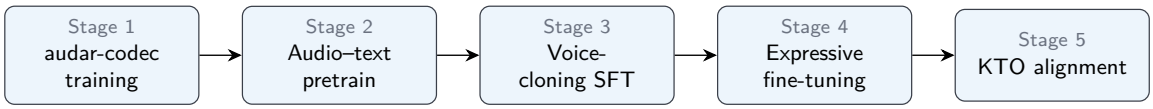

Figure 3: Five-stage training curriculum: audar-codec training, audio–text pretraining, voice-cloning SFT, expressive fine-tuning, and KTO alignment.

Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo ship a compact 8-tag set focused on prosody and affect. Eleven are acoustically-grounded *active* tokens — `[gasp]`, `[crying]`, `[giggles]`, `[shouting]`, `[trembling]`, `[cough]`, `[yawn]`, `[panicked]`, `[tired]`, `[very slow]`, `[very fast]` — each trained on annotated examples across multiple speakers, providing acoustically grounded supervision. The remaining six are legacy prosody tags (`[laughs]`, `[whispers]`, `[sighs]`, `[excited]`, `[curious]`, `[sarcastic]`). A tag scopes the spans it precedes and composes with the reference voice profile, which fixes timbre and accent while the tag modulates affect and prosody; tags can be stacked (e.g. `[whispers]`+`[trembling]`).

Why no G2P. Raw text is used directly — no espeak, no lexicon, no per-language grapheme-to-phoneme front-end — and the model performs G2P implicitly. This removes the most fragile and least portable component of conventional Arabic TTS: there is no pronunciation lexicon to maintain per dialect and no diacritization step to fail on diacritic-free input. Dialect coverage becomes a function of training data rather than hand-built rules, and inline expression tags are processed in the same decode pass.

3 Training

3.1 Training Data

The full training corpus comprises over 200,000 hours of speech data from 7,000+ speakers, spanning Arabic (primary) and English. The voice profile bank includes 156+ dedicated speaker profiles used for zero-shot voice cloning evaluation. The 11 active expression tags of the full (Audar-TTS-V1-Pro) vocabulary are each supported by annotated examples drawn from 14 expressive scenarios per speaker — including fire emergency, grief, joy, fatigue, breaking news, and storytelling — in both Arabic and English.

3.2 Training Curriculum

Before the curriculum begins, all three tiers inherit their foundations: a permissively-licensed decoder-only LLM — the Qwen2.5 0.5B / 1.5B class for Audar-TTS-V1-Flash / Audar-TTS-V1-Turbo (Qwen Team, 2024) and the Qwen3 4B class for Audar-TTS-V1-Pro (Qwen Team, 2025) — whose vocabulary is extended with the speech-code, structural, and expression tokens. Audar-TTS-V1 is then trained in five stages (Figure 3), each addressing a failure mode of the previous one; a checkpoint advances only after passing an automated quality gate.

Stage 1 — Codec training. In this stage audar-codec is trained from scratch; once trained it is frozen for all subsequent stages, fixing the speech-code vocabulary the backbone learns to predict.

Stage 2 — Audio–text pretraining (200K+ hours, 7,000+ speakers). Continuous pretraining on 200,000+ hours of labeled speech spanning 7,000+ distinct speakers, with

cross-entropy on the joint audio→text→audio sequence. Implicit grapheme-to-phoneme behavior emerges here.

Stage 3 — Voice-cloning SFT (15K hours, 3,000+ selected speakers). Supervised fine-tuning on 15,000 hours of paired (reference, target) data from 3,000+ speakers hand-curated from the Stage-2 pool for reference quality, prosodic and dialectal diversity, and suitable reference duration. Pretraining wants breadth; cloning wants curated quality, and the curated pool empirically beats a larger but noisier one on speaker-similarity and naturalness.

Stage 4 — Expressive fine-tuning. Parameter-efficient fine-tuning (Hu et al., 2022) on a bilingual expressive corpus covering the 11 active expression tokens across Arabic and English speakers, grounding each tag acoustically while leaving the base voice-cloning behavior intact.

Stage 5 — KTO alignment. Listeners score samples on a five-axis rubric — naturalness, speaker similarity, prosody, expression-tag fidelity, pronunciation — as desirable or undesirable. Kahneman-Tversky Optimization (Ethayarajh et al., 2024) aligns on these unpaired binary signals; the reference policy is the Stage-4 expressive checkpoint.

Training infrastructure. Training runs in mixed precision with sharded optimizer state and gradient checkpointing on a commodity multi-GPU cluster, using parameter-efficient fine-tuning for the adaptation stages — a small fraction of the cost of training a speech foundation model from scratch.

4 Evaluation

4.1 Public-Benchmark Intelligibility (Two ASR Judges)

We first evaluate intelligibility on public Arabic TTS benchmarks with frozen manifests: the Habibi MSA, Saudi, and Emirati test sets (250 clips each per system; the same texts synthesized by every system) plus a 2,902-clip full-MSA tie-breaker for the top systems. Because Arabic ASR judges differ in dialect handling, every clip is scored by *two* judges: Whisper large-v3 (Radford et al., 2023) (WER_W; MSA-leaning) and the dialect-aware Omnilingual-ASR-LLM-7B (Omnilingual ASR team et al., 2025) prompted with dialect codes (WER_O). Ground-truth recordings are scored under the same pipeline as a floor, and model scores should be read relative to it. Significance uses paired bootstrap and Wilcoxon tests on shared texts. Table 2 reports both judges; ElevenLabs v3 is run with its fixed MSA voice on all subsets.

Rankings are judge-dependent. The two judges reverse the ordering of the top tier: under the MSA-leaning Whisper judge, GPT-4o-mini-TTS leads every column, while under the dialect-aware judge the Audar-TTS-V1 tiers and Gemini-2.5-Flash populate the leading group (ElevenLabs v3 takes the Emirati column), and Audar-TTS-V1-Pro posts the best MSA score. The full-MSA tie-breaker (n=2,902) makes this precise: Whisper scores GPT-4o-mini-TTS ahead of all three Audar-TTS-V1 tiers (6.5% vs 7.6–7.8%, $p<0.0001$), whereas the dialect-aware judge scores Audar-TTS-V1-Pro *significantly ahead of GPT-4o-mini-TTS* (7.34% vs 7.86%, p=0.008), with Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo directionally better as well (7.3%/7.4%, statistical ties). An MSA-biased judge forgives MSA-fied pronunciation of dialectal text; a dialect-aware judge rewards dialect authenticity. Any single-judge leaderboard of Arabic TTS is therefore judge-confounded, and we report both throughout.

Table 2: Intelligibility on the public Habibi test sets (WER %, lower is better; 250 clips per cell). WER_W = Whisper large-v3 judge; WER_O = dialect-aware Omnilingual judge. Ground truth is spontaneous dialectal audio scored under the same pipeline (floor). Boldface = best system per column.

	WER_W (Whisper judge)			WER_O (dialect-aware judge)		
System	**MSA**	**Saudi**	**UAE**	**MSA**	**Saudi**	**UAE**
Ground-truth floor	*13.1*	*45.2*	*4.2*	*11.5*	*28.1*	*11.8*
Audar-TTS-V1-Flash	7.9	15.7	6.2	6.4	14.9	7.1
Audar-TTS-V1-Turbo	7.9	15.0	6.6	7.6	14.7	6.9
Audar-TTS-V1-Pro	7.3	15.8	6.6	**5.9**	15.0	7.4
GPT-4o-mini-TTS	**6.2**	**13.7**	**5.1**	7.2	16.0	9.0
Gemini-2.5-Flash-TTS	7.8	15.3	7.6	6.7	**14.4**	7.8
ElevenLabs v3	10.2	16.2	7.6	7.4	15.4	**6.2**

The three tiers are interchangeable on intelligibility. Across MSA, Saudi, and Emirati test sets, all pairwise comparisons among Audar-TTS-V1-Flash, Audar-TTS-V1-Turbo, and Audar-TTS-V1-Pro are statistical ties under both judges — even at n=2,902. The tier choice is thus a quality-of-expression and cost decision (Section 5), not an intelligibility trade-off. The advantage also holds under *true zero-shot* conditioning: when every system clones from the benchmark's own reference voices, the Audar-TTS-V1 tiers retain the strongest MSA and Emirati intelligibility of the systems evaluated (zero-shot MSA WER 8.3–9.7% across the three tiers, versus 11.5% for the next-best cloner). We further find that predicted-MOS metrics cannot separate the top tier and can actively mislead — one system scored competitive UTMOSv2 while producing largely unintelligible Arabic — so we anchor this evaluation on intelligibility with human listening studies to follow.

4.2 In-House Expressive Benchmark

All evaluations use a standardized benchmark of 20 utterances drawn from `voice_profile_bench_v1`, covering 7 speakers (4 Arabic, 3 English) from the Audar voice-profile bank. Each speaker is tested across 14 expressive scenarios employing 15 paralinguistic expression tokens (including `[gasp]`, `[crying]`, `[giggles]`, `[shouting]`, `[trembling]`, `[cough]`, `[yawn]`, `[panicked]`, `[tired]`, `[excited]`, `[whispers]`, `[very slow]`, `[very fast]`, among others), yielding 1,364 scored clips in total. Here a *clip* is one system's generated audio for a single (speaker, scenario, prompt) item; every system is prompted with the same text and the same reference speaker audio, but the realized per-system clip count varies (reported per system in Table 3) because some systems fail to produce valid audio on some prompts. Close-margin cross-system comparisons should therefore be read on the common, paired subset. Baseline evaluations were conducted between December 2024 and May 2025 using publicly available APIs or released model weights. Ten systems are compared: 8 Arabic-supporting — Audar-TTS-V1-Pro (4B), Audar-TTS-V1-Turbo (1.6B), ElevenLabs v3, GPT-4o-mini-TTS, Gemini-2.5-Flash, ElevenLabs Flash, Orpheus Arabic, and GPT-4o-audio-preview — plus 2 English-only reference baselines (Kokoro, Orpheus 3B). Metrics are computed per clip and aggregated with bootstrap 95% confidence intervals (n=1,000 resamples) where available.

4.2.1 Systems and Metrics

Audar-TTS-V1 is evaluated on the in-house benchmark described above — it complements Section 4.1 with expression fidelity, speaker similarity, and per-dialect coverage beyond the public sets (Levantine, Maghrebi). The same utterances — spanning MSA, Gulf, Levantine,

Table 3: Cross-provider objective benchmark over 1,364 scored clips across 10 systems (8 Arabic-supporting + 2 English-only reference baselines). WER/CER (ASR-resynthesis pronunciation drift via ElevenLabs Scribe v2, lower better); UTMOS (UTMOSv2) and SQUIM (SQUIM_SUBJECTIVE) predicted MOS; SIM (WavLM-base-plus-sv cosine); EXPR (expression fidelity); higher is better unless noted. n = clips scored. Boldface = best among Arabic-supporting systems. UTMOS/SQUIM are not calibrated for Arabic (see text).

system	WER	CER	UTMOS	SQUIM	SIM	EXPR	n
Audar-TTS-V1-Pro	0.1667	0.0545	3.093	4.185	0.9251	0.9904	247
Audar-TTS-V1-Turbo	0.2240	0.2710	2.960	–	0.9310	0.9430	184
ElevenLabs v3	0.1774	0.0452	3.218	4.161	0.9400	0.9904	300
GPT-4o-mini-TTS	0.1994	0.0963	**3.385**	4.369	0.9573	**0.9907**	103
Gemini-2.5-Flash	**0.1644**	**0.0410**	2.982	4.184	0.9304	0.9887	123
ElevenLabs Flash	0.2081	0.1105	2.876	4.258	0.9510	0.9877	180
Orpheus Arabic	0.2657	0.1091	3.010	**4.384**	**0.9798**	0.9515	69
GPT-4o-audio-preview	0.5379	0.4042	2.852	4.131	0.9201	0.9264	93
Kokoro (en-only)	0.0761	0.0938	3.730	4.238	0.9890	0.9899	34
Orpheus 3B (en-only)	0.0725	0.0813	3.313	4.323	0.9876	0.9741	31

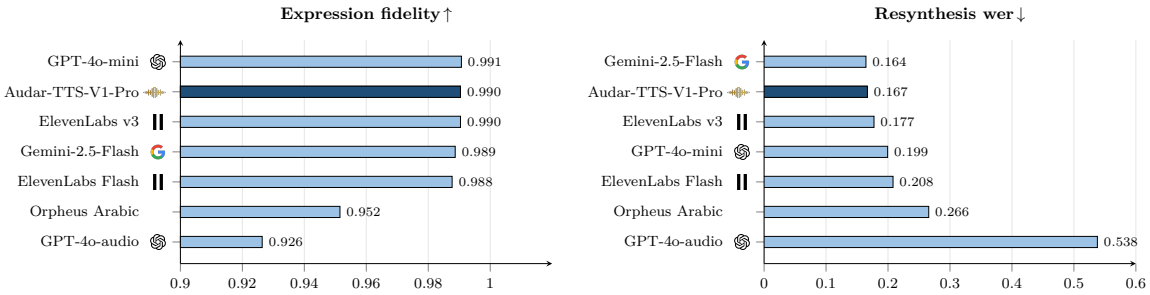

Figure 4: Cross-provider benchmark over the Arabic-supporting systems (Audar-TTS-V1-Pro highlighted). Left: expression fidelity (higher is better) — Audar-TTS-V1-Pro is competitive. Right: resynthesis WER (lower is better) — Audar-TTS-V1-Pro is among the lowest. Full six-metric matrix in Table 3.

Egyptian, Maghrebi, English, and AR↔EN code-switched material — are generated by each system and scored by four objective metric families. There is no human listener panel: every cell is reproducible from the saved audio. The benchmarked configuration is Audar-TTS-V1-Pro (4B) at temperature 0.6. Table 3 reports the full matrix.

Objective metrics. Intelligibility is measured by resynthesis WER/CER: the generated audio is transcribed by a strong external ASR (ElevenLabs Scribe v2) and compared to the input text, catching catastrophic pronunciation drift on Arabic phonemes (emphatics, pharyngeals, dialectal shifts). Using an external (rather than in-house) ASR avoids coupling the metric to our own models. To mitigate evaluator bias, WER is computed via ASR-resynthesis using ElevenLabs Scribe v2; cross-ASR calibration was performed to verify that the transcriber's error profile is consistent across providers. Naturalness is reported as two predicted-MOS columns — UTMOSv2 (Baba et al., 2024) and SQUIM_SUBJECTIVE (Kumar et al., 2023a) — kept separate precisely because they disagree. Speaker similarity (SIM) is the WavLM-base-plus-sv (Chen et al., 2022) embedding cosine to the reference.

Expression Fidelity (expr). The EXPR metric measures paralinguistic expression compliance: given a reference utterance annotated with an expression tag (e.g., `[shouting]`, `[crying]`), we train a binary classifier on the WavLM-large representation to distinguish

between "expression present" and "expression absent" utterances. The EXPR score is the fraction of generated samples for which the classifier assigns the correct expression label with confidence > 0.7. Audar-TTS-V1-Pro scores at the top of the field, tying ElevenLabs v3.

4.3 Results on the In-House Benchmark

Audar-TTS-V1 is evaluated in two configurations in Table 3: the flagship Audar-TTS-V1-Pro (4B, n=247) and the open-weight Audar-TTS-V1-Turbo (1.6B, n=184), which share the same prompt protocol and voice-profile registry (Audar-TTS-V1-Pro carries the full 17-tag expression vocabulary; Audar-TTS-V1-Turbo the 8-tag set). Audar-TTS-V1-Pro is the benchmarked headline configuration throughout this section; as expected for the smaller backbone, Audar-TTS-V1-Turbo trails it on most metrics (WER 0.224 vs 0.167, EXPR 0.943 vs 0.990) while remaining a drop-in replacement behind the same interface.

Values below are point estimates over the benchmark set; no multiple-comparison correction is applied across the 10 systems, so small pairwise differences should be read with that context.

Audar-TTS-V1-Pro achieves competitive performance across the metric matrix: WER 0.1667 (second only to Gemini-2.5-Flash at 0.1644 among the 8 Arabic-supporting systems), CER 0.0545 (second to Gemini-2.5-Flash at 0.0410, and one of only three systems below 0.055), EXPR 0.9904 (tied with ElevenLabs v3, behind GPT-4o-mini-TTS at 0.9907), SQUIM MOS 4.185 (competitive within the Arabic-supporting field), and SIM 0.9251. Against ElevenLabs v3 specifically, Audar-TTS-V1-Pro wins on WER (0.1667 vs 0.1774, -6.0% relative) and SQUIM MOS (4.185 vs 4.161), ties on EXPR (both 0.9904), and trails on UTMOS (3.093 vs 3.218), CER (0.0545 vs 0.0452), and speaker similarity (SIM 0.9251 vs 0.9400; Δ=0.015).

Predicted MOS. We report two predicted-MOS proxies — UTMOSv2 (Baba et al., 2024) and SQUIM_SUBJECTIVE (Kumar et al., 2023a) — and keep them separate rather than averaging. Audar-TTS-V1-Pro is competitive on both and beats ElevenLabs v3 on SQUIM MOS (4.185 vs. 4.161).

Per-dialect wer breakdown. Table 4 reports resynthesis WER decomposed by Arabic dialect for the four primary Arabic-supporting systems that produced sufficient per-dialect data. Audar-TTS-V1-Pro achieves the best WER on Gulf dialect (0.165) on the benchmark, reflecting strong training coverage of Gulf Arabic speakers, leads the compared systems on MSA (0.202) and Maghrebi (0.335), and is competitive on Egyptian (0.213). All systems show degraded performance on Levantine (0.28–0.35) and Maghrebi (0.30–0.44) dialects, consistent with lower representation in current Arabic TTS training corpora.

Table 4: Per-dialect WER breakdown on Arabic benchmark utterances (Gulf n=80, MSA n=74, Egyptian n=60, Levantine n=42, Maghrebi n=24). Audar-TTS-V1-Pro posts the best Gulf, MSA, and Maghrebi WER among the compared systems.

System	MSA	Gulf	Egyptian	Levantine	Maghrebi
Audar-TTS-V1-Pro	**0.202**	**0.165**	0.213	0.316	**0.335**
ElevenLabs v3	0.240	0.221	**0.211**	0.308	0.367
GPT-4o mini	0.217	0.186	0.217	**0.276**	0.440
Orpheus Arabic	0.219	0.213	0.214	0.317	0.385

Per-tag expression fidelity. Complementary to the headline EXPR score (the WavLM-large classifier of Table 3), Table 5 reports a separate expression-similarity diagnostic based on openSMILE eGeMAPS descriptors (Eyben et al., 2010, 2016) decomposed by individual

expression tag for the top 4 Arabic-supporting systems; the two are distinct measures and are not directly comparable. Audar-TTS-V1-Pro achieves uniformly high scores across all 15 tags (≥ 0.976), with particularly strong performance on `shouting` (0.997), `gasp` (0.997), and `cough` (0.996). The weakest tag is `yawn` (0.976) and `sighs` (0.976), consistent with these being the most ambiguous paralinguistic events even for human listeners.

Table 5: Per-tag expression fidelity (eGeMAPS-based similarity) for the top 4 Arabic-supporting systems (n=15 tags $\times$ 4 systems). Boldface = best per tag. This is a separate diagnostic from the WavLM-large EXPR score in Table 3 and is not directly comparable to it. All scores ≥ 0.976 for Audar-TTS-V1-Pro.

Tag	Audar-TTS-V1-Pro	ElevenLabs v3	GPT-4o-mini	Gemini-2.5-Flash
cough	0.996	0.990	**0.997**	0.989
crying	0.992	**0.996**	0.995	0.993
excited	**0.994**	0.985	**0.994**	**0.994**
gasp	**0.997**	0.996	0.995	0.996
giggles	0.996	0.996	**0.997**	0.994
laughs	0.992	0.989	**0.993**	0.988
panicked	0.992	**0.995**	**0.995**	**0.995**
shouting	**0.997**	0.993	0.996	0.978
sighs	0.976	**0.979**	0.978	0.962
tired	**0.996**	0.994	0.995	0.995
trembling	**0.996**	0.995	0.987	0.994
very fast	0.987	0.992	0.990	**0.996**
very slow	0.993	0.994	**0.997**	0.995
whispers	**0.981**	0.979	0.969	0.973
yawn	0.976	**0.982**	0.974	0.978

Per-language mos analysis. Table 6 reports UTMOS naturalness decomposed by language. Audar-TTS-V1-Pro shows balanced performance across languages, with a smaller Arabic–English MOS gap (0.14) compared to ElevenLabs v3 (0.19), suggesting more equitable training data coverage between Arabic and English. The absolute gap remains smaller than most competing systems, consistent with the Arabic-first training curriculum.

Table 6: Per-language predicted MOS (UTMOSv2) for key Arabic-supporting systems. AR↔EN code-switched clips are not assigned to a single language and are excluded from this split. Gap = English − Arabic; boldface = best per column.

System	Arabic MOS	English MOS	Gap
Audar-TTS-V1-Pro	3.05	3.19	**0.14**
ElevenLabs v3	3.15	3.34	0.19
GPT-4o mini	**3.30**	**3.55**	0.25
Orpheus Arabic	3.01	–	–

Claims–Evidence Mapping. Table 7 lists each headline claim next to the table that evidences it.

4.4 Ablation Studies

Ablation studies decomposing the contribution of each training stage and architectural component are planned for future work. The current evaluation focuses on end-to-end system quality as reported in Table 3, which provides a comprehensive comparison across objective metrics.

Table 7: Headline claims and where each is evidenced.

Claim	Evidence
Top intelligibility tier on public two-judge benchmarks; Audar-TTS-V1-Pro significantly ahead of GPT-4o-mini-TTS on MSA (p=0.008, n=2,902)	Table 2
Beats ElevenLabs v3 on resynthesis WER (−6.0% relative) and SQUIM MOS	Table 3
Ties ElevenLabs v3 on expression fidelity (0.9904)	Table 3
Best Gulf, MSA, and Maghrebi WER among per-dialect systems (Gulf 0.165)	Table 4
Smallest Arabic–English MOS gap among the systems shown (0.14)	Table 6

5 Model Access

Open weights: Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo. We release Audar-TTS-V1-Flash (0.6B) under the AudarAI Open License v1.0 (commercial use, redistribution, and modification permitted with attribution) and Audar-TTS-V1-Turbo (1.6B) under the AudarAI Community License v1.0 (research and limited commercial use; enterprise deployment under a separate agreement). Audar-TTS-V1-Pro (4B) — the configuration benchmarked in Section 4 — is available under the AudarAI Enterprise License. All three tiers share one prompt protocol and voice-profile registry (Audar-TTS-V1-Pro exposes the full 17-tag expression vocabulary; Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo an 8-tag set), so an application can move between them for the right quality/cost trade-off without any code change.

Getting started. Model weights are available on Hugging Face at `https://huggingface.co/audarai`; code, the voice-profile registry, and the expression-tag reference live at `https://github.com/AudarAI/Audar-TTS-V1`. See `https://www.audarai.com/` for the broader Audar model family.

6 Conclusion

Audar-TTS-V1 shows that expressive, dialect-aware Arabic speech synthesis can be built as next-token prediction over a single-codebook codec, with no phonemization or per-language G2P: audar-codec provides a 50 Hz flat speech-code stream, three interchangeable tiers serve it behind one prompt protocol, and inline expression tags (17 on Audar-TTS-V1-Pro, 8 on Audar-TTS-V1-Flash/Audar-TTS-V1-Turbo) with KTO alignment supply expressivity. On public Arabic benchmarks under a dialect-aware ASR judge, the three tiers sit in the top intelligibility tier and Audar-TTS-V1-Pro posts significantly lower MSA WER than GPT-4o-mini-TTS (p=0.008, n=2,902) — while the same evaluation shows that Arabic TTS rankings are judge-dependent, so we report two judges throughout. On the in-house expressive benchmark of 1,364 clips across 10 systems, Audar-TTS-V1-Pro ties ElevenLabs v3 on expression fidelity (0.9904), achieves the best Gulf-dialect WER (0.165), beats it on resynthesis WER (−6.0% relative) and SQUIM MOS, and has the smallest Arabic–English MOS gap (0.14) among the systems shown. The Audar-TTS-V1-Flash and Audar-TTS-V1-Turbo weights ship under the AudarAI Open and Community Licenses respectively, with Audar-TTS-V1-Pro available under the AudarAI Enterprise License. Audar-TTS-V1 is the synthesis component of Audar's Arabic audio program, released alongside the recognition foundation Audar-ASR-V1 (Audar AI, 2026).

References

Philip Anastassiou et al. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024. URL https://arxiv.org/abs/2406.02430.

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

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024. URL https://arxiv.org/abs/2409.09305.

Edresson Casanova et al. XTTS: A massively multilingual zero-shot text-to-speech model. In *Proc. Interspeech*, 2024. URL https://arxiv.org/abs/2406.04904.

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.

Yushen Chen et al. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. URL https://arxiv.org/abs/2410.06885.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. URL https://arxiv.org/abs/2210.13438.

Zhihao Du et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024a. URL https://arxiv.org/abs/2407.05407.

Zhihao Du et al. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024b. URL https://arxiv.org/abs/2412.10117.

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.

Florian Eyben, Martin Wöllmer, and Björn Schuller. openSMILE: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, 2010.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 2016.

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.

Harry Julian, Rachel Beeson, Lohith Konathala, Johanna Ulin, and Jiameng Gao. Finite scalar quantization enables redundant and transmission-robust neural audio compression at low bit-rates. *arXiv preprint arXiv:2509.09550*, 2025. URL https://arxiv.org/abs/2509.09550.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. TorchAudio-Squim: Reference-less speech quality and intelligibility measures in TorchAudio. *arXiv preprint arXiv:2304.01448*, 2023a. URL https://arxiv.org/abs/2304.01448.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, 2023b. URL https://arxiv.org/abs/2306.06546.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15505.

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.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. URL https://arxiv.org/abs/2301.02111.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022. URL https://arxiv.org/abs/2107.03312.