MENASpeechBank: A Reference Voice Bank with Persona-Conditioned Multi-Turn Conversations for SpeechLLMs

Anonymous ACL submission

Abstract

Audio large language models (AudioLLMs) enable instruction following over speech and general audio, but progress is limited by the scarcity of *diverse, conversational, and instruction-aligned* speech–text data. This gap is particularly pronounced for persona-grounded and dialectal interactions, where collecting real multi-speaker recordings remains costly and slow. We introduce MENASPEECHBANK, a reference speech bank comprising ∼**18K high-quality utterances** from **124 speakers** spanning multiple MENA countries, covering English, Modern Standard Arabic (MSA), and regional Arabic varieties. We develop a controllable data pipeline that *(i)* constructs persona profiles enriched with World Values Survey (WVS) inspired attributes, *(ii)* defines a taxonomy driven $\sim 5K$conversational scenarios, *(iii)* matches personas to scenarios via semantic similarity, *(iv)* generates ∼417K role-play conversations with an LLM where the user speaks as the persona and the assistant behaves as a helpful agent, and *(v)* produces speaker-conditioned user-turn audio (synthetic) from reference recordings to preserve speaker diversity. We evaluate synthetic and human recorded conversations and provide an analysis. We will make the MENASPEECHBANK available for the community.[1]

1 Introduction

Speech is the dominant interface for human communication, yet most large language models are still optimized for text. Recent Audio or Speech LLMs move beyond cascaded "ASR → LLM" systems by directly conditioning instruction following on audio inputs, enabling richer interaction patterns such as multi-turn spoken dialogue,

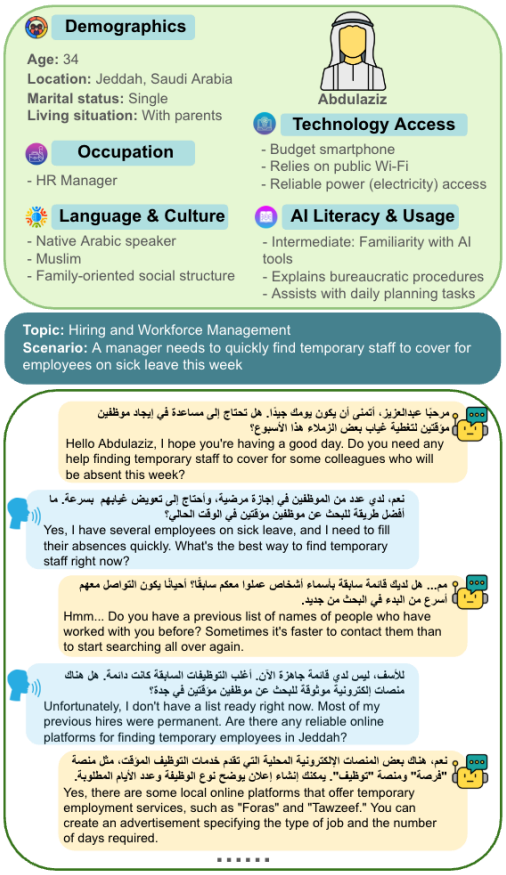

Figure 1: An example of a persona, a conversational scenario, and conversational turns. Persona attributes are derived from basic demographic information, country-specific WVS values, and heuristics.

speech-grounded reasoning, and paralinguistic understanding. Representative systems include Qwen-Audio (Chu et al., 2023), SpeechGPT (Zhang et al., 2023), AudioGPT (Huang et al., 2024), and SALMONN (Tang et al., 2024). In parallel, evaluation suites such as AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025a) increasingly emphasize open-ended, instruction-following behavior conditioned on diverse audio, surfacing persistent limitations in robustness, domain coverage, and conversational grounding. A central bottleneck is *data*. Compared to text-only instruction tuning,

[1] anonymous.com

SpeechLLMs (Wang et al., 2025c) require aligned audio–text pairs, and they benefit disproportionately from *(i) conversational structure* (multi-turn context), *(ii) task and intent diversity* (e.g., QA, classification, reasoning, dialogue acts), and *(iii) speaker and dialect diversity* (accent, dialect, sociolect, bilingual/code-switched usage, and channel conditions). These desiderata are difficult to satisfy with real recordings alone: collecting persona-rich, long-form conversations is labor-intensive; obtaining broad dialect and speaker coverage is expensive and often infeasible for low-resource varieties; and privacy/licensing constraints frequently limit the release and reuse of speech data. This bottleneck is particularly pronounced in the Middle East and North Africa (MENA) region, where dialect diversity is high and resources remain limited.

Synthetic data pipelines are an attractive alternative, but their impact on SpeechLLMs is still under-explored. In text-only LLMs, self-generated instruction corpora can align models at scale (e.g., Self-Instruct (Wang et al., 2023)). Persona conditioning has likewise been used to promote consistent dialogue (e.g., PersonaChat (Zhang et al., 2018)), and LLM-based generation with critics can improve persona faithfulness (Jandaghi et al., 2024). In parallel, high-fidelity multi-speaker TTS and voice cloning now enable controllable, speaker-consistent synthesis (Wang et al., 2025b; Le et al., 2023), and combining LLM-generated text with TTS has begun to support speech-language instruction corpora (Noroozi et al., 2024). However, it remains unclear *when* persona-grounded synthetic speech improves SpeechLLM performance, *which* data factors drive gains (faithfulness, speaker/dialect diversity, synthesis quality, filtering), and *how* to design end-to-end pipelines for downstream conversational and speech understanding tasks.

We address this gap with MENASPEECHBANK (*MenaSpeechBank*), a data-centric study that pairs a *multi-speaker MENA reference voice bank* with an *end-to-end persona-to-speech generation pipeline* for SpeechLLM adaptation. Using speaker profiles as anchors, we construct controllable personas by completing attributes with World Values Survey (WVS) inspired values (Haerpfer et al., 2020), design a taxonomy driven conversational scenarios, and assign scenarios to personas via semantic similarity in an embedding space (Reimers and Gurevych, 2019). We then generate role-play dialogues with an LLM (persona as user; assistant as helpful agent), synthesize user turns from reference voices, and fine-tune an LLM on the resulting audio–text instruction pairs to evaluate gains on scenario-driven conversation and spoken QA. Our work makes the following key contributions:

- We propose *MenaSpeechBank*, a multi-speaker MENA reference voice bank (∼**18K utterances; 124 speakers**) covering **English and Arabic variants**.
- We enrich speaker profiles with WVS inspired value attributes to build controllable, behaviorally grounded personas.
- We propose an end-to-end persona → dialogue → speech pipeline using LLM role-play generation and speaker-conditioned voice cloning to produce scalable audio–text instruction data.
- We benchmark on synthetic and human recorded conversations.

2 Related Work

Speech, Audio LLMs. A growing line of work equips LLMs with audio encoders and bridging modules to enable audio-conditioned instruction following. Examples include Qwen-Audio (Chu et al., 2023), SpeechGPT (Zhang et al., 2023), AudioGPT (Huang et al., 2024), and SALMONN (Tang et al., 2024); related efforts such as Spectron (Nachmani et al., 2024) adapt LLMs for spoken question answering and speech generation using spectrogram-based representations. Benchmarking has evolved alongside these models. AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025a) evaluate AudioLLMs under open-ended, chat-style instruction following across diverse audio types and task families, highlighting gaps in robustness and coverage and motivating data-centric improvements.

Synthetic Instruction Data. Synthetic instruction tuning has been central to aligning text-only LLMs. Self-Instruct (Wang et al., 2023) demonstrates that models can bootstrap large instruction corpora via iterative prompting and filtering. For dialogue, persona conditioning provides a lightweight mechanism for consistency and engagement. PersonaChat (Zhang et al., 2018) formalizes persona-grounded conversation and remains a standard testbed. More recently, LLMs have been used to generate persona-based conversational datasets at scale with explicit controls for adherence and faithfulness. Jandaghi et al. (2024) highlight both the benefits of synthetic persona data and

the importance of careful filtering. *MenaSpeechBank* builds on these ideas, however, differs in two ways: *(i)* we treat persona dialogue generation as *scenario-conditioned* instruction generation for speech-centric evaluation, and *(ii)* we render persona conversations into *speech* with controllable speaker/dialect factors to study their effect on SpeechLLMs.

Synthetic Speech and Training Data. Generative speech models now support high-quality multi-speaker TTS and voice cloning. Neural codec language models enable prompt-based voice cloning through discrete codec tokens (Wang et al., 2025b), while Voicebox extends multilingual generation and speech editing (Le et al., 2023). Synthetic speech has also been used to create instruction data for adapting speech-language models with less reliance on labeled real audio (Noroozi et al., 2024). However, the role of *persona-grounded conversational speech* in SpeechLLM adaptation remains underexplored, especially for SpokenQA and multi-turn dialogue where semantic understanding and speaker variation both matter. We study this setting through controlled speaker-conditioned data generation and conversational setups.

3 MenaSpeechBank

In Figure 2, we present the *MenaSpeechBank* dataset development pipeline, which we describe in the following subsections.

3.1 Reference Audio Collection

We developed a reference speech bank with audio samples in MSA, dialectal arabic, and English for neural voice cloning based speech generation.

Data Sources. MSA reference audio was collected from different sources: QASR dataset (Mubarak et al., 2021), and from *in-house dataset*. In addition to MSA, dialectal Arabic reference audio was collected from the *in-house set*, and the ADI-17 dataset (Shon et al., 2020), covering four major dialect regions: Egyptian, North African, Gulf, and Levantine Arabic. Finally, English reference audio was sourced from the English recordings available in the in-house dataset.

Filtering & selection. To ensure high audio quality and accurate text alignment, we applied dataset-specific filtering steps. For datasets with available reference scripts (our in-house dataset), we computed the word error rate (WER) between automatically generated transcriptions and the original

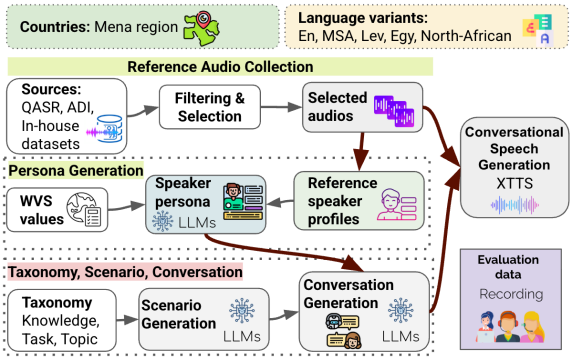

Figure 2: An overview of *MenaSpeechBank* development pipeline.

text. Transcriptions were produced using the Fanar Transcription tool (Aura)[2] for Arabic speech and Whisper-Small[3] for English speech. We chose Fanar due to its free access and competitive performance compared to other state-of-the-art systems, as reported in (Alam et al., 2025). We retained only samples with minimal transcription errors (WER ≤ 0.05). For the QASR dataset, we randomly selected 10 segments per speaker, each with a duration between 5 to 8 seconds, and used the automatic dialect identification tool Tamyiz (Abdullah et al., 2025) to verify that all selected samples were in MSA. For the ADI-17 dataset, we manually inspected all segments and selected 56 high-quality samples from five speakers, ensuring that each sample contained speech from a single speaker. These steps ensure that the reference audios are clean, well-aligned with text, and consistent in language variety.

3.2 Persona Generation

For persona profile generation, we curated basic demographic information whenever it was available in the dataset. In addition, we used country-specific WVS values, simple heuristics, and LLMs. Below, we discuss these components in detail.

Reference speaker profile. For the reference speaker profile, in some sources we had country, gender, age and educational information. Additionally for the paid speakers who recorded the audio and signed NDA for all permitted use of the data we asked them to provide basic demographic information. However, we do not ask or collect any personally identifiable information. Not for all speakers all of these information were available. For such cases we have derived them based on some heuristics (see Appendix B.1 for details).

[2] https://api.fanar.qa/docs
[3] https://huggingface.co/openai/whisper-small

WVS value grounding. We have collected country specific WVS values (Wave 7)[4] (EVS/WVS, 2022)to include additional or missing attributes with the persona. We use country-level aggregates and cultural dimension summaries (e.g., traditional vs. secular-rational; survival vs. self-expression) to capture cross-country value differences. WVS value information is matched by matching speaker country and adjusted based on the age information. More details on WVS incorporation are proivded in Section B.2.

Speaker persona profile. We construct persona profiles by combining *(i)* basic metadata from reference speaker profiles, *(ii)* WVS inspired value attributes, and *(iii)* lightweight heuristics capturing technology access, AI literacy, and AI usage. Concretely, each persona is instantiated by sampling a country- and gender-consistent *name* from curated name lists and a *city* uniformly from a country-specific city inventory. We sample *profession* uniformly from an age-bucketed occupation list, and for synthetic personas we draw *age* $\in [18, 40]$ and *gender* $\in \{\text{male}, \text{female}\}$, then sample *marital status* and *household type* uniformly from predefined categorical sets. To model digital and AI-related variation, we independently sample *device* and *connectivity* from fixed categorical inventories and assign an *AI-competence level* from an ordinal set. *AI use cases* are selected as a uniform sample of k=2 distinct items from a predefined list. Finally, we generate a continuous OCEAN (Big Five) personality vector by perturbing fixed base values.

We use the original 124 speaker profiles as *seeds* and expand them to create additional personas. To mitigate near-duplicates, we compute pairwise semantic similarity between persona descriptions using cosine similarity over multilingual Sentence-BERT embeddings (Reimers and Gurevych, 2019). We discard any candidate whose similarity to an existing profile exceeds 0.80, resulting in a final set of 469 personas.

Persona summary. Although personas are stored as structured attribute profiles, we convert each profile into a concise first-person *persona summary* to better support LLM role-play generation and quality control. Persona-grounded dialogue work shows that conditioning on explicit persona descriptions improves consistency and personalization (Li et al., 2016; Zhang et al., 2018). In our pipeline, the summary provides a compact interface for prompting and critic-based faithfulness checks (Jandaghi et al., 2024), and a single text representation for embedding-based persona scenario matching (Reimers and Gurevych, 2019). We generate summaries with GPT-4.1 using profile attributes; Appendix Figure 5 lists the full prompt.

Persona quality. We automatically assess persona summaries using a deterministic, rule-based *Persona Quality Index* (PQI) that counts satisfied checks over *(i)* static attribute consistency with the persona profile (e.g., no unsupported numerals/proper nouns; no explicit mention of value terminology) and *(ii)* narrative compliance (e.g., first-person voice, 90–180 words, grounded opening, single-paragraph prose, and a grounded ending). A summary is considered fully compliant only if it passes all checks (PQI $= 12$). On $n = 469$ personas, 86.6% achieve the maximum score, while the remaining cases are primarily attributable to a conservative *consistency* flagging heuristic (11.1%); full check definitions are provided in the Section C.2.

3.3 Taxonomy, Scenario and Conversations

Taxonomy. For taxonomy development, we focused on *everyday* user intents and information needs. We began with *(i)* our predefined task/service domains (e.g., public sector & utilities; travel & mobility) and *(ii)* knowledge-oriented topic categories drawn from (Alam et al., 2026) and the everyday topics defined in (Hasan et al., 2025). We harmonized overlapping labels, merged near-duplicates, and organized the result into a hierarchical taxonomy with two top-level branches: *(i)* task/service domains and *(ii)* knowledge/topic domains. Overall, the taxonomy is designed to reflect the *service- and tool-mediated* interactions that dominate modern task-oriented conversational assistants, while maintaining broad coverage of major industries (e.g., finance, commerce, travel/-mobility, health, public services, and productivity). The complete domain hierarchy is shown in Figure 6. This design choice is consistent with established multi-domain dialogue benchmarks created to approximate real assistant behavior (Byrne et al., 2019; Rastogi et al., 2020).

Scenario. Although the taxonomy provides struc-

[4]World Values Survey (WVS) values capture population-level attitudes and beliefs about major aspects of social life, such as religiosity, family and gender norms, trust in others and institutions, political preferences and support for democracy, tolerance toward out-groups, work and economic views, and subjective well-being.

tured domain coverage, its leaf nodes remain too coarse to consistently elicit diverse, realistic conversations. Prompting directly from a leaf-domain label often produces repetitive or overly generic interactions, leaving *common user intents* underspecified. We therefore introduce an intermediate *topic* layer that operationalizes each root-to-leaf domain path. Specifically, for each domain path d, we use an LLM to generate $N = 10$ topics (Listing 5). We then generate scenarios for each domain path-topic pair (d, t) as $\mathcal{S}_{d,t} \leftarrow \text{LLM}_\theta(d, t; N)$, where $\mathcal{S}_{d,t} = \{s_i\}_{i=1}^{N}$ and $N = 10$. In Listing 6, we report the prompt. This intermediate expansion improves within-domain diversity and coverage prior to full conversation generation, consistent with observations in prior work (Rebedea et al., 2024). In this step we also applied near duplicate filtering following the same similarity based approach mentioned in Section 3.2 with a threshold of 0.85. Using this procedure, we obtained 900 distinct topics and 4,521 scenarios.

Conversations. Because personas differ in roles, goals, and constraints, they are typically most plausible in different domain scenarios. Therefore, we align each persona with the scenario(s) that best match its summary. For each persona–scenario pair, we compute a hybrid similarity score that combines semantic similarity (cosine similarity of sentence-transformer embeddings; `sentence-transformers/all-MiniLM-L6-v2`)[5] with lexical similarity (token-overlap Jaccard similarity), and retain only pairs with a score of at least 0.05. This procedure yields scalable and reproducible persona–scenario pairings, ensuring that personas are grounded in an explicit context.

After matching personas to scenarios, we generate persona-profiled assistant-user conversations conditioned on the selected persona attributes and scenario context. To streamline this process, we use GPT-4.1 with a structured prompt that instantiates the persona and scenario for each dialogue. The full prompt is provided in Listings 7 and 8 .

3.4 Conversational Speech Generation

Following persona—scenario-–conditioned dialogue generation, we use neural voice closing approach to generate speech for the MSA conversations using the XTTS-v2 (Casanova et al., 2024)[6] model. In this study, we focus on MSA because dialectal support in Arabic voice-cloning systems was limited at the time of our experiments; extending the pipeline to Arabic dialects is left for future work. We generated audio for user turns only, while assistant turns are kept as text. For each conversation, a single reference audio corresponding to the persona's speaker profile is used. In total, we generate ∼2.1M audio samples across 416K conversations, Table 3 summarizes the corresponding conversation statistics.

Quality assessment. To assess audio quality, we randomly select 100 synthesized samples per speaker and evaluate them using standard objective metrics: *(i)* Word Error Rate (WER) to measure transcription accuracy, computed using Fanar Aura; *(ii)* Speaker Cosine Similarity (SpkCos) (Ravanelli et al., 2021) to assess speaker consistency, based on embeddings from the `spkrec-ecapa-voxceleb` model; and *(iii)* NISQA (Mittag et al., 2021), which predicts overall perceptual speech quality.

The generated speech exhibits a moderate word error rate (WER = 10%), while maintaining good perceptual quality (NISQA = 3.6) and moderate speaker consistency across conversations (SpkCos = 0.49). To better interpret the SpkCos similarity score, we compare it agains two baselines using the same speaker encoder: same-speaker human-human similarty accross different recordings, as an empirical upper bound under our setup, and different-speaker similarity. Our results show that the synthetic-human speaker similarity (0.49) is substantially higher than different-speaker similarity (0.12) and is closer to the natural same-speaker similarity (0.57), indicating that generated audios reflect natural within-speaker variations and are separated from different-speaker pairs.

3.5 Dataset Statistics

Reference speakers. *MenaSpeechBank* covers a wide range of speakers, comprising *124 unique speakers* from 18 different countries. Tables 1 and 3 offers a detailed breakdown of the number of speakers, utterances by mother tongue and conversation statistics. Of these, 56% are male, and the dominant age group is 20—29 years, with 50 speakers in this range (see details in Appendix D).

3.6 Speakers Statistics

In Table 2, we summarizes the number of speakers for each language or dialect, together with the average number of utterances per speaker and the average utterance duration. The speaker counts

[5] all-MiniLM-L6-v2
[6] `https://huggingface.co/coqui/XTTS-v2`

Mother tongue	Spk.	Utt.	Mother tongue	Spk.	Utt.
Arabic (Egypt)	31	6,440	Arabic (Jordan)	5	690
Arabic (Iraq)	11	1,861	Urdu	2	694
Arabic (Morocco)	10	688	Arabic (UAE)	2	26
Arabic (Syria)	10	1,081	Arabic (Libya)	2	1,549
Arabic (Algeria)	8	788	Arabic (Qatar)	3	227
Arabic (Lebanon)	8	198	Hindi	2	1,377
Arabic (Palestine)	7	111	Arabic (Kuwait)	2	12
Arabic (Saudi Arabia)	6	333	Arabic (Yemen)	1	10
Arabic (Sudan)	6	959	Arabic (Tunisia)	5	571

Table 1: Distribution of speakers and utterances in the *MenaSpeechBank* dataset by mother tongue.

Lang/Dial	#Spk	#Utt	Utt/Spk	Dur/Utt (s)
MSA	101	10,777	133	6
English	29	5,548	192	6
Gulf	15	287	17	5
Egyption	12	515	45	5
North African	10	168	17	4
Levantine	10	346	35	5
Total	177	**17,641**	-	-

Table 2: Speaker and utterance statistics in the dataset by language or dialect. #Spk denotes the number of speakers, #Utt the number of utterances, Utt/Spk the average number of utterances per speaker, and Dur/Utt the average utterance duration (in seconds).

are reported *per language variant*: many speakers contributed recordings in both Arabic and English. Consequently, the per-variant counts sum to 177 speaker entries, but these include overlaps across variants. In total, the dataset contains **124 unique speakers** as mentioned earlier. As MSA and English dominate the collection, they naturally include more speakers and a higher average number of utterances per speaker.

Reference Speech Samples. *MenaSpeechBank* consists of **17,641 speech utterances** comprising ≈26.4 hours of speech. Figure 7 illustrates the distribution of speech utterances by language and by dataset. The majority of utterances are in MSA and English, and notably, our in-house dataset comprises 98% of the total corpus.

Domain and subcategory wise distribution. In Figure 3, we summarize the distribution of conversations across domains and subcategories. A large portion of interactions fall under *Knowledge/Topic* (80%), alongside a substantial share of *Task/Service* dialogues (20%). This mix reflects the natural roles and intents of the reference speakers, which frequently elicit information-seeking and discussion-oriented scenarios while still capturing many actionable, service-oriented requests. Within *Knowledge/Topic*, a small set of everyday themes contributes the most instances. *Culture & Society* accounts for 24.7% of all samples (30.7% within-domain), followed by *Travel, Transport & Places* (11.6%) and *Religion & Spirituality* (11.2%). These three subcategories comprise 47.5% of the dataset. *Task/Service* conversations are led by *Productivity, Media & Education* (7.0% overall; 35.6% within-domain) and *Health & Wellness* (5.5% overall; 28.1% within-domain), with *Travel & Mobility* adding a further 3.0%. Overall, this distribution highlights strong coverage of culturally grounded interactions and tasks.

Metric	Value	Metric	Value
Conversations	416,297	Asst. turns	2,431,479
User turns	2,144,048	Asst. turns/conv.	5.84
User turns/conv.	5.15	Asst. words/turn	32.04
User words/turn	18.13	User duration/turn	11.01s

Table 3: Statistics of MSA conversations. Asst.: Assistant.

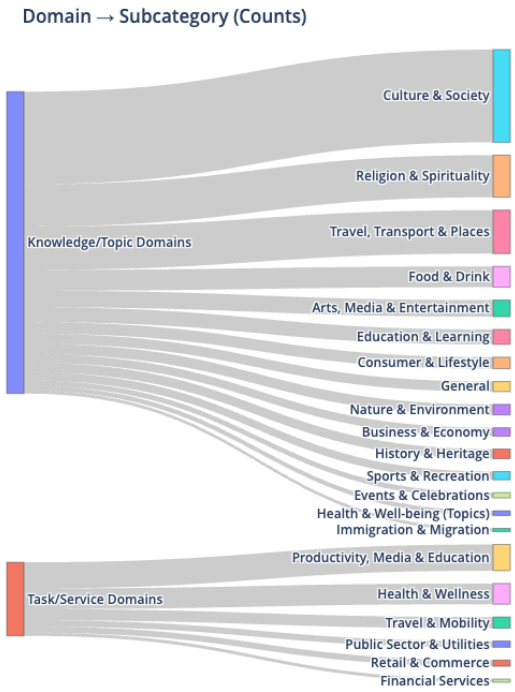

Figure 3: Distribution of conversation domains and their respective subcategories.

4 Experimental Setup

Data splits. For training and evaluation, we split the corpus *by speaker profile* to ensure that conversations associated with the same underlying speaker/persona do not appear in multiple splits, thereby preventing overlap of speaker-grounded patterns (e.g., style, lexical preferences, or recurring personal facts) from train to test. We first reserve 12.3% of the data as a test set at the profile level. The remaining 87.7% is then partitioned into training and development sets in a 90/10 ratio using stratified sampling over scenarios, preserving the scenario distribution across splits. This procedure

results in ∼328K training, ∼36K development, and ∼51K test conversations.

Models. For baseline comparisons, we include strong closed-source models such as **Gemini-2.5 Pro** and **GPT-audio**, evaluated both in audio-native setting and within an ASR→LLM pipeline. We also include open-weight models such as **Qwen** and Arabic-centric models including **Fanar-C-1-8.7B** (Team et al., 2025), **Fanar-C-2-27B** (Team et al., 2026), and **Allam-7B** (Bari et al., 2025).

Fine-tuning. We fine-tuned an in-house Qwen2.5-Omni-7B model that had been previously trained on 10K hours of ASR data. We fine-tuned this model on approximately 330K multiturn conversations for 2 epochs. More details are provided in Appendix F.

Evaluation data. For benchmarking, we selected a subset of the *MenaSpeechBank* test split comprising 206 MSA conversations. Only the user turns were recorded, spoken by 10 paid Arabic-speaking annotators. Recording quality was verified using WER computed on automatic transcriptions, generated using Fanar Aura, yielding an average WER of 12%.

Evaluation setup. We evaluate conversational response quality in an *assistant-initiated* multi-turn setting where user inputs are provided as *audio* (both synthetic and human-recorded). For each conversation, the model first receives the initial assistant and user-audio turns and produces an assistant response. For subsequent turns, the model is conditioned on the accumulated interaction history consisting of prior user-assistant pairs (user audio and assistant text) to generate the next response. We report turn and conversation wise results. Audio models are evaluated under two input conditions: (i) direct audio input and (ii) an ASR→LLM cascaded pipeline, while text-only models are evaluated using the ASR→LLM pipeline. Both settings are applied to speaker-conditioned synthetic and human-recorded conversations.

LLM-as-a-Judge scoring. To score generated responses, we use an LLM-as-a-judge that evaluates *only the final candidate assistant turn*, while using the full preceding transcript, the provided profile memory, and the session scenario as context. The judge outputs a structured checklist over ten rubrics: *request understanding*, *dialogue coherence*, *context tracking*, *profile memory consistency*, *task completion*, *constraint satisfaction*, *specificity actionability*, *calibration honesty*, *language tone match* and *cultural pragmatic fit*. We summarize quality using two metrics: *Average Pass Rate (APR)*, the fraction of turns that jointly satisfy all required rubric checks, and *Average Rubric Score (ARS)*, the mean pass rate across individual rubric checks. We use ***APR as the primary metric*** in the main paper because it provides a stricter measure of overall response quality, while ARS is used for diagnostic analysis and reported in the Appendix G. Similar rubric-based evaluation setup have been adopted in prior work (Guo et al., 2025; Gosai et al., 2025).

	Conversation		Turn	
Model	**Synth**	**Human**	**Synth**	**Human**
Large Closed Models				
Gemini-Audio	0.777	0.806	0.833	0.871
Gemini (ASR)+ Gemini	0.828	0.815	0.877	0.853
GPT-audio	0.767	0.728	0.801	0.821
GPT (ASR)+GPT-5.2	0.845	0.816	0.888	0.866
Open and Fine-tuned Models				
Qwen2.5-Omni-7B	0.024	0.034	0.256	0.226
Qwen2.5-Omni-7B FT	0.252	0.204	0.559	0.534
Arabic-centric Cascaded Models				
Fanar(Aura)+Allam-7B	0.262	0.248	0.562	0.508
Fanar(Aura)+Fanar	0.366	0.353	0.616	0.632
Fanar(Aura)+Fanar-2	0.434	0.371	0.669	0.646

Table 4: Conversation and turn level APR results on synthetic and human-recorded inputs. Full APR/ARS results are provided in Appendix Table 12.

5 Results and Discussion

Table 4 reports conversation- and turn-level APR on speaker-conditioned synthetic and human-recorded inputs, while APR/ARS results, extended ablations, and fine-grained breakdowns are reported in Appendix G.

Overall performance. Closed models provide the strongest reference results. GPT (ASR)+GPT-5.2 achieves the highest conversation-level APR on both synthetic and human inputs (0.845 and 0.816), while Gemini (ASR)+Gemini-2.5-Pro performs best at the turn level on synthetic recordings (0.877). These results indicate that both direct audio modeling and cascaded ASR→LLM pipeline remain competitive for this task, with the cascaded system benefiting from stronger text-side reasoning after transcription.

Turn-level vs. conversation-level quality. Across all systems, turn-level APR is higher than conversation-level APR. This is expected because conversation-level evaluation is stricter and reflects accumulated failures across multiple turns. The gap

is especially large for weaker open-weight models, indicating that isolated turns may pass some requirements, while maintaining speakers' profile-aware, and scenario-consistent behavior over a *full conversation remains challenging*.

Does performance differ between synthetic and human speech inputs? Human recordings do not uniformly degrade performance, but they introduce greater variability across systems. For example, Gemini-2.5 Pro improves from 0.777 to 0.806 at the conversation level and from 0.833 to 0.871 at the turn level, suggesting robust handling of human-recorded speech. In contrast, GPT-audio and GPT (ASR)+GPT-5.2 show lower conversation-level APR on human inputs. This pattern suggests that *human speech affects systems differently depending on their audio understanding and transcription quality*.

Does fine-tuning help? Fine-tuning substantially improves the open-weight Qwen2.5-Omni-7B model. At the conversation level, APR improves from 0.024 to 0.252 on synthetic inputs and from 0.034 to 0.204 on human recordings. At the turn level, the gains are even larger, increasing from 0.256 to 0.559 on synthetic inputs and from 0.226 to 0.534 on human recordings. These results show that task-specific *fine-tuning improves both turn and conversational responses*.

Arabic-centric cascaded models. Since no end-to-end AudioLLM is specifically designed for Arabic conversational speech in our setting, we evaluate Arabic-centric ASR→LLM pipelines using Fanar-Aura ASR followed by ALLaM and Fanar models. Among these systems, Fanar(Aura)+Fanar-C-2-27B performs best, reaching 0.434 conversation-level APR on synthetic inputs and 0.371 on human recordings, with turn-level APR of 0.669 and 0.646, respectively. Across all four evaluation settings, Fanar-C-2-27B outperforms both ALLaM and Fanar-C-1-8.7B when paired with the same Fanar-Aura ASR frontend. This suggests that stronger Arabic-centric text reasoning improves the cascaded pipeline after ASR. However, the remaining gap to strong closed-model performance highlights *the need for stronger Arabic speech-native models*.

ARS as diagnostic evidence. While APR is our primary metric, ARS provides useful diagnostic information by measuring average rubric-level satisfaction. In Appendix Table 12, we observe cases where ARS is substantially higher than APR, indicating that models often satisfy several rubric dimensions but fail at least one required criterion. This supports our choice of APR as the main metric, since it better captures complete response success in multi-turn conversational evaluation.

LLM-judge agreement with humans. To assess evaluation reliability, we manually annotated 10 human-recorded conversations comprising 50 turns. The LLM judge shows strong agreement for the fine-tuned model (AC1 = 0.83; observed agreement = 0.86).[7] These results provide additional support for using the LLM judge as an evaluator for response quality checking.

Scope of the evaluation. We evaluate a focused benchmark subset to establish initial baselines under compute, latency, and API-cost constraints. The subset retains key properties of the benchmark, including multi-turn interactions with both speaker-conditioned synthetic and human-recorded inputs, enabling controlled comparison across direct-audio, cascaded, closed, open-weight, and fine-tuned systems.

6 Conclusions

We introduced *MenaSpeechBank*, to our knowledge the first bilingual Arabic–English MENA-centric speech bank and speaker-conditioned multi-turn conversation resource for evaluating and training speech-capable conversational models. We also presented an LLM-as-judge evaluation protocol with APR as the primary metric and ARS as a diagnostic measure, covering relevance, completeness, specificity, coherence, context tracking, calibration, language, and safety. Our experiments show that strong closed-model perform best, while fine-tuning substantially improves open-weight Speech-LLMs and Arabic-centric ASR→LLM pipelines provide competitive cascaded baselines. Future work will scale evaluation to the full benchmark and study training strategies for improving specificity, persona consistency, and robustness to human speech.

7 Limitations

MenaSpeechBank is intended as a controlled resource for studying persona-conditioned and speaker-conditioned conversational speech, rather than an exhaustive representation of all MENA populations, dialects, or speaking styles. Coverage

[7] We report both Gwet's AC1 (Gwet, 2008) and raw agreement.

across countries and dialect varieties can be further expanded and balanced, especially as more diverse reference speakers become available. The current persona schema also simplifies some sociolinguistic attributes through bounded age ranges and categorical fields. Similarly, the WVS-inspired attributes and heuristics for technology access and AI use provide controllability, but should be interpreted as modeling choices rather than demographic claims. Future work can refine these attributes to better capture social and linguistic nuance.

The generated conversations and speaker-conditioned audio may differ from naturally occurring dialogue in spontaneity, disfluency, turn-taking, prosody, and recording conditions. Although we use deterministic PQI checks and conservative checks to reduce common artifacts.

Our evaluation includes both speaker-conditioned and human-recorded inputs, and it is conducted on a focused subset of the benchmark and emphasizes scenario-driven conversation and spoken QA.

Ethics Statement

Our dataset is designed to minimize privacy risk: we do not collect or store personally identifiable information, and the released artifacts are intended for research use under clear documentation. We use only high-level demographic cues from reference speakers as seeds and extend them with WVS-inspired attributes and heuristic signals, so the resulting persona profiles are synthetic and are not intended to correspond to any real individual. While synthetic personas enable controllable generation at scale, they can still surface representational biases or stereotypes inherited from the underlying schema, value priors, or scenario design. To mitigate this, our pipeline conditions summaries strictly on explicit profile fields and applies deterministic checks for groundedness and constraint compliance; we additionally recommend routine checks for representational harms (e.g., gender- or nationality-linked stereotypes) before release or deployment, and we position persona summaries as conditioning signals for data generation rather than as faithful depictions of real people.

Broader Impact

MenaSpeechBank enables controllable, large-scale synthetic speech instruction data that can accelerate AudioLLM adaptation for underrepresented languages and speech varieties, supporting more inclusive spoken assistants, educational applications, and improved information access for MENA speakers while enabling systematic robustness studies across accents, dialects, and bilingual usage. To ensure these benefits are realized responsibly, we promote clear documentation and intended-use guidance, appropriate licensing and access controls, and routine check of personas and scenarios to reduce representational bias and discourage misuse (e.g., impersonation or misinformation), and we encourage future work to broaden evaluations across demographic groups and dialects alongside targeted risk assessments before deployment.

References

Badr M. Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification. In *Interspeech 2025*, pages 2790–2794.

Firoj Alam, Md Arid Hasan, and Shammur Absar Chowdhury. 2025. SpokenNativQA: Multilingual Everyday Spoken Queries for LLMs. In *Interspeech 2025*, pages 2685–2689.

Firoj Alam, Ali Ezzat Shahroor, Md. Arid Hasan, Zien Sheikh Ali, Hunzalah Hassan Bhatti, Mohamed Bayan Kmainasi, Shammur Absar Chowdhury, Basel Mousi, Fahim Dalvi, Nadir Durrani, and Natasa Milic-Frayling. 2026. Oasis: A multilingual and multimodal dataset for culturally grounded spoken visual qa. *Preprint*, arXiv:2510.06371.

M Saiful Bari, Yazeed Alnumay, Norah Alzahrani, Nouf Alotaibi, Hisham Alyahya, AlRashed AlRashed, Faisal Mirza, Shaykhah Alsubaie, Hassan Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman I Alsubaihi, Maryam Al Mansour, Saad Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. Allam: Large language models for arabic and english. In *International Conference on Learning Representations*, volume 2025, pages 34179–34214.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and

Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

EVS/WVS. 2022. European values study and world values survey: Joint evs/wvs 2017-2022 dataset (joint evs/wvs). *JD Systems Institute and WVSA. Dataset Version 3.0. 0*.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and 1 others. 2025. Audio MultiChallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *arXiv preprint arXiv:2512.14865*.

Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Jayeon Park, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and 1 others. 2025. Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning. *arXiv preprint arXiv:2510.12712*.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, and 1 others. 2020. World values survey: Round seven country-pooled datafile. *World Values Survey*.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. AudioGPT: Understanding and generating speech, music, sound, and talking head. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23802–23804.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15245–15270, Bangkok, Thailand. Association for Computational Linguistics.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech 2021*, pages 2127–2131.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered LLM. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Vahid Noroozi, Zhehuai Chen, Somshubra Majumdar, Steve Huang, Jagadeesh Balam, and Boris Ginsburg. 2024. Instruction data generation and unsupervised adaptation for speech language models. In *Interspeech 2024*, pages 4049–4053.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8689–8696.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Traian Rebedea, Makesh Sreedhar, Shaona Ghosh, Jiaqi Zeng, and Christopher Parisien. 2024. CantTalkAboutThis: Aligning language models to stay on topic in dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12232–12252, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

FANAR Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Asim Ersoy, Masoomali Fatehkia, Mohammed Qusay Hashim, and 18 others. 2026. Fanar 2.0: Arabic generative ai stack. *arXiv preprint arXiv:2603.16397*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2025b. Neural codec language models are zero-shot text to speech synthesizers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:705–718.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. 2025c. Benchmarking contextual and paralinguistic reasoning in Speech-LLMs: A case study with in-the-wild data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14133–14148, Suzhou, China. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A *MenaSpeechBank*: Applications

MenaSpeechBank is, to the best of our knowledge, **the first** *bilingual* (Arabic–English) and *MENA-centric* resource that jointly provides reference speech bank, profile-conditioned multi-turn conversations and corresponding speech. This combination enables controlled studies of spoken conversational behavior while remaining close to real voice-assistant interaction patterns. Below we outline several representative use cases.

Audio-to-text multi-turn dialogue benchmarking. *MenaSpeechBank* supports turn-level evaluation of AudioLLMs, measuring response quality, coherence, and context tracking when user inputs are provided as speech rather than transcripts.

Long-context spoken dialogue memory evaluation. The multi-turn structure enables stress-testing of long-horizon conversational memory: whether models retain and correctly reuse constraints, preferences, and entities introduced in earlier *spoken* turns, and whether they avoid hallucinating user-specific facts beyond the provided profile memory.

Speech-conditioned personalization modeling. As each dialogue is grounded in an explicit persona profile, *MenaSpeechBank* can be used to study personalization under speech input, including consistent use of profile attributes, stable interaction style, and appropriate register across turns.

Robustness and invariance probing under speech variability. By pairing controlled scenarios with diverse speakers and recordings, *MenaSpeechBank* facilitates analyses of robustness to dialect/accent variation, speaker identity, speaking rate, and recording conditions, while holding the underlying semantic intent fixed.

Multi-turn conversational alignment for AudioLLMs. Finally, *MenaSpeechBank* can support preference-based optimization (e.g., reranking, rejection sampling, or preference tuning) using rubric- and pairwise-judged outputs to explicitly target multi-turn dialogue qualities such as coherence, calibration, and profile-consistent context tracking, rather than single-turn helpfulness alone.

B Details on Persona Generation

B.1 Handling of Speaker Missing Persona attributes

If iso3 or country name is missing, we derive it from the normalized country name (and vice-versa). We then sample culturally aligned name and city from country-specific lists; if unavailable, we fall back to regional/default lists. Profession is sampled from age-bucketed occupation pools, which are optionally constrained by education level. Household/digital-access fields (device/connectivity/AI competence) and 1–2 AI use cases are sampled from small categorical pools conditioned on age/education to ensure plausibility.

Figure 4 shows an example illustrating the input persona attributes and the corresponding derived outputs. We omit the descriptive persona summary for brevity.

B.2 WVS Incorporation

We retrieve a country-level WVS vector (via ISO3/country name) as normalized value axes and use it as a soft conditioning prior that guides conversational framing (e.g., trust/tradition/self-expression) without asserting individual beliefs. If WVS is unavailable, we use a default axis vector with a small age-based adjustment.

C Prompts

C.1 Prompt Details

Figure 5 presents the system and user prompts used for persona-to-summary generation. Table 5 shows the prompt used to generate domain-specific conversation topics for each taxonomy path. Table 6 provides the prompt used to generate user-centric conversational scenarios for each topic. Tables 7 and 8 present the system and user prompts used for dialogue generation with strict JSON output and language constraints.

C.2 Persona Quality Assessment (PQI).

To verify that each generated persona summary is (a) grounded in the underlying attribute record and (b) compliant with the required narrative constraints, we define a deterministic, rule-based *Persona Quality Index* (PQI). PQI decomposes quality into two components: *static fidelity* (PQI_S), which measures whether the summary stays within the information supported by the persona attributes, and *narrative compliance* (PQI_N), which measures adherence to format and style constraints (e.g., first-person voice, grounded opening, length bounds, and a non-moralizing ending). For each persona i, we compute a set of binary checks $\{S_{i,j}\}_{j=1}^{J}$ (static) and $\{N_{i,k}\}_{k=1}^{K}$ (narrative), where each indicator equals 1 if the corresponding requirement is satisfied and 0 otherwise. We then define:

$$\begin{aligned} \mathrm{PQI}_S(i) &= \sum_{j=1}^{J} S_{i,j}, \\ \mathrm{PQI}_N(i) &= \sum_{k=1}^{K} N_{i,k}, \\ \mathrm{PQI}(i) &= \mathrm{PQI}_S(i) + \mathrm{PQI}_N(i). \end{aligned} \tag{1}$$

Here, j and k index the individual static and narrative checks, respectively, and J and K denote the total number of checks in each group.

In our implementation, PQI_S comprises $J{=}5$ grounding checks and PQI_N comprises $K{=}7$ narrative checks, yielding $\mathrm{PQI} \in [0, 12]$ (higher is better). We consider a summary *fully compliant* iff $\mathrm{PQI}(i) = 12$, i.e., all checks pass.

Narrative compliance (PQI_N) uses deterministic pattern-based tests to enforce: *(i)* first-person voice, *(ii)* length compliance (90–180 words), *(iii)* grounded opening (name and age appear in the first sentence; country may optionally appear early),

The speaker input attributes.


```
{
  "speaker_id": "QASR_9b101a25d6ef3a5f300ac4c8fce2058f",
  "nationality": "Algeria",
  "age": 52,
  "gender": "Female",
  "mother_tongue": "Arabic",
  "education_level": "bachelor",
  "languages": ["msa"],
  "iso3": "DZA"
}
```


The generated persona profile (illustrative excerpt) preserves the core attributes and adds controlled extensions:


```
{
  "speaker_id": "QASR_9b101a25d6ef3a5f300ac4c8fce2058f",
  "speaker_age": 52,
  "speaker_nationality": "Algeria",
  "speaker_mother_tongue": "Arabic",
  "native": true,
  "persona_name": "Samira",
  "gender": "female",
  "city": "Ain Oulmene",
  "education_level": "bachelor",
  "profession": "chief pharmacist",
  "marital_status": "married",
  "household_type": "lives with family",
  "digital_access": {
    "device": "smartphone and laptop",
    "connectivity": "public Wi-Fi occasionally",
    "ai_competence_level": "advanced"
  },
  "ai_use_cases": [
    "writing or editing professional emails",
    "trip or visa document checklist help"
  ],
  "wvs_profile": {
    "axis_rel_trad": 0.8989932771328996,
    "axis_self_expr": 0.22982508432763685,
    "axis_trust": 0.2151343635877581,
    "axis_gender_eq": 0.6573941789147184,
    "axis_inst_trust": 0.5202429027162727,
    "axis_auth": 0.7220759785629295,
    "axis_econ_egal": 0.7188702133402142,
    "axis_env": 0.4258046789134873,
    "axis_civic": 0.22560079951657136
  },
  "ocean": {
    "openness": 0.5705403422556412,
    "conscientiousness": 0.8554382388838186,
    "extraversion": 0.5467230344675954,
    "agreeableness": 0.6208736731466347,
    "neuroticism": 0.3090192732191873
  }
}
```


Figure 4: Input persona attributes and their derived outputs.

System Prompt.

```
You are a specialized AI "Persona Profiler." Your sole task is to take a single JSON object
    containing persona attributes-aligned with a WVS-inspired persona schema---and produce a
    coherent, natural, first-person persona summary.

Input: JSON object with persona attributes (schema below).

Output: strict JSON only:
{"summary_first_person":"<120--180 words>"}

Hard constraints:
- One paragraph. First-person only ("I..."). Natural, grounded tone.
- 90--180 words.
- Opening sentence MUST include: persona_name + speaker_age + (city if present) +
    speaker_nationality.
- Use ONLY explicit JSON facts. No invention, inference, stereotypes, or extra backstory.
- Do not output IDs (speaker_id, script_id, region) unless user explicitly asks.

Content selection (use if present):
- Identity/core: persona_name, speaker_age, gender (only if it fits naturally),
    speaker_mother_tongue, native, city, speaker_nationality.
- Life context: Use marital_status and household_type to describe living arrangements and family
    situation, ensuring they are mutually consistent.
- Work/education: Use education_level and profession together, ensuring the stated role is
    plausible and consistent with the education level.
- Digital/AI: Reflect technology access and AI familiarity using digital_access.device,
    digital_access.connectivity, and digital_access.ai_competence_level; also consider life
    context, education, and profession.
- AI Use Cases: Integrate them naturally based on the ai_use_case information and other personal
    attributes. Do not assume enthusiasm for AI---reflect only what the fields support.
- Religion: Mention only if religion exists; keep neutral.
- Include small, ordinary, concrete details (routines, habits, regular activities).
- Interests, activities and hobbies must appear organically within the flow of daily life and
    they should be consistent with age.

Values/profiles:
- wvs_profile: NEVER mention numbers, axes names, "WVS", or "scores". If used at all, reflect
    implicitly via everyday preferences/actions. If a value is null, ignore it. Do not "explain"
     or psychoanalyze.
- ocean: NEVER mention trait names or numbers. Do not write "I am X" traits. If used, show via
    behavior (e.g., planning habits) without labeling.

Negative/contradictions:
- If conflicts/negatives exist, preserve them plainly; no "growth/lesson" framing.

Ending:
- No abstract/moral reflections ("I've learned...", "In the end..."). End with a routine, near-
    term goal, or ongoing interest.

Failure rule:
- If you cannot satisfy opening/length/JSON-only, output {"summary_first_person":""}.

Internal self-check before final:
[ ] first-person; [ ] 120--180 words; [ ] opening fields present; [ ] no IDs; [ ] no invented
    facts;
[ ] no WVS/OCEAN labels or numbers; [ ] no "I am <trait>"; [ ] grounded ending; [ ] strict JSON
    only.
```

User Prompt Template.

```
Generate a first-person persona summary from the following JSON data consisting of persona
    attributes.

Input JSON:
{json_object}
```

Figure 5: System and user prompts used for persona-to-summary generation.

```
prompt = """You are an expert in conversational AI.
Domain: "{domain_path}"

Task:
Generate exactly {num_topics} specific, relevant conversation topics for this domain.

Requirements:
1) Topics must be specific and actionable user-intent areas.
2) Topics must represent common user intents and conversation flows.
3) Topics must be diverse and cover different aspects of the domain (avoid near-duplicates).
4) Use clear, concise topic names (2--6 words each).
5) Prefer topics commonly relevant in MENA contexts when applicable. Do NOT invent country-specific
    facts unless implied.

Output format (STRICT):
Return ONLY valid JSON (no markdown, no explanations).
The JSON must be an object with a single key "topics" whose value is an array of strings.
Example:
{{"topics": ["Topic 1", "Topic 2"]}}

Generate exactly {num_topics} topics.
"""
```

Table 5: Prompt used to generate domain-specific conversation topics for each taxonomy path.

```
PROMPT = """You are an expert in conversational AI and user experience design.
Given:
- Domain path: "{domain_path}"
- Topic: "{topic}"

Generate {num_scenarios} realistic conversational scenarios that users might have when interacting
    with a system for this topic.

Each scenario should be:
1. A natural, user-focused description of what the user wants to accomplish
2. Specific and actionable (not vague)
3. Representative of real-world user intents
4. Written in a conversational, user-centric way (e.g., "A customer wants to...", "A parent needs
    to...")

Return ONLY a JSON array of scenario strings, no explanations, no markdown.
"""
```

Table 6: Prompt used to generate user-centric conversational scenarios for each topic.

```
SYSTEM_PROMPT = """
You are a helpful assistant.

OUTPUT FORMAT (STRICT)
- Return exactly one JSON object with a "messages" array.
- Each item in "messages" must be an object with:
  - "role": "user" or "assistant"
  - "content": message text
- No extra text, no markdown, no code fences.

INITIATION (VARY THIS)
- Vary who starts the conversation:
  - ~50-60%: start with "user" (inbound)
  - ~40-50%: start with "assistant" (outbound call)

LANGUAGE (CRITICAL)
- All user messages MUST be in PURE {language}.
- No English words, no code-switching, no transliterated English.
- If {language} is Arabic, use ONLY Modern Standard Arabic.
  - Do NOT use dialect words/particles.
  - If any dialect word appears, rewrite that message into MSA before output.
- If the assistant starts (outbound), its first message MUST also be in PURE {language} (and in MSA
    if Arabic).

VALID JSON EXAMPLE:
{{"messages":[{{"role":"user","content":"..."}},{{"role":"assistant","content":"..."}}]}}
"""


DIALOGUE_GENERATION_PROMPT = """
You are an expert in conversational AI and natural language generation.
Your task is to generate a realistic, natural dialogue between a user and an AI assistant.

USER INFORMATION (Persona):
- Name: {user_name}
- Age: {user_age}
- Nationality: {user_nationality}
- Language: {language}  (If {language} is Arabic, it MUST be Modern Standard Arabic / MSA)
- Gender: {user_gender}
- City: {user_city}
- Education Level: {user_education_level}
- Profession: {user_profession}
- Marital Status: {user_marital_status}
- Household Type: {user_household_type}
- Religion: {user_religion}
- Digital Access: {user_digital_access}
- AI Use Cases: {user_ai_use_cases}
- Values Profile: {user_wvs_profile}
- Persona Summary: {summary}

TOPIC INFORMATION:
- Domain: {domain}
- Topic: {topic}
- Scenario: {scenario}
- Max Messages: {max_messages}
```

Table 7: System and user prompts for dialogue generation with strict JSON output and language constraints. (Part 1)

```
CRITICAL REQUIREMENTS:
1. Generate a natural, realistic SPOKEN conversation between a user and an AI assistant.
2. The user should speak in the target language and reflect their persona characteristics.

3. CRITICAL LANGUAGE REQUIREMENTS:
   - The user MUST speak in pure {language} without mixing English words, code-switching,
     transliterations, or dialectal variations.
   - Use ONLY native words from {language}.
   - Replace English technical terms with native equivalents.
     - Arabic examples (MSA): use هاتف محمول / هاتف instead of transliterations like موبايل.
     - Use على الإنترنت sparingly and prefer native descriptive phrases when possible.
   - Avoid transliterated English words; use native equivalents.
   - Example (Arabic): instead of آب / أب for app use تطبيق instead of بلاتفورم use منصة إلكترونية or
     any appropriate Arabic description.

4. ARABIC MODE (STRICT) - Applies ONLY if {language} is Arabic / MSA:
   - Use ONLY Modern Standard Arabic (العربية الفصحى الحديثة) for BOTH user and assistant.
   - Do NOT use dialect particles/negations/words such as:
   (إيش، شو، شنو، وين، ليش، مو، مش، عايز، بدّي، حابب، كده، كذا، هلا، هلّق، رح، لسا، بزاف، برشا).

   - Spoken MSA style: keep sentences short and natural for speech, but still MSA.
     - Allowed MSA fillers: (مم ، حسنًا، دعني أفهم، لحظة، على مهلك).

   - SELF-CHECK: If any dialect word appears, rewrite that message into MSA before output.

5. SPOKEN CONVERSATION STYLE:
   - The dialogue must sound like real spoken conversation, not written text.
   - Use short, direct messages, clarifications, and back-and-forth.
   - Avoid overly formal or written-style language.
   - The conversation should feel authentic and not scripted.

6. INITIATION (VARY THIS):
   - Sometimes the USER initiates (about 50-60%).
   - Sometimes the ASSISTANT initiates (about 40-50%) with a polite greeting.
   - If assistant starts, the greeting and initial message MUST be in {language}.
   - If {language} is Arabic/MSA, the greeting MUST be in MSA (not dialect).

OUTPUT FORMAT (STRICT):
Return ONLY a JSON object with a "messages" key containing an array of message objects.
Each message object must have:
- "role": "user" or "assistant"
- "content": message text

Inbound example:
{{"messages":[{{"role":"user","content":"Message in
    {language}"}},{{"role":"assistant","content":"Response"}}, ...]}}

Outbound example:
{{"messages":[{{"role":"assistant","content":"Greeting in
    {language}"}},{{"role":"user","content":"Response in {language}"}}, ...]}}

IMPORTANT:
- First message can be from either "user" or "assistant" (vary this).
- User must always use PURE {language} with no English mixing.
- If {language} is Arabic/MSA, BOTH sides must be strictly MSA with no dialect words.
- Generate a realistic SPOKEN conversation with natural flow. Maximum {max_messages} messages.
"""
```

Table 8: System and user prompts for dialogue generation with strict JSON output and language constraints. (Part 2)

(iv) single-paragraph prose (no bulleting or list formatting), *(v) show-don't-tell* style (no explicit "I am ⟨trait⟩" self-ascriptions), *(vi)* presence of everyday routine markers, and *(vii)* a grounded, non-moralizing ending.

Static fidelity (PQI_S) uses conservative heuristics to reduce hallucination risk and enforce attribute grounding, including: *(i)* detecting unsupported numerals and suspicious proper nouns not present in the source JSON, *(ii)* preventing explicit leakage of value terminology (e.g., "WVS", "score", or axis names), and *(iii)* allowing religion mentions only when religion is explicitly provided in the source JSON. Full definitions (regexes, token windows, and lexicons) are provided in the appendix to ensure exact reproducibility.

PQI Results.

Overall quality is high. Across $n = 469$ personas, PQI is strongly concentrated near the maximum ($\mathrm{PQI} \in [0, 12]$), and all summaries are successfully generated (no missing/empty outputs). In total, $406/469$ personas (86.6%) achieve $\mathrm{PQI} = 12$ and $468/469$ (99.8%) score ≥ 10 (mean $= 11.85$; median $= 12$).

Component scores are consistently strong. Static fidelity is near-perfect, with $\mathrm{PQI}_S = 5$ for $464/469$ personas (98.9%). Narrative compliance is also high: $\mathrm{PQI}_N = 7$ for $410/469$ (87.4%) and $\mathrm{PQI}_N = 6$ for $57/469$ (12.2%), suggesting that non-max PQI values mainly reflect minor narrative deviations rather than grounding issues. In addition to PQI, we emit conservative, deterministic audit flags: $64/469$ personas (13.6%) trigger at least one flag, while individual triggers are rare (unsupported-content: $4/469$, 0.9%; first-person: $2/469$, 0.4%; and length, religion-policy, and value-leakage: $1/469$, 0.2% each). Overall, these results suggest the pipeline produces grounded, stylistically compliant persona summaries at scale, with most flags driven by audit-oriented conservatism rather than systematic issues.

C.3 Persona–Scenario Matching

We hypothesize that personas are most plausible in different domain scenarios. To operationalize this, we automatically align each persona with the most relevant domain scenario(s) by comparing the persona summary text to each scenario text and scoring their similarity. We use a hybrid similarity because embedding-based similarity is robust to paraphrase (capturing semantic relatedness even when surface wording differs), while token overlap anchors the match when domain-specific terms must appear explicitly. This combination produces more robust and realistic persona-scenario pairings.

Hybrid similarity. Let p denote a persona summary and d denote a domain scenario text. We compute *(i)* embedding cosine similarity and *(ii)* token-overlap Jaccard similarity, then combine them with a weighted average:

$$s_{\mathrm{emb}}(p, d) = \cos(\mathbf{e}_p, \mathbf{e}_d)\,, \tag{2}$$

$$\mathbf{e}_p = \mathrm{Embed}(p), \quad \mathbf{e}_d = \mathrm{Embed}(d), \tag{3}$$

$$s_{\mathrm{jac}}(p, d) = \frac{|T(p) \cap T(d)|}{|T(p) \cup T(d)|}, \tag{4}$$

$$s_{\mathrm{hyb}}(p, d) = w\, s_{\mathrm{emb}}(p, d) + (1 - w)\, s_{\mathrm{jac}}(p, d). \tag{5}$$

We set $w = 0.5$ (equal weighting of embedding and token-overlap signals). Finally, we retain only sufficiently similar pairs using a minimum similarity threshold τ:

$$\mathrm{keep}(p, d) \iff s_{\mathrm{hyb}}(p, d) \geq \tau, \tag{6}$$

with $\tau = 0.05$ in our experiments.

D Data Statistics

Table 10 summarizes the distribution of speakers across different age groups in the *MenaSpeechBank* dataset, highlighting the diversity of the collected speakers.

Figure 7 illustrates the distribution of speech utterances across languages or dialects, further grouped by dataset origin to show the contribution of each source to the overall dataset.

Age group	# Speakers
<20	4
20-29	50
30-39	19
40-49	20
50-59	16
60+	12

Table 10: Distribution of speakers across age groups in the *MenaSpeechBank* dataset.

Figure 6: Root-left taxonomy with two highlighted branches: task/service domains (blue) and knowledge/topic domains (green).

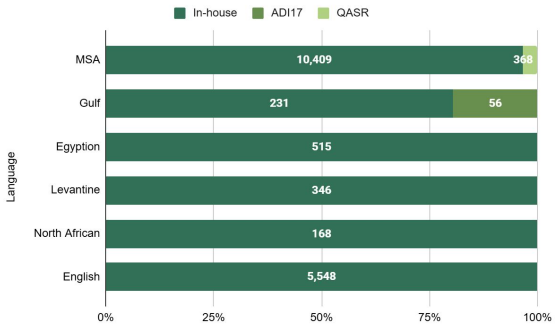

Figure 7: Number of speech utterances in the *MenaSpeechBank* dataset, grouped by language or dialect and by dataset origin.

E Data Recording

We hired annotators for recording through a third-party company and compensated at the standard hourly rate that is reasonable for their location. All annotators were professionals fluent in both Arabic and English and held at least a bachelor's degree. Each annotator signed a non-disclosure agreement (NDA) specifying all permitted uses of the data.

E.1 Annotation Guidelines

Voice Recording Instructions

1. **Review the conversation:** For each turn, you will see an Assistant message (context) and a User message (the text you must record).
2. **Record each User turn**: Click Record, read the User text aloud exactly as written, then click Stop.
3. **Playback:** Click Play to listen and confirm your audio is clear and complete.
4. **Re-record if needed:** If you made a mistake or the audio is noisy/cut off, click Delete and record again.
5. Check your recording: Always press Play. If the audio is too quiet, clipped, or incorrect, Delete and re-record. After recording all 3 user turns, select a conversation-flow rating from 1 (Bad) to 5 (Good).
6. **Submit:** You can submit only after all 3 recordings are present and a rating is selected.

E.2 Instructions

What to record: Record only the text shown in the User (Record this) box. The Assistant text is provided for context only and should not be recorded.

Recording rules

- Read exactly: Do not paraphrase, translate, correct grammar, or add extra words. Read the user text word-for-word.
- Speak naturally and clearly: Use a normal speaking pace and clear pronunciation.
- Follow punctuation: Questions should sound like questions. Pause naturally at commas and sentence boundaries.
- One speaker, low noise: Record in a quiet place. Avoid TV/music, loud fans, or other people speaking.
- Check your recording: Always press Play. If the audio is too quiet, clipped, or incorrect, Delete and re-record.

Conversation Flow Quality (1–5) After you finish recording all 3 turns, rate the overall conversation flow for the snippet shown (assistant + user turns). Flow quality reflects how natural and coherent the back-and-forth is.

Rate based on: relevance between turns, logical coherence, consistency, and naturalness of the dialogue.

Do NOT rate based on: whether you agree with the content, factual correctness, or your own recording quality.

- **5 (Good):** Very natural and coherent; user replies clearly match the assistant context.
- **4:** Mostly coherent; only minor awkwardness or small mismatch.
- **3:** Mixed; noticeable issues but still understandable overall.
- **2:** Poor; multiple mismatches, unclear transitions, or unnatural dialogue.
- **1 (Bad):** Very incoherent; turns are disconnected, contradictory, or confusing.

Troubleshooting

- If you cannot record, make sure your browser has microphone permission enabled, then refresh the page.
- If playback is silent, check your microphone input and re-record closer to the mic.

	Conv.-level		Turn-level	
Models	**APR**	**ARS**	**APR**	**ARS**
Allam	0.325	0.659	0.585	0.898
Fanar-C-1-8.7B	0.502	0.789	0.698	0.941
Fanar-C-2-27B	0.410	0.752	0.652	0.911

Table 11: Conversation- and turn-level results across text-only inputs for Arabic-centric models.

F Fine-tuning Parameters

LoRa fine-tuning was performed with rank $r = 32$, scaling factor $\alpha = 64$, and dropout $p = 0.05$. Training used a per-device batch size of 2 with gradient accumulation over 16 steps, a learning rate of 3×10^{-5}, and a warmup ratio of 0.2. The maximum sequence length was set to 8192. All experiments were conducted on 8 $\times$ H100 GPUs.

G Experimental Results

Table 12 presents a comprehensive summary of all experiments conducted on audio and transcribed inputs across closed-source, open-source, and Arabic-centric models. We additionally evaluate Arabic-centric models using text-only inputs; these results are reported in Table 11.

Statistical Significance. To assess whether differences between systems are reliable, we apply paired significance tests over 206 conversations using three methods: paired bootstrap (10,000 resamples), Wilcoxon signed-rank test on ARS, and McNemar's exact test on APR. All tests use the conversation as the resampling unit since turns within a conversation are not independent. We compare 64 system pairs across three conditions: models on generated speech, models on human recordings, and generated vs. human speech for the same model.

Of the 64 pairs, 54 differ significantly at p<0.05, with most at $p < 0.001$. The 10 non-significant pairs involve top-performing systems (e.g., Gemini vs. GPT-Audio on generated speech; Gemini vs. ASR+GPT on human speech), indicating comparable performance among the strongest models. For 5 of 8 models, the difference between generated and human speech is not significant, supporting the validity of using synthesized speech in the benchmark.

G.1 Speaker-wise Experimental Results

Table 13 presents the speaker-wise evaluation results for recorded conversations accross audio models (GPT-audio and Gemini). We report both con-

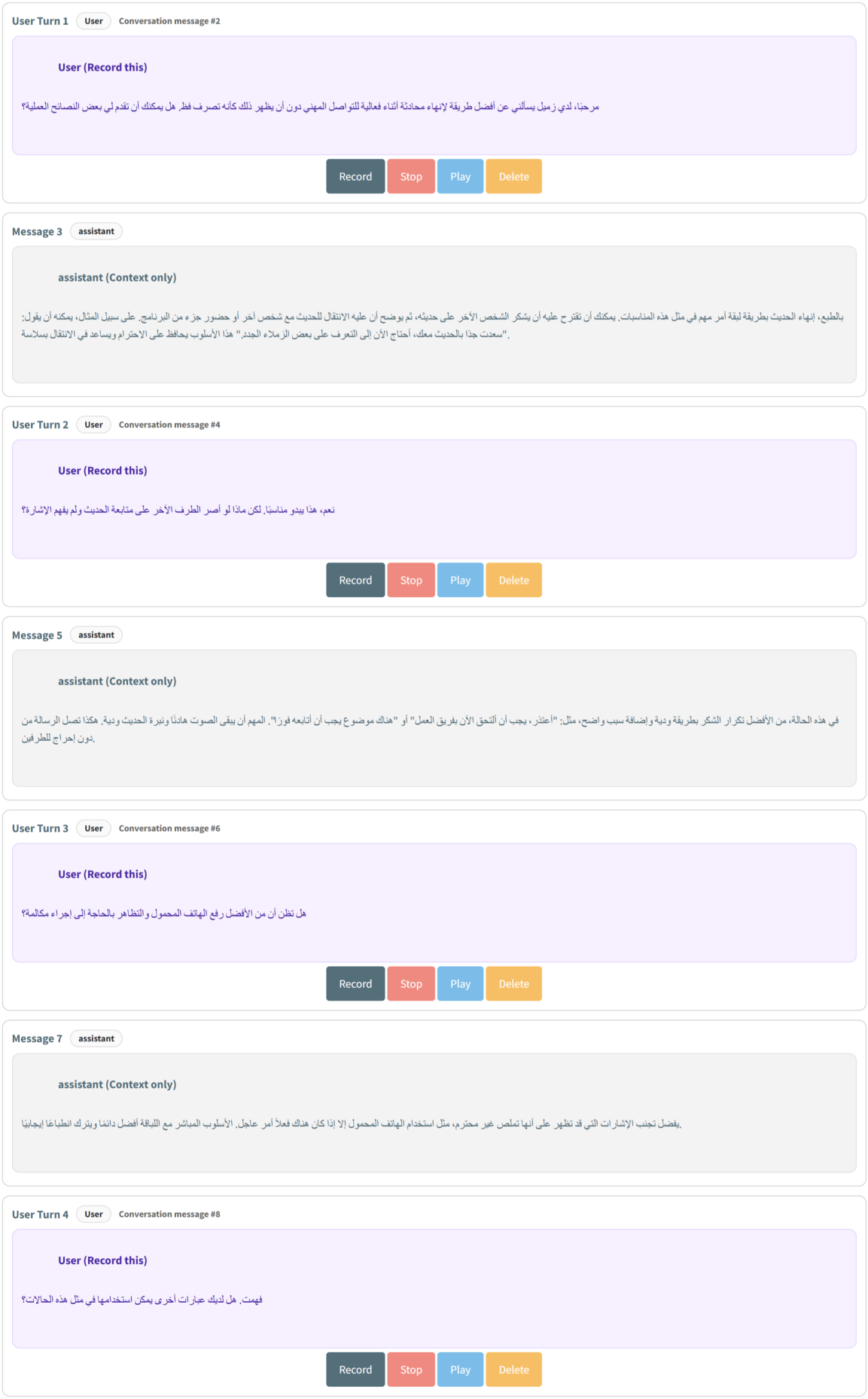

Figure 8: Screenshot of the annotation platform used for conversation recording.

Model	Input	Conversation-level				Turn-level			
		Synthetic		Human		Synthetic		Human	
		APR	ARS	APR	ARS	APR	ARS	APR	ARS
Large Closed Models									
Gemini-2.5 pro	Audio	0.777	0.892	0.806	0.926	0.833	0.971	0.871	0.980
	Transcription (Fanar)	0.820	0.925	0.850	0.937	0.850	0.976	0.859	0.979
	Transcription (Gemini)	0.828	0.937	0.815	0.925	0.877	0.979	0.853	0.978
GPT-audio	Audio	0.767	0.893	0.728	0.894	0.801	0.964	0.821	0.965
GPT-5.2	Transcription (GPT-audio)	0.845	0.947	0.816	0.947	0.888	0.979	0.866	0.977
Open & FT Models									
Qwen2.5-Omni-7B	Audio	0.024	0.314	0.034	0.271	0.256	0.734	0.226	0.712
	Transcription (Fanar)	0.010	0.266	0.024	0.267	0.234	0.716	0.228	0.709
Qwen2.5-Omni-7B FT	Audio	0.252	0.624	0.204	0.597	0.559	0.887	0.534	0.871
	Transcription (Fanar)	0.228	0.616	0.218	0.589	0.523	0.865	0.528	0.860
Arabic-centric Models									
Allam	Transcription (Fanar)	0.262	0.631	0.248	0.587	0.562	0.886	0.508	0.871
Fanar-C-1-8.7B	Transcription (Fanar)	0.366	0.708	0.353	0.727	0.616	0.894	0.632	0.907
Fanar-C-2-27B	Transcription (Fanar)	0.434	0.759	0.371	0.734	0.669	0.923	0.646	0.921

Table 12: Conversation- and turn-level evaluation results across synthetic and human conversations for audio and transcription-based inputs.

versation and turn level APR and ARS scores. Performance was stable across speakers with sufficient data (more than 17 conversations) but varied for those with limited data. Overall, there is no clear indication of systematic bias toward specific speakers. GPT-audio showed greater robustness to speaker variability, particularly for underrepresented speakers. However, speaker 5 showed lower scores on both models, suggesting potential challenges related to audio quality or speech patterns.

G.2 Topic-wise Experimental Results

Table 14 presents topic-wise evaluation results for Gemini-2.5 pro, the best-performing audio model, on human audio and Fanar-C-2-27B, the best-performing Arabic-centric model on transcibed text of human audio. Fanar-C-2-27B achieves its best performance on structured, task-oriented topics such as Events & Celebrations (APR: 0.80) and Travel & Mobility (APR: 0.75), while failing on descriptive domains like Sports & Recreation and Nature & Environment (APR: 0.00). Gemini-2.5 pro excels on conversationally rich topics including Food & Drink, Arts & Entertainment, and Business & Economy (APR: 1.00), but underperforms on service-oriented topics such as Travel & Mobility and Public Sector & Utilities (APR: 0.20-0.25). Both models share a common weakness on Travel & Mobility and Retail & Commerce, suggesting these domains present inherent challenges for conversational evaluation regardless of model architecture.

G.3 Gender-wise Experimental Results

Table 15 presents gender-wise evaluation results. For synthetic audio, *gender* refers to the reference audio used for voice cloning, which matches the assigned persona gender. For human recordings, *gender* refers to the speaker's self-identified gender.

Both models exhibit higher performance on female voices across synthetic and human audio inputs. The performance gap is most pronounced for Gemini on human recordings (APR: 0.79 female vs 0.58 male), while GPT-Audio demonstrates more balanced performance across genders (APR: 0.83 female vs 0.73 male). This disparity may be influenced by sample imbalance (143 female vs 60 male conversations in human recordings) or differences in acoustic characteristics between genders.

H Data and Resource Release

We will release the supporting code, documentation, and dataset under the CC BY-NC-SA 4.0 license to facilitate adoption, reproducibility, and future extensions.

ID	#Conv	#Turns	Gemini-2.5 pro				GPT-audio			
			Conversation-level		Turn-level		Conversation-level		Turn-level	
			APR	ARS	APR	ARS	APR	ARS	APR	ARS
0	52	260	0.712	0.858	0.827	0.968	0.712	0.885	0.808	0.962
1	51	254	0.824	0.910	0.823	0.973	0.765	0.878	0.756	0.957
2	49	245	0.755	0.865	0.792	0.966	0.735	0.861	0.771	0.960
3	17	84	0.824	0.938	0.905	0.988	0.824	0.965	0.835	0.984
4	17	85	0.882	0.976	0.965	0.995	0.882	0.941	0.859	0.962
5	8	40	0.625	0.825	0.700	0.945	0.750	0.875	0.800	0.968
6	4	20	1.000	1.000	0.800	0.975	1.000	1.000	1.000	1.000
7	3	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	1	5	0.000	0.600	0.600	0.860	1.000	1.000	1.000	1.000
9	1	5	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.980

Table 13: Speaker-wise evaluation results for Recorded Conversations across Gemini and GPT-Audio, reported at both conversation and turn levels using APR and ARS metrics. ID indicates the speaker identifier; #Conv denotes the number of conversations per speaker; #Turns denotes number of turns per speaker.

Topic	#Conv	#Turns	Fanar-C-2-27B				Gemini-2.5 pro			
			Conv.-level		Turn-level		Conv.-level		Turn-level	
			APR	ARS	APR	ARS	APR	ARS	APR	ARS
Culture & Society	47	235	0.447	0.762	0.727	0.941	0.894	0.974	0.898	0.984
Travel, Transport & Places	24	120	0.333	0.717	0.583	0.893	0.875	0.942	0.900	0.986
Productivity, Media & Education	18	90	0.222	0.689	0.500	0.893	0.722	0.900	0.800	0.977
Religion & Spirituality	17	85	0.471	0.812	0.753	0.936	0.941	0.988	0.965	0.996
Health & Wellness	15	74	0.467	0.770	0.676	0.936	0.667	0.853	0.824	0.965
Food & Drink	13	65	0.231	0.769	0.692	0.935	1.000	1.000	0.954	0.995
Education & Learning	11	55	0.455	0.791	0.750	0.944	0.909	0.964	0.927	0.989
General	9	45	0.333	0.711	0.644	0.904	0.667	0.844	0.844	0.967
History & Heritage	8	40	0.375	0.750	0.750	0.938	0.875	0.975	0.975	0.993
Arts, Media & Entertainment	7	35	0.286	0.514	0.429	0.849	1.000	1.000	0.829	0.983
Events & Celebrations	5	25	0.800	0.840	0.640	0.964	0.600	0.880	0.880	0.984
Sports & Recreation	5	25	0.000	0.440	0.440	0.824	0.800	0.960	0.840	0.968
Travel & Mobility	5	25	0.750	0.950	0.550	0.910	0.200	0.520	0.440	0.940
Consumer & Lifestyle	4	20	0.250	0.650	0.600	0.925	0.750	0.950	0.850	0.960
Nature & Environment	4	20	0.000	0.650	0.650	0.940	0.750	0.950	0.950	0.990
Public Sector & Utilities	4	19	0.250	0.500	0.450	0.900	0.250	0.650	0.650	0.940
Retail & Commerce	4	20	0.000	0.600	0.200	0.865	0.250	0.850	0.800	0.945
Business & Economy	3	15	0.667	0.933	0.933	0.987	1.000	1.000	0.867	0.987
Health & Well-being (Topics)	3	15	0.333	0.800	0.800	0.947	0.667	0.733	0.733	0.947

Table 14: Topic-wise evaluation results for Gemini-2.5 Pro, on human audio, and Fanar-C-2-27B, on Fanar (Aura)-transcribed human audio. Reported at both conversation and turn levels using APR and ARS metrics.

Input Audio	Gender	#Conv	#Turns	Gemini-2.5 Pro				GPT-Audio			
				Conversation-level		Turn-level		Conversation-level		Turn-level	
				APR	ARS	APR	ARS	APR	ARS	APR	ARS
Synthetic	Female	83	415	0.807	0.896	0.786	0.968	0.795	0.901	0.836	0.970
	Male	123	614	0.740	0.891	0.811	0.960	0.764	0.886	0.830	0.972
Human	Female	143	715	0.790	0.922	0.853	0.973	0.832	0.943	0.892	0.985
	Male	60	299	0.583	0.830	0.742	0.946	0.733	0.883	0.813	0.967

Table 15: Gender-wise evaluation results for Gemini-2.5 Pro and GPT-Audio accorss both synthetic speech and human audio inputs, reported at both conversation and turn levels using APR and ARS metrics.