Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic AudioLLMs

Anonymous ACL submission

Abstract

Audio large language models (LLMs) enable unified speech understanding and generation, but adapting them to linguistically complex and dialect-rich settings such as Arabic–English remains challenging. We present a controlled study of multi-task instruction tuning for an Arabic-centric audio LLM across generative tasks including ASR and speech and text summarization, and discriminative tasks including dialect and emotion recognition, in a resource-constrained setting. To support end-to-end Arabic speech summarization, we introduce **AraMega-SSum**, a *first* speech summarization resource for training and benchmarking Arabic-centric Audio-LLMs. We compare four training strategies *(i) Uniform Task Mixing*, *(ii) Task-Progressive Curriculum (TPC)*, *(iiii) Aligner-Based Diverse Sampling (ADS)* for training-time batch construction, and *(iv)* A two-stage *TPC→ADS* strategy. Our results show a clear *efficiency–robustness trade-off*. ADS speeds up early convergence and improves paralinguistic performance, however, it hurts other tasks. A two-stage *TPC→ADS* strategy gives the most reliable overall balance across tasks, offering practical guidance for adapting omni audio LLMs to low-resource, dialect-rich environments. We will make **AraMega-SSum** and all experimental resources publicly available to the community.[1]

1 Introduction

Large language models (LLMs) are rapidly evolving into native multimodal systems capable of unified speech understanding and generation (Xu et al., 2025a). However, adapting these "omni" models to linguistically complex and low-resource settings such as the Arabic–English bilingual space remains challenging. This setting combines dialectal variation, frequent code-switching, and paralinguistic cues such as emotion, which are often difficult for general-purpose audio backbones to model effectively. Instruction tuning for this setting should jointly handle *generative* tasks such as automatic speech recognition and summarization, and *discriminative* tasks such as dialect and emotion recognition. In practice, this is further complicated by severe imbalance across tasks and labels. Large ASR corpora can dominate optimization, while smaller paralinguistic datasets and minority classes may remain under-trained, resulting in negative transfer and reduced robustness (Chen et al., 2024).

We study how **data scheduling** and **within-batch task mixing** affect transfer, training stability, and final performance when adapting Qwen2.5-Omni (7B) to an Arabic-centric task suite, while keeping the total number of training steps the same. *Data scheduling* specifies how tasks and examples are sampled over training time. *Within-batch mixing* specifies which tasks, labels, and acoustic conditions co-occur in the same batch.

To support end-to-end Arabic speech summarization, we introduce **AraMega-SSum**, a *first* Arabic speech summarization benchmark developed from translated summaries and synthesized speech. We emphasize that AraMega-SSum is used for the summarization task, while our ASR, dialect, and emotion evaluations are conducted on real speech benchmarks.

We instruction-tune Qwen2.5-Omni (7B) with LoRA and compare four training strategies. *(i)* **Uniform mixing** serves as the standard baseline and samples uniformly from the pooled multi-task data. *(ii)* **Task-Progressive Curriculum (TPC)** introduces tasks gradually, moving from lower-level acoustic objectives to higher-level objectives. *(iii)* **Aligner-Based Diverse Sampling (ADS)** constructs training-time batches that preserve task proportions, balance labels for classification tasks,

[1] anonymous.com

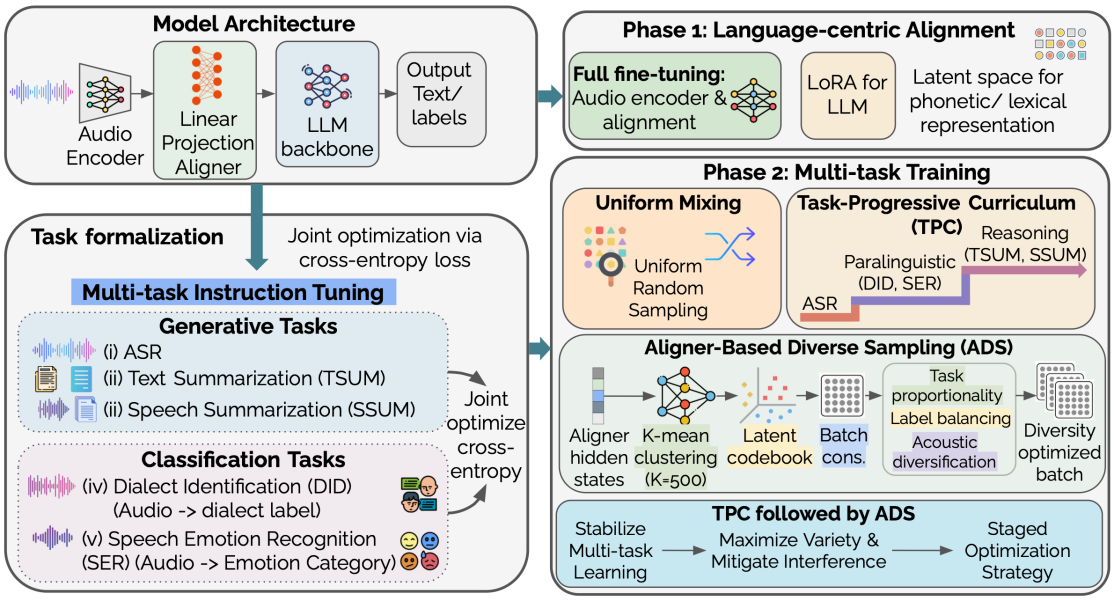

Figure 1: Overview of the proposed methodology. The framework utilizes a Whisper-v3 audio encoder and a Qwen2.5-7B LLM backbone to handle a unified space of generative and classification audio tasks. Training is conducted in two phases: an initial language-centric alignment (Phase 1), followed by a comparative study of four multi-task training strategies (Phase 2), including Uniform Mixing, TPC, ADS, and a hybrid TPC-ADS approach.

and promote cluster diversity in the aligner embedding space, which maps speech encoder representations to the LLM. Finally, *(iv)* a two-stage **TPC→ADS** strategy combines early-stage stabilization with later diversity-oriented training. Our contributions are the following:

- We provide a controlled study of scheduling and batch composition for Arabic-centric omni audio instruction tuning across ASR, summarization, dialect identification, and emotion recognition.
- We introduce **AraMega-SSum**, a novel dataset for Arabic speech summarization.
- We show a consistent efficiency–robustness trade-off across schedules and identify **TPC→ADS** as a reliable balanced strategy for low-resource tasks.

To the best of our knowledge, this is the *first* work to study specialized training strategies for an Arabic-centric omni-model spanning acoustic, linguistic, and paralinguistic tasks. It is also the *first* study to show how diversity in aligner embeddings can guide training-time batch construction while promoting task balance, label balance, and acoustic diversity.

Our findings reveal a clear efficiency–robustness trade-off. TPC improves core acoustic mapping for ASR, but often at the expense of robustness on paralinguistic and data-imbalanced tasks as higher-level objectives are introduced. In contrast, ADS accelerates early convergence, yet tends to saturate earlier than uniform task mixing. Overall, **TPC→ADS** offers the most balanced strategy, combining early-stage stabilization with later diversity-based training, and improving low-resource tasks such as emotion recognition, dialect identification, and Arabic text summarization.

2 Methodology

In Figure 1, we present an overview of the proposed methodology. We study data scheduling and within-batch task mixing for adapting an omni audio LLM to Arabic-centric speech understanding. We utilized the **Qwen2.5-7B-Omni** architecture (Xu et al., 2025a). It combines a Whisper-v3 audio encoder (Radford et al., 2023a) with a 7B transformer decoder. A linear projection layer, which we call the **aligner**, maps encoder features into the LLM embedding space.

2.1 Tasks Formalization

We instruction-tune the model on a unified task, $\mathcal{T} = \{\mathcal{T}_g, \mathcal{T}_c\}$, set that includes generative ($\mathcal{G}$) tasks and classification ($\mathcal{C}$) tasks. Each training example consists of an audio input x (or a transcript for text-only summarization), a task prompt p, and a target output sequence y. We optimize the negative log-likelihood of the target tokens using cross-entropy.

Generative ($\mathcal{G}$) tasks We include automatic speech recognition (ASR), text summarization (TSUM) given the transcript, and speech summarization (SSUM) given the speech input. For TSUM and SSUM, the target is a short summary in Arabic or English depending on the prompt.

Classification ($\mathcal{C}$) tasks We include dialect identification (DID) and speech emotion recognition (SER). We formulate classification in the same instruction format by predicting a canonical label string, so the loss remains token-level cross-entropy.

2.2 Two-phase Adaptation

We use two training phases to separate modality alignment from multi-task scheduling effects.

Phase 1: Language-centric alignment. We fine-tune the audio encoder and aligner on large-scale bilingual ASR to align the audio representation with Arabic and English phonetic and lexical structure. During this phase, the LLM is adapted using LoRA.

Phase 2: Multi-task instruction tuning. We freeze the audio encoder and the aligner, and we update only the LoRA parameters in the LLM. All scheduling strategies in this phase use the same compute budget, meaning the same total number of training steps, so differences are attributable to the sampling strategy.

2.3 Data Scheduling and Batch Construction

A. Uniform Mixing Baseline (UM): We sample training instances uniformly from the pooled multi-task set $\mathcal{D}$. This is a standard baseline for instruction tuning.

B. Task-Progressive Curriculum (TPC): TPC trains in different stages ordered by task abstraction. The model is exposed to tasks in five sequential stages: **Acoustic**: ASR $\rightarrow$ **Paralinguistic**: DID, SER $\rightarrow$ **Reasoning**: TSUM, SSUM. At each stage, we retain a fixed fraction of data from earlier stages to reduce forgetting.

C. Aligner-Based Diverse Sampling (ADS): To mitigate task interference and label imbalance, particularly the dominance of MSA over regional dialects such as Gulf Arabic, we propose the ADS method, presented in Algorithm 1. The method promotes batch diversity by sampling from a discretized latent space obtained through K-means clustering of the aligner's output representations.

Algorithm 1 ADS for Batch creation: Dataset $\mathcal{D} = \{(x_i, y_i, \tau_i, l_i)\}_{i=1}^{N}$, Codebook $\mathcal{C}$, Batch size M

$h_i = A_\phi(x_i)$
$\mathcal{C} \leftarrow$ K-Means$(\{h_i\}, K = 500)$
$\forall i,\ x_i \in \mathcal{D}$ to nearest centroid $c_k \in \mathcal{C}$
while Training **do**
  $\mathcal{B} \leftarrow \emptyset$
  for each task $t \in \mathcal{T}$ **do**
    $n_t \leftarrow M \times PriorDist(t)$
    $\mathcal{L}_t \leftarrow UniqueLabels(t)$
    for each label $l \in \mathcal{L}_t$ **do**
      $\mathcal{S}_{t,l} \leftarrow \{x_i \in \mathcal{D} \mid \text{task} = t, \text{label} = l\}$
      Sample x from $\mathcal{S}_{t,l}$ via *Round-Robin* over assigned clusters $\{c_k\}$
      $\mathcal{B} \leftarrow \mathcal{B} \cup UpsampleMinor(x)$
    end for
  end for
  yield $\mathcal{B}$
end while

Latent representation. We first define a representation space using the hidden states of the linear aligner, A_ϕ. Given that the aligner bridges the acoustic encoder and the LLM, its embeddings capture the specific phonetic-semantic features relevant for downstream reasoning. To minimize computational overhead, we perform clustering on a 3% representative subset of the total corpus. For each sample i, we extract the aligner hidden states and apply max-pooling over the temporal dimension to obtain a fixed-size vector $h_i \in \mathbb{R}^d$. We then apply K-Means clustering to generate a global acoustic-semantic codebook $\mathcal{C}$ with $K = 500$ centroids.

Batch construction and upsampling. During the SFT phase, each "effective batch" $\mathcal{B}$ is constructed to satisfy three constraints:

1. *Task proportionality:* The relative frequency of tasks in batch $\mathcal{B}$ follows the original distribution of the dataset to maintain stable convergence on high-volume tasks like ASR.
2. *Label balancing:* For discriminative tasks (e.g., Emotion, Dialect), we upsample minority classes ($UpsampleMinor(.)$) such that all labels, l, within a task appear with prior frequency per batch.
3. *Acoustic diversification:* Samples are selected using a Round-Robin traversal across the K clusters within each task-label pair. This ensures that the model is exposed to a

maximally diverse set of speakers and acoustic environments in every gradient update.

D. TPC followed by ADS: We also evaluate a two-stage training strategy that runs TPC for an initial portion of training to stabilize the shared representation, then switches to ADS for the remaining steps to emphasize label coverage and cluster diversity. This strategy uses the same total number of steps as the other strategies.

3 Datasets

To facilitate our language-centric alignment and multi-task training objectives, we curate training and evaluation data covering MSA, multiple Arabic dialects, and English. The development of Arabic-centric audio LLMs is challenging due to dataset scarcity. Large-scale ASR data are relatively available, whereas high-level spoken understanding and paralinguistic tasks remain limited and often exhibit strong label imbalance. A major gap is Arabic end-to-end speech summarization, for which paired speech and abstractive summaries are scarce in the public domain.

3.1 AraMega-SSum

To address current gaps in the literature, we introduce and publicly release **AraMega-SSum**, a large-scale Arabic speech summarization dataset with 50,618 training samples (222 hours) and 4,000 test samples (16.98 hours). AraMega-SSum (Arabic Short Speech Summarization) serves as the first end-to-end dataset for semantic *understanding* directly from Arabic audio, providing a foundation for evaluating high-level reasoning in Arabic-centric multimodal systems. A key contribution is the dataset's scale and quality for Arabic spoken short summarization. Given the extreme scarcity of paired audio-to-summary data in Arabic, we construct AraMega-SSum using a semi-synthetic pipeline that preserves semantic consistency while maximizing acoustic diversity. Our goal is to provide a large dataset that enables reproducible Arabic speech summarization evaluation, together with quality controls that reduce noise.

Translation. We leverage the **MegaSSUM** corpus (Matsuura et al., 2024) as our source, a large-scale English sentence-wise speech summarisation dataset based on the Gigaword dataset, comprising more than 3.8 million synthesized speech, transcription, and summary triplets. English speech is generated using a multi-speaker text-to-speech model trained on LibriTTS-R to produce natural-sounding utterances from the first sentences of news articles paired with their headlines, enabling broad coverage and consistent speech-text alignment. We translated the English source transcripts and their corresponding human-written summaries into MSA using **Gemini-2.5-flash**. To ensure the naturalness of the translations, we utilise GPT-4.1 as judge to evaluate the quality of translations with respect to *semantic equivalence*, *information preservation*, *contextual accuracy*, *completeness*, and *coherence*. The prompt is provided in Listing 7. Table 6 (Appendix) reports absolute LLM-as-a-judge scores on a 0--10 scale for EN→AR transfer. Under our current setup and rubric, most outputs are fluent and semantically aligned with the source, leading to a narrow score range with values clustered near the top of the scale. Moreover, we added the same measures with Human-as-a-judge for translation quality. Table 7 show a *similar pattern* with ***human evaluation*** on a subset of 200 pairs.

Audio Synthesis via Neural Voice Cloning. To reconstruct the speech modality, we use **XTTS-v2** (Casanova et al., 2024), a state-of-the-art multi-speaker latent diffusion model for speech synthesis. To reduce the risk of overfitting to static synthetic acoustic profiles, we leverage the model's zero-shot **voice cloning** capability. We use reference speaker audio from *MenaSpeechBank* (Ali et al., 2026) and extract speaker embeddings from **47 distinct speakers** representing **10 Arab countries**. This geographic breadth of the speakers audios helps ensure that the generated speech captures a wide range of regional prosody, pitch variation, and glottal characteristics across the Arab world. For each reference speaker, we maintain at least 10 high-quality voice clips, and randomly sample one reference clip to increase intra-speaker acoustic variability. This approach provides the necessary acoustic diversity for the LLM to generalize across heterogeneous recording conditions.

The resulting dataset, **AraMega-SSum**, contains approximately **50,000** pairs and provides a strong foundation for learning semantic compression.

3.2 Training Datasets

Datasets for language-centric alignment. For the model alignment experiments, we curate a high-quality ASR dataset of ≈**10K hours** from multiple sources, as shown in Table 1. We select

Obj.	Task	Data	Total
		Phase 1: Language-centric Alignment	
ASR	ASR	QASR (Mubarak et al., 2021), MASC (Al-Fetyani et al., 2023), MGB-3/5 (Ali et al., 2017b,a), GigaSpeech (Chen et al., 2021), GALE,[1] TED-LIUM 3 (Hernandez et al., 2018), LibriSpeech (Panayotov et al., 2015), CV (ar/en) (Ardila et al., 2020), SPGISpeech (O'Neill et al., 2021), + in-house	10,000h
ASR	ASR	Augmentation	≈800h
Total			**≈10,800h**
		Phase 2: Multi-task Training	
Gen.	ASR	20% subset of Phase 1	≈1,740h
Gen.	SSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	378h
Gen.	TSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	101,242#
Disc.	DID	ADI-17 (Shon et al., 2020) + ADI-5 (MSA only) (Ali et al., 2017b)	≈2,983h
Disc.	Emo.	ANAD (Klaylat et al., 2018) + EAED (Safwat et al., 2023) + KEDAS (Belhadj et al., 2022) + KSU (Meftah et al., 2021) + BAVED (Aouf, 2020) + YSED (Derhem et al., 2025)	12.17h
Total			**5,113h + 101,242#**

Table 1: Datasets used in each training phase. Obj.denotes task objective, Gen.: generative, Disc.: discriminative, SSUM: speech summarization, TSUM: text summarization, DID: dialect identification, Emo.: emotion recognition, CV: Common Voice. "h" denotes audio duration and "#" denotes the number of samples.

Task	Benchmarks	Metric
ASR (Nat.)	LibriSpeech (Panayotov et al., 2015), MGB-2 (Ali et al., 2016)	WER
ASR (Dial.)	MGB-3 (Ali et al., 2017b), SADA (Alharbi et al., 2024)	WER
ASR (L2/CS)	L2-ARCTIC (Zhao et al., 2018), ESCWA (Chowdhury et al., 2021), DACS (Chowdhury et al., 2020)	WER
Summ.	MegaSUM-SSum (Matsuura et al., 2024), AraMega-SSum (**ours**)	RL, BSc., Lj
Emotion	KSU (Meftah et al., 2021), MELD (Poria et al., 2019)	w-F1
DID	ADI-17 (Shon et al., 2020)	w-F1

Table 2: Evaluation datasets and metrics. Nat. denotes native, Dial.: dialectal, L2: second-language, and CS: code-switching. Summ.: summarization, RL: ROUGE-L, BSc.: BERTScore, Lj: LLM-judge, and w-F1: weighted F1.

these sources to ensure balanced linguistic coverage while maintaining computational efficiency. To improve acoustic robustness under real-world variability, we apply a targeted data augmentation strategy. Specifically, we select a representative 100-hour subset and apply **speed perturbation** (with factors of 0.9× and 1.1×), together with **additive noise augmentation** using the MUSAN (Snyder et al., 2015) and RIR (Ko et al., 2017) corpora. This process yields an additional ≈800 hours of synthetic high-variance audio. The final dataset consists of ≈**10,800 hours**.

Datasets for multi-task training. For the multi-task learning experiments, we selected several datasets covering both generative and discriminative tasks, including ASR, summarization, emotion recognition, and dialect identification. Below, we briefly describe each dataset.

- **ASR:** Given the high-resource nature of ASR compared to downstream semantic tasks, we performed a controlled **subsampling of the ASR dataset**. We selected 20% of the dataset curated for Phase 1.
- **Summarization:** For English, we use the standard training partitions of the Mega-SSum dataset. For Arabic, we use the newly developed **AraMega-SSum** dataset. Examples from the dataset are presented in Appendix Figure 5.
- **Emotion Recognition:** We curated publicly available Arabic affective corpora, which covers spontaneous emotional prosody.
- **Dialect Identification (DID):** We unified the ADI-17 (Shon et al., 2020) corpus with the MSA partitions from the ADI-5 dataset, providing a robust taxonomy covering six regional variants.

Our final multi-task dataset aggregates approximately **5,113** hours of speech. To ensure training stability and optimize memory utilization, we constrained the duration of all speech instances to $t \leq 180$ seconds. Table 1 reports the statistics of the curated datasets. Per-task, label-wise distributions are shown in Figures 4–3 in the Appendix.

3.3 Evaluation Datasets

To assess the multi-task capabilities we curate a comprehensive evaluation suite spanning generative and discriminative tasks from publicly available test sets. Below we describe each of them as in also reported in Table 2.

ASR. We benchmark transcription performance across three dimensions: native English, L2 speech, MSA and dialectal Arabic. For native English, we utilize the **LibriSpeech** (Panayotov et al., 2015) *test-clean* and *test-other* partitions. To assess the model's proficiency with non-native accents, we evaluate on the **L2-ARCTIC** test split (Zhao et al., 2018), specifically filtering for speakers with Arabic as their first language (L1) to measure L2-English phonetic adaptation. For Arabic, we employ the widely recognized **MGB-2** benchmark (Ali et al., 2016) for Modern Standard Arabic (MSA). We further evaluate the models robustness to dialectal variation and code-switching through a suite of challenging datasets:

- **Dialectal Diversity:** We use the **MGB-3** Egyptian test set (Ali et al., 2017b) and the **SADA**

(Alharbi et al., 2024) Saudi dialectal corpus to measure regional linguistic adaptation;

- **Code-Switching (CS):** We evaluate intra-sentential and inter-sentential dynamics using **ESCWA** (Arabic-English CS) (Chowdhury et al., 2021) and **DACS** (within-dialect CS) (Chowdhury et al., 2020).

Overall, this setup supports broad evaluation across English and Arabic, including native, non-native, dialectal, and code-switched speech.

Emotion Recognition (ER). We evaluate emotion recognition on two benchmarks to measure robustness and cross-language transfer:

- **KSU Emotions:** acted Arabic speech (Meftah et al., 2021);
- **MELD:** English conversational audio (Poria et al., 2019), used to assess zero-shot cross-language generalization and prior knowledge retention.

Dialect Identification (DID). We use the **ADI17** test set (Shon et al., 2020) to evaluate classification across 17 Arabic dialects. This serves as a high-granularity evaluation of the model's ability to distinguish subtle phonetic and lexical variations across the Arab world.

Summarization (SSum and TSum). We evaluate summarization using the test sets of **Mega-SSum** for English (Matsuura et al., 2024) and our proposed **AraMega-SSum** for Arabic. These benchmarks measure the models ability to generate concise, faithful summaries from short-form audio. To separate acoustic errors from summarization ability, we also include a *text-only* setting that feeds gold transcripts directly to the LLM, providing an oracle upper bound on summarization performance.

4 Experimental Setup

Baseline. For baseline, we opt for **Qwen2.5-7B-Omni** (Xu et al., 2025a), which combines a Whisper-v3 audio encoder with a linear aligner that projects acoustic features into the 4096-dimensional LLM embedding space.

Two-Phase Training. We train in two phases and keep all hyperparameters fixed across all training strategies so differences are attributable to scheduling and batch construction.

- *Phase 1 Language-centric Alignment (LA).* We perform foundational alignment for 1 epoch using the ASR task. In this phase, the audio encoder and linear aligner are fully fine-tuned while for the LLM, we adapted LoRA based training.
- *Phase 2 Multi-task instruction tuning* We run the four training strategies, UM, TPC, ADS, and TPC+ADS, for ≈10K steps. During this stage, the encoder and aligner remain frozen, and only the LoRA adapters in the LLM are updated. We use a learning rate of 3×10^{-5}, with linear warmup over the first 30% of the first epoch, followed by cosine annealing. Training is distributed across 24 H100 GPUs with a per-device batch size of 2. With 16 gradient accumulation steps, the effective global batch size is 768.
  For ADS, we build a global codebook with $K = 500$ clusters from a 3% subset ($\approx$ 75 hours) of the training data to capture fine-grained paralinguistic variation.

Evaluation We evaluate generative and discriminative tasks with standard metrics. For **ASR**, we report WER, applying Arabic normalization by unifying Alef/Hamza variants and removing diacritics. For **summarization**, we report ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and GPT-4.1 as an LLM judge (Zheng et al., 2023). The judge scores outputs from 1--10 on clarity, conciseness, coherence, completeness, semantic alignment, accuracy, relevance, and information density (see prompts used in Figure 6). For **classification**, we use weighted F1 and normalize label variants (e.g., mapping "KSA" to "Saudi Arabia") during post-processing.

5 Results and Discussion

5.1 Effect of Language-centric Alignment

Our results show that *language-centric alignment* (**LA**) shifts the models acoustic space toward Arabic while largely retaining English performance (Table 3). It also improves generalization to *unseen* tasks, including SSUM, SER, and DID (Tables 4–5). Better TSUM scores from our GPT-4.1 judge further indicate that **LA** produces stronger representations for downstream reasoning. We do not compare with specialized state-of-the-art systems such as Fanar (Team et al., 2025); instead, we focus on available multimodal or audio-LLM baselines, such as Gemini, and compare DID with Althubaiti et al. (2025).

5.2 Effect of Multi-task Instruction Tuning

We compare all four Phase2 strategies under the same backbone, optimization settings, and train-

Model	Arabic ASR			CS		English ASR		L2-EN
	MGB2	**MGB3**	**SADA** (OOD)	**ESCWA** (OOD)	**DACS** (OOD)	**Libri-c**	**Libri-o**	**L2-Arctic (Ar)**
Base	34.02	55.53	79.97	60.51	45.58	4.37	6.71	5.51
LA	13.72	28.89	46.06	**37.01**	21.92	1.46	**2.88**	4.39
UM	12.61	**27.80**	**43.06**	42.39	**20.56**	1.45	2.96	**3.04**
TPC	**12.49**	28.06	49.02	38.96	20.76	1.45	2.93	3.15
ADS	13.63	29.22	45.66	41.72	22.19	**1.43**	3.04	3.65
TPC + ADS	12.77	28.49	44.65	42.02	22.66	1.46	3.01	3.24
Gemini	11.91	21.77	53.79	–	–	4.41	7.98	9.90

Table 3: ASR WER (%) across Arabic (MSA/dialectal), code-switched speech (ESCWA/DACS), English (LibriSpeech), and L2 English (L2-Arctic). Lower is better. Best results are highlighted in **blue** (excluding Gemini). We used Gemini results to show the upperbound performance.

	Arabic			English		
Model	**R-L**	**BERT**	**Judge**	**R-L**	**BERT**	**Judge**
SSUM (Speech Summary)						
Base	10.53	42.24	6.73	28.94	57.63	8.48
LA	23.04	54.28	7.64	23.23	55.86	7.89
UM	35.84	63.65	**7.83**	42.12	66.74	**8.54**
TPC	**35.92**	**63.74**	7.80	**44.49**	**68.83**	8.51
ADS	34.04	62.68	7.54	42.92	68.01	8.40
TPC + ADS	35.37	63.44	6.73	42.31	67.80	8.41
Gemini	24.34	55.28	8.08	24.19	55.07	8.43
TSUM (Text Summary)						
Base	29.14	58.59	8.12	34.36	61.44	8.75
LA	24.63	55.19	8.23	25.54	55.77	8.50
UM	35.17	62.84	**8.00**	43.35	67.28	8.70
TPC	35.37	62.99	7.96	**46.31**	**69.73**	8.67
ADS	37.14	64.19	7.91	45.18	69.34	8.64
TPC + ADS	**38.04**	**64.63**	7.95	45.51	69.46	8.64
Gemini	29.05	58.42	8.29	30.83	58.78	8.82

Table 4: Summarization results for speech (SSUM) and text (TSUM). R-L = ROUGE-L F1 (%), BERT = BERTScore F1 (%), Judge = LLM-judge average (/10). Best scores are highlighted in **blue** (excluding Gemini which is hypothised as upperbound performance).

Model	SER (w-F1)		DID (Acc)
	KSUEmo (Ar)	**MELD** (En-OOD)	**ADI17**
Base	13.71	**52.98**	9.05
LA	24.43	50.61	13.97
UM	59.99	50.82	86.97
TPC	59.39	50.93	87.12
ADS	72.01	40.52	77.25
TPC + ADS	**75.94**	49.02	**87.17**
Gemini/SOTA	19.33	41.76	88.70

Table 5: Paralinguistic results for speech emotion recognition (SER) and dialect identification (DID). SER is reported as weighted F1 (%) and DID as accuracy (%). Best scores are highlighted in **blue** (excluding Gemini/SOTA). Higher is better. For DID we used results presented in (Althubaiti et al., 2025).

ing budget. Thus, performance differences can be attributed to the sampling strategy alone.

Uniform mixing, **UM**, remains the most stable for resource-rich generative tasks (ASR, SSUM). While convergence is slower (Appendix Figure 2), sampling from the natural distribution provides a consistent gradient that reaches a balanced global minimum. However, this approach fails to capture niche paralinguistic nuances, yielding lower F1-scores (in Table 5).

Does a curriculum order improve optimization and transfer? TPC improves early optimization by starting with ASR and introducing higher-level tasks later. It performs well on ASR, but it can reduce robustness on paralinguistic tasks. Delaying DID and SER encourages representations that prioritize lexical reconstruction. Later stages must then adapt to dialectal and emotional cues from a less suitable starting point, which appears as negative transfer in Table 5.

Can within-batch diversity and label balancing improve paralinguistic robustness? ADS improves paralinguistic learning by increasing label coverage and acoustic diversity in each batch, and it accelerates early training progress. At the same time, we observe a systematic cross-task tradeoff. Gains in SER can come with regressions in DID and in some generative metrics. The training curves show higher loss volatility under ADS and a clear loss jump at the TPC→ADS switching point in Appendix Figure 2. For generative tasks that benefit from repeated exposure to canonical patterns, reduced redundancy in ADS batches can also be less favorable during longer training.

Can hybrid strategy balance generative stability and paralinguistic robustness? TPC→ADS combines the strengths of both approaches. The initial TPC phase stabilizes the audio-to-text mapping and provides a strong starting point. The ADS phase then increases label coverage and acoustic diversity, which improves paralinguistic robustness without the larger regressions observed under ADS-only. Across our evaluations, TPC→ADS provides the most consistent balance between generative quality and paralinguistic robustness.

When diversity hurts vs. helps: ADS helps discriminative tasks find class boundaries but hurts generative tasks (such as SSUM and ASR). Generative tasks require canonical, redundant patterns

to learn the "rules" of language; ADS starves the model of these by focusing on exceptions. It is essentially trying to learn the "advanced" task before it is mastering the "basics" knowledge. The low-resource tasks then become highly sensitive to the dominant tasks (e.g., ASR), which can weaken other learning signal.

Does data scheduling matters? Scheduling choices matter under the low-resource and same compute setup. UM and TPC favor generative stability, while ADS improves paralinguistics but can introduce trade-offs. The TPC→ADS schedule is more reliable when the goal is balanced performance across ASR, summarization, dialect, and emotion. For ASR, a dominant and well-represented task, UM is sufficient. For generative tasks such as SSUM, both UM and TPC perform well, underscoring the benefit of stable early training. However, for imbalanced or low-resource tasks, including SER, DID, and Arabic text summarization (Ar-TSUM), **TPC→ADS** provides the best trade-off. Relative to **TPC→ADS**, UM shows only small differences on ASR (average WER gap of ≈ 1.25) and SSUM (Δjudge < 1.0), but substantially larger gaps on SER ($\Delta = 15.95$) and smaller yet consistent gains on DID ($\Delta = 0.2$). These results suggest that **TPC→ADS** improves robustness by combining early-stage stabilization with later diversity-aware sampling.

6 Related Work

Audio and multimodal LLMs. Recent work has extended LLMs to speech and other modalities by aligning pretrained encoders with language decoders. Early models such as SpeechT5 learned multiple speech tasks in a shared encoder–decoder framework (Ao et al., 2022). Later systems, including AudioPaLM, SpeechGPT, and Qwen-Audio, combined LLM backbones with speech tokenization and instruction tuning for unified audio understanding and generation (Rubenstein et al., 2021; Zhang et al., 2023; Chu et al., 2023). More recent omni models further scale this paradigm across modalities (Xu et al., 2025a,b; Rouditchenko et al., 2025). However, prior work gives limited attention to how data mixing affect stability and cross-task generalization in end-to-end audio instruction tuning.

Multi-task training. Multi-task audio instruction tuning is often affected by negative transfer, where dominant tasks suppress weaker or higher-level ones. Octavius addresses this problem with a LoRA-based mixture-of-experts design that reduces cross-task interference (Chen et al., 2024). MINT shows that naive task aggregation can harm generalization and that structured grouping improves transfer across tasks (Shan et al., 2025). Curriculum learning has also been shown to improve optimization when the training order is well chosen (Bengio et al., 2009). These findings motivate a systematic comparison between curriculum-style ordering and diversity-oriented sampling.

Arabic speech understanding. Arabic has benefited from multilingual speech foundation models such as Whisper and XLS-R, as well as LLM-based systems for Arabic ASR and dialect identification such as Octopus (Radford et al., 2023b; Ng et al., 2026; Althubaiti et al., 2025). However, higher-level Arabic speech understanding remains underdeveloped. Arabic summarization research is still largely text-based, for example AraBART (Kamal Eddine et al., 2022). Speech emotion recognition also continues to rely on relatively small task-specific datasets such as BAVED and KSUEmotions (Aouf, 2020; Meftah et al., 2018). Unified instruction-following audio models for Arabic remain largely unexplored.

7 Conclusion

We presented a controlled study of how data mixing and batch construction affect multi-task instruction tuning for an Arabic-centric omni audio LLM. We evaluate both **generative tasks**, including ASR and speech summarization, and **discriminative tasks**, including dialect and emotion recognition. We also introduce **AraMega-SSum**, a new benchmark for end-to-end Arabic speech summarization. Across benchmarks, we find a clear efficiency–robustness trade-off. **ADS** improves paralinguistic performance through better label coverage and acoustic diversity, but can hurt other objectives when used throughout training. **TPC** benefits early ASR learning, but may weaken paralinguistic robustness when higher-level tasks are introduced later. Overall, the two stage **TPC→ADS** strategy provides the best balance, combining stable acoustic grounding with later diversity-oriented refinement. These results show that data mixing is a key design choice for low-resource, dialect-rich omni audio tuning. Future work will test generality across AudioLLMs, budgets, and fine-tuning settings.

Limitations

Our study focuses on multi-task audio instruction tuning for a single AudioLLM, Qwen2.5-Omni 7B, in an Arabic–English setting. While this setup reflects a realistic and challenging deployment scenario, the findings may not fully generalize to models with substantially different architectures, scales, or training objectives. In particular, models with explicit task-specific heads or more decoupled audio–text pathways may respond differently to data scheduling strategies.

We evaluate scheduling strategies under a fixed training and a parameter-efficient fine-tuning setup. This choice is intentional and reflects practical compute constraints, but different conclusions may emerge under substantially larger budgets or full-parameter fine-tuning.

We compare four training strategies: uniform mixing, TPC, ADS, and TPC→ADS. However, due to limited compute resources, we do not exhaustively evaluate all sampling combinations and ADS ablations, such as per-task uniform and label-balanced-only variants.

Societal/Broader Impact

This work can improve Arabic speech technology in dialect-rich, code-switched settings, supporting accessibility (e.g., captions), education, and media/knowledge access with better transcription and spoken summarization. However, societal risks exist. Performance may vary across dialects, accents, and speaking styles, potentially creating unequal error rates and exclusion. Paralinguistic tasks (emotion/dialect recognition) can be misused if treated as definitive signals, and should not inform consequential decisions without strong validation and human oversight. Finally, AraMega-SSum relies on translation and TTS/voice cloning, which raises source-traceability and misuse concerns (e.g., impersonation) if not clearly documented. We therefore emphasize responsible release and use: transparent dataset documentation (licensing, consent, and synthetic labeling), subgroup evaluation, and explicit restrictions against surveillance or high-stakes deployment without safeguards.

Ethical Considerations

Our work leverages large-scale Arabic and English speech corpora, including synthetic audio generated via neural voice cloning. While these methods enable the creation of diverse and balanced datasets, they raise important ethical considerations. First, the use of voice cloning and synthesized speech could potentially be misused for impersonation, fraud, or other malicious purposes. Second, even when using publicly available or synthetic data, issues of privacy and consent must be carefully managed, particularly when real speaker recordings are involved. We mitigate these concerns by curating datasets with explicit permissions where applicable, ensuring anonymization of speaker identities, and restricting dataset release to research purposes only. Future work should continue to address these risks, especially in low-resource and dialect-rich contexts where consent practices may vary.

References

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006--1013.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286--10290.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279--284.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017a. The MGB-5 challenge: Arabic multi-dialect broadcast media recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316--322. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017b. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic*

Speech Recognition and Understanding Workshop (ASRU), pages 316--322.

Zien Sheikh Ali, Hunzalah Hassan Bhatti, Rabindra Nath Nandi, Shammur Absar Chowdhury, and Firoj Alam. 2026. MENASpeechBank: A reference voice bank with persona-conditioned multi-turn conversations for audio-llms. *arXiv preprint arXiv:2602.07036*.

Sara Althubaiti, Vasista Sai Lodagala, Tjad Clark, Yousseif Ahmed Elshahawy, Daniel Izham, Abdullah Alrajeh, Aljawahrah Bin Tamran, and Ahmed Ali. 2025. Octopus: Towards building the Arabic speech LLM suite. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 425--435, Suzhou, China. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723--5738, Dublin, Ireland. Association for Computational Linguistics.

Ali Aouf. 2020. Basic arabic vocal emotions dataset (baved). Kaggle dataset. Accessed 2026-01-06.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218--4222.

Mourad Belhadj, Ilham Bendellali, and Elalia Lakhdari. 2022. Kedas: A validated arabic speech emotion dataset. In *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*, pages 1--6.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41--48. ACM.

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 *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*. ISCA.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. 2024. Octavius: Mitigating task interference in mllms via lora-moe. *Preprint*, arXiv:2311.02684.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards One Model to Rule All: Multilingual Strategy for Dialectal Code-Switching Arabic ASR. In *Proc. of INTERSPEECH*.

Shammur Absar Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020. Effects of dialectal code-switching on speech modules: A study using Egyptian Arabic broadcast speech. In *Interspeech*.

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. *Preprint*, arXiv:2311.07919.

Somia Derhem, Eiad AL-Mekhlafi, Nashwan Ahmed AL-Majmar, and Moeen AL-Makhlafi. 2025. Ysed: Yemeni speech emotion dataset. *Data in Brief*, 63:112233.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments*

on Speaker Adaptation, page 198208. Springer International Publishing.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31--42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Samira Klaylat, Ziad Osman, Rached Zantout, and Lama Hamandi. 2018. Arabic natural audio dataset.

Tom Ko, Vijayaditya Peddinti, Michael Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. pages 5220--5224.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74--81.

Koki Matsuura, Takuya Ashihara, Tetsuro Moriya, Mamoru Mimura, Takafumi Kano, Atsushi Ogawa, and Marc Delcroix. 2024. Sentence-wise speech summarization: Task, datasets, and end-to-end modeling with lm knowledge distillation. In *Proceedings of Interspeech 2024*, pages 1945--1949.

Ali H. Meftah, Yousef A. Alotaibi, Sid-Ahmed Selouani, Mustafa A. Qamhan, and Mohammed Zakariah. 2018. Evaluation of an arabic speech corpus of emotions: A perceptual and statistical analysis. *IEEE Access*, 6:72845--72861.

Ali Hamid Meftah, Mustafa A. Qamhan, Yasser Seddiq, Yousef A. Alotaibi, and Sid Ahmed Selouani. 2021. King saud university emotions corpus: Construction, analysis, evaluation, and comparison. *IEEE Access*, 9:54201--54219.

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.

Kwok-Ho Ng, Tingting Song, Yongdong WU, and Zhihua Xia. 2026. Xlsr-mambo: Scaling the hybrid mamba-attention backbone for audio deepfake detection. *Preprint*, arXiv:2601.02944.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *Preprint*, arXiv:2104.02014.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527--536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023a. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492--28518. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492--28518. PMLR.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalan Borsos, Felix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2021. Audiopalm: A large language model that can speak and listen (2023). *arXiv preprint arXiv:2306.12925*.

Sarah Safwat, Mohammed Salem, and Nada Sharaf. 2023. Egyptian arabic emotional dataset (eaed).

Xiaojun Shan, Qi Cao, Xing Han, Haofei Yu, and Paul Pu Liang. 2025. Mint: Multimodal instruction tuning with multimodal interaction grouping. *Preprint*, arXiv:2506.02308.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020 - Proceedings*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 8244--8248, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2020 IEEE.; 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020 ; Conference date: 04-05-2020 Through 08-05-2020.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Multipurpose Corpus for Music and Noise. In *arXiv preprint arXiv:1510.08484*.

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. *Preprint*, arXiv:2501.13944.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, page 27832787.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595--46623.

Appendix

A Translation Quality for AraMega-SSum

In Table 6, we present English-to-Arabic summary translations achieve near-perfect evaluation scores from a GPT-4.1 judge, consistently exceeding 9.95 out of 10 across all measured quality metrics, including semantic equivalence and coherence. On the human evaluation we observe similar scores in all rubrics as shown in Table 7.

Metric (EN→AR transfer)	Score (/10)
Semantic Equivalence	9.960
Information Preservation	9.958
Contextual Accuracy	9.960
Completeness	9.956
Coherence	9.987

Table 6: LLM-as-a-judge (GPT-4.1) scores for English→Arabic summary translation by Gemini. Scores are out of 10.

Metric (EN→AR transfer)	Score (/10)
Semantic Equivalence	9.931
Information Preservation	9.950
Contextual Accuracy	9.990
Completeness	9.950
Coherence	10.0

Table 7: Human scores for English→Arabic translation quality. Scores are out of 10.

B Training Loss

Figure 2 compares training loss trajectories for different multi-task training regimes over three epochs. Stochastic Multi-tasking (SM, red) and Task-Progressive Curriculum (TPC, green) exhibit smooth and stable convergence. Aligner-Based Diverse Sampling (ADS, brown) starts with a substantially higher loss but rapidly converges, reaching competitive loss values early in training. Hybrid strategies that transition from SM or TPC to ADS (orange and blue, respectively) show brief loss spikes at the regime switch, followed by continued loss reduction. By step 7,600, single-regime training approaches have converged, while multi-regime schedules achieve lower final loss values following the transition to ADS.

C Dataset Details

Figure 3 presents the distribution of emotion labels across multiple Arabic emotion datasets used for training, including KEDAS, EAED, KSU, YSED, BAVED, and ANAD. Each dataset exhibits a distinct class composition, with emotions such as anger, happiness, neutrality, and sadness consistently represented, while fear, questioning, and surprise appear less frequently and vary across datasets. KEDAS and EAED contribute the largest number of samples with a broad coverage of emotion categories.

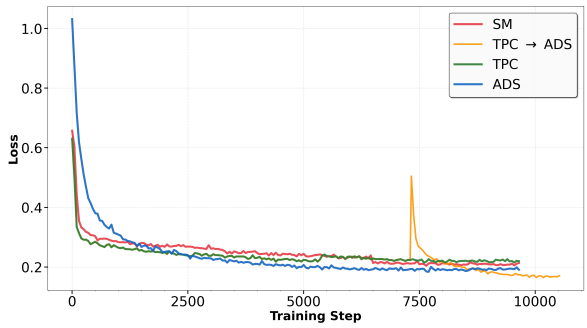

Figure 2: Training loss comparison across multi-task training setup.

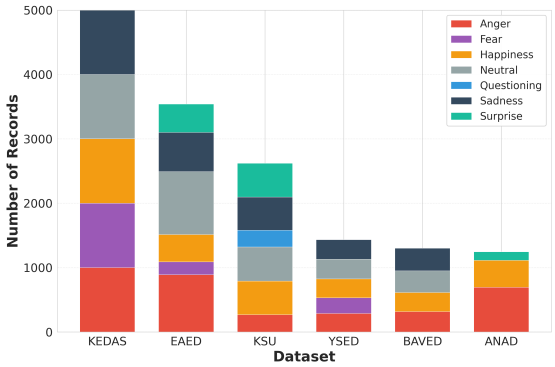

Figure 3: Emotion label distribution across training datasets.

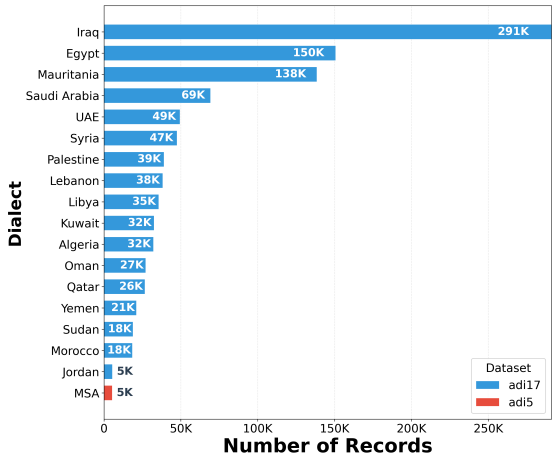

Figure 4: Dialect distribution in the training set.

Figure 4 shows the training data covers dialects across 17 countries using the **ADI-17** dataset, along with MSA from **ADI-5**. The distribution is dominated by Egyptian, Iraqi, and Mauritanian dialects, while the remaining dialects are represented with varying but comparatively smaller sample sizes.

In Figure 5, presents example samples from the MegaSSUM dataset, showing English transcriptions and summaries, alongside corresponding Arabic transcriptions and summaries from the proposed AraMega-SSum dataset. Each instance is paired with both English and Arabic speech through text-to-speech generation, enabling cross-

English Transcription *Mega-SSum*	the yugoslav justice ministry on friday issued subpoenas to a total of ## leaders of nato and its major member countries , including u.s. president bill clinton , ordering them to appear in court on war crimes charges
English Summary *Mega-SSum*	yugoslav court issues subpoenas to nato leaders
Arabic Transcription *AraMega-SSum*	أصدرت وزارة العدل اليوغوسلافية يوم الجمعة مذكرات استدعاء إلى ما مجموعه 74 من قادة ناتو ودوله الأعضاء الرئيسية، بمن فيهم الرئيس الأمريكي بيلل كلينتون، تأمرهم فيها بالمثول أمام المحكمة بتهم ارتكاب جرائم حرب
Arabic Summary *AraMega-SSum*	محكمة يوغوسلافية تصدر أوامر استدعاء بحق قادة ناتو

English Transcription *Mega-SSum*	a man from the italy 's south city of bari committed suicide in vatican 's famous basilica of st. peter in rome thursday , shooting himself in the head while being filmed by an australian tourist
English Summary *Mega-SSum*	a man suicides in vatican 's basilica of st. peter
Arabic Transcription *AraMega-SSum*	رجل من مدينة باري جنوب إيطاليا انتحر في كنيسة القديس بطرس الشهيرة في الفاتيكان بروما يوم الخميس، مطلقاً النار على رأسه بينما كان يصوره سائح أسترالي
Arabic Summary *AraMega-SSum*	رجل ينتحر في كاتدرائية القديس بطرس بالفاتيكان

Figure 5: Sample speech summarization instances from AraMega-SSum.

lingual and cross-modal speech summarization training and evaluation.

D Prompts

In Figures 6 and 7 and Listing 8, we present the prompts used for LLM-as-judge summarization evaluation, LLM-as-judge translation evaluation and multitask training.

```
SYSTEM_PROMPT = """You are a reference-grounded
    summarization quality evaluator.
Grade a predicted summary by comparing it with
    the reference summary.
Do NOT use outside knowledge. Judge in the same
    language as the summaries
(Arabic or English). Return ONLY a valid JSON
    object that matches the
schema exactlyno extra text.

Score each criterion as an INTEGER from 1 to 10
    (1=poor, 10=excellent):

- Clarity
- Conciseness
- Coherence
- Completeness
- Semantic_Alignment
- Accuracy
- Relevance
- Information_Density

"""

USER_PROMPT_TEMPLATE = """Language: {language}
Reference summary: {reference_summary}
Predicted summary: {predicted_summary}

Evaluate the predicted summary by comparing it
    with the reference summary.
Output only the JSON described in the system
    prompt."""
```

Figure 6: **Prompt used for LLM-as-a-judge summarization evaluation.**

```
SYSTEM_PROMPT = """You are an expert translation evaluator.

You will be given:
- An English transcription (may contain anonymized tokens such as ###, ####,
  and other placeholders)
- An Arabic transcription translated from the same English content

Your task is to evaluate how semantically equivalent the two transcriptions are.

Evaluation Rules (STRICT):
1. Ignore all anonymization tokens in English.
2. Ignore number differences caused by anonymization.
3. Ignore name differences and transliteration variations.
4. Be lenient with phonetic spellings in Arabic.
5. Focus only on semantic meaning and core information.
6. Judge based on events, facts, actions, relationships, and intent.

Score each criterion as an INTEGER from 1 to 10:

- semantic_equivalence
- information_preservation
- contextual_accuracy
- completeness
- coherence

Return ONLY a valid JSON object matching the schema exactly.
"""

USER_PROMPT_TEMPLATE = """Arabic translation:
{arabic_transcription}

Original English source:
{english_transcription}

Evaluate the quality of the Arabic translation relative to the English source.
Follow the evaluation rules defined in the system prompt.
Return only the required JSON output."""
```

Figure 7: **Prompt used for LLM-as-a-judge translation evaluation.**

```
System
You are an Arabic audio-language assistant. You will receive audio and a user instruction
    describing the task. Your capabilities include: (1) Accurate speech transcription with proper
    formatting, (2) Dialect and emotion identification from speech, (3) Understanding
    conversational context and dialogue acts, (4) Generating concise, coherent summaries from both
     speech and text. Always follow instructions precisely, maintain linguistic accuracy, and
    format outputs exactly as requested. For ASR/transcription tasks, output the spoken words
    verbatim in Arabic as written, preserving numbers, names, and code-switching without
    paraphrasing or summarizing.

ASR
<audio>
Task: Transcription.
Transcribe the audio accurately in its original language.
Respond with a single-line JSON object only:
{"transcription":"<text>"}

Dialect
<audio>
Task: Dialect Identification.
Identify the dialect spoken in the audio.
Choose exactly one label from the following list:
Algeria, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Modern Standard Arabic, Morocco,
    Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, United Arab Emirates, Yemen

Respond with a single-line JSON object only:
{"dialect":"<dialect>"}

Emotion
<audio>
Task: Emotion Identification.
Identify the primary emotion expressed in the audio.
Choose exactly one emotion from the following list:
Anger, Fear, Happiness, Neutral, Questioning, Sadness, Surprise

Respond with a single-line JSON object only:
{"emotion":"<emotion>"}

SSUM
<audio>
Task: Speech Summarization.
Summarize the main content of the audio concisely.
Preserve the original language of the speech.

Respond with a single-line JSON object only:
{"summary":"<text>"}

TSUM
Task: Text Summarization.
Read the following text: "{text}"

Summarize it concisely in the same language.
Respond with a single-line JSON object only:
{"summary":"<text>"}
```

Figure 8: **System and user prompts used for multi-task training.**

,