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

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 Mixing (UM)*, *(ii) Task-Progressive Curriculum (TPC)*, *(iii) Aligner-Based Diverse Sampling (ADS)* for training-time batch construction, and *(iv)* A two-stage *TPC→ADS* strategy. Our results reveal a clear *efficiency–robustness trade-off*. TPC yields the strongest generative performance on ASR and summarization. ADS improves paralinguistic tasks but reduces generative stability when used alone. The two-stage *TPC→ADS* strategy provides the best overall balance, delivering the strongest DID and SER results while outperforming large proprietary models such as Gemini-2.5-Pro on discriminative tasks. 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, 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 ASR 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. *Data scheduling* controls how tasks and examples are sampled over training time, whereas *within-batch mixing* controls which tasks, labels, and acoustic conditions co-occur in the same batch.

Our main experiments use **Qwen2.5-Omni-7B** (Xu et al., 2025a), a dense open-weight omni model and our primary Arabic-centric adaptation target. We additionally include **Gemma-4-E4B-it**[2] as a secondary contrastive model to test whether the observed scheduling trends persist under a different architecture with stronger Arabic priors, including prior Arabic ASR exposure. Thus, Qwen captures the main low-resource Arabic adaptation setting, while Gemma serves as an architecture-sensitivity check.

To support end-to-end Arabic speech summarization, we introduce **AraMega-SSum**, a *first* Arabic speech summarization dataset for training and benchmarking Arabic-centric AudioLLMs. It contains 49,995 training samples and 4,000 test samples, corresponding to 222 and 16.98 hours

[1] anonymous.com

[2] https://huggingface.co/google/gemma-4-E4B-it

of speech, respectively. The dataset is built from MSA-translated summaries and speaker-conditioned synthesized speech from 47 speakers across 10 Arab countries. A subset of the data is human-evaluated. AraMega-SSum is used only for summarization, while ASR, dialect identification, and emotion recognition are evaluated on real-speech benchmarks.

We instruction-tune Qwen2.5-Omni-7B and Gemma-4-E4B-it with LoRA and compare four training strategies. *(i)* **Uniform Mixing (UM)** is the standard baseline, sampling 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)** preserves task proportions, balances labels for classification tasks, and *promotes cluster diversity* in the **aligner embedding space** that maps speech encoder representations to the LLM. *(iv)* **TPC→ADS** combines early-stage curriculum stabilization with later diversity-oriented training.

Our contributions are the following:

- We provide a controlled and comparative study of scheduling and batch composition for Arabic-centric omni audio instruction tuning across Generative tasks: *ASR and summarization*, and Discriminative tasks: *dialect identification, and emotion recognition*.
- We introduce **AraMega-SSum**, a novel and first dataset for Arabic speech summarization, with human evaluation on a subset of the dataset.
- We show a consistent efficiency–robustness trade-off across training schedules. **TPC→ADS** provides a reliable balance for discriminative tasks with imbalanced classes, whereas **TPC** is more effective for generative 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. We emphasize that our contribution is not generic multi-task data mixing, but a controlled study of how task scheduling and batch construction affect Arabic-centric AudioLLM adaptation under limited resource.

Our findings reveal a consistent trade-off across training schedules, with variation across model families. For **Qwen2.5-Omni-7B**, TPC gives the best ASR and matches UM on summarization, confirming that curriculum ordering stabilizes acoustic-to-text mapping. For **Gemma-4-E4B-it**, UM gives the best ASR, likely because its multilingual pretraining already provides a strong Arabic prior, reducing sensitivity to curriculum ordering. For discriminative tasks, TPC→ADS gives the best DID for both models and the best SER for Qwen2.5-Omni-7B, with a notable $+16.6$ gain over TPC, showing that diversity-oriented sampling is most effective for paralinguistic tasks once the shared representation has been stabilized. Overall, **TPC** is preferable for generative tasks, while **TPC→ADS** provides the strongest balance for discriminative tasks with imbalanced classes.

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* and *Gemma-4-E4B-it* architecture. For Qwen2.5-7B-Omni combines a Whisper-v3 audio encoder (Radford et al., 2023) with a 7B transformer decoder. A linear projection layer, called the **aligner**, maps encoder features into the LLM embedding space. We did the same for Gemma-4-E4B-it with USM-style audio encoder (Zhang et al., 2023b).

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

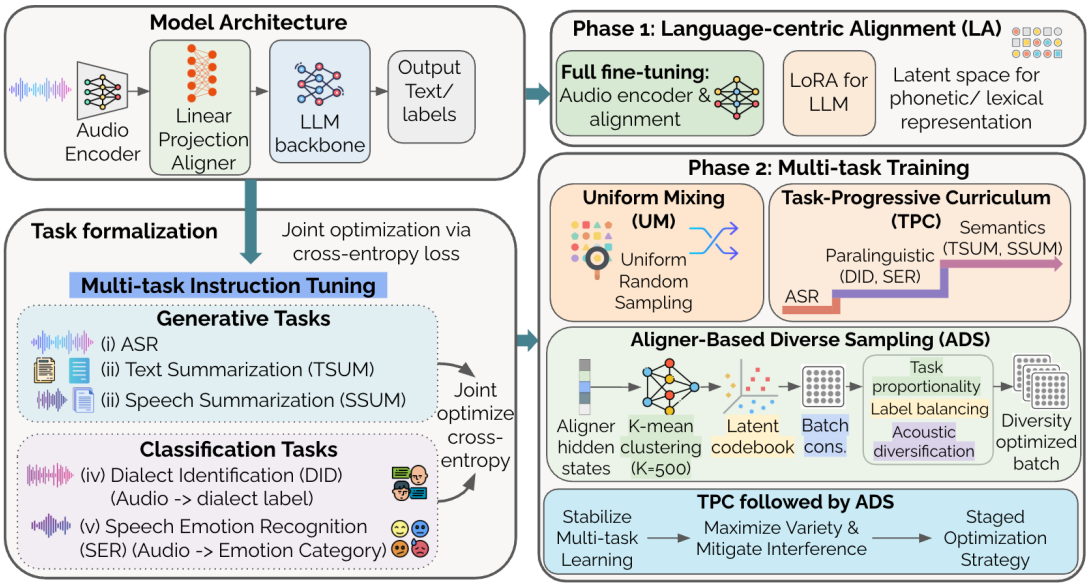


Figure 1: Overview of the proposed methodology. The framework adapts two AudioLLM backbones for a unified set of generative and classification audio tasks. The Qwen-based model uses a Whisper-v3 audio encoder with a Qwen2.5-7B LLM backbone, while the Gemma-based model uses a USM-style encoder with a Gemma-4 LLM backbone. Training proceeds in two stages: language-centric alignment (Phase 1), followed by multi-task adaptation (Phase 2), where we compare UM, TPC, ADS, and the two-stage TPC-ADS strategy.

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

In this section, we compare four strategies under the same model, optimization settings, and fixed computational setup.

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.

Algorithm 1 ADS batch construction

$h_i \leftarrow A_\phi(x_i)$ for all $x_i \in \mathcal{D}$
$\mathcal{C} \leftarrow \mathrm{KMeans}(\{h_i\}, K = 500)$
Assign each x_i to nearest centroid $c_k \in \mathcal{C}$
while training **do**
  $\mathcal{B} \leftarrow \emptyset$
  for each task $t \in \mathcal{T}$ **do**
    $n_t \leftarrow M \cdot \mathrm{PriorDist}(t)$
    for each label $l \in \mathrm{UniqueLabels}(t)$ **do**
      $\mathcal{S}_{t,l} \leftarrow \{x_i : \tau_i = t,\ l_i = l\}$
      Sample $x \sim \mathcal{S}_{t,l}$ by round-robin over clusters
      $\mathcal{B} \leftarrow \mathcal{B} \cup \mathrm{UpsampleMinor}(x)$
    end for
  end for
  yield $\mathcal{B}$
end while

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 → **Paralinguistic**: DID, SER → **Semantic abstraction**: 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.

Latent representation. We first define a repre-

sentation 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 semantic understanding. 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. AraMega-SSum is the only newly developed semi-synthetic dataset in our study, introduced to support this missing Arabic speech summarization setting. The remaining datasets for ASR, emotion recognition, and dialect identification are curated from public real-speech benchmarks.

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 49,995 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 generative understanding 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*.

Translation Quality. 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 Appendix Listing 7. Appendix Table 4 reports LLM-as-a-judge scores on a 0--10 scale. Moreover, we also performed a human evaluation to validate the translation quality.

Table 5 shows a *similar pattern* based on ***human evaluation*** of a subset of 200 pairs. Under our current setup and rubric, most outputs are fluent and semantically aligned with the source.

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 **≈50K** pairs and provides a strong foundation for learning semantic representation.

Audio Quality. We validated the generated speech using both human assessment and standard automatic quality metrics. Three paid annotators rated 100 randomly sampled test-set clips for *naturalness*, *audio quality*, and *intelligibility* (Ju et al., 2024; Du et al., 2025). The clips achieved average scores of 3.98, 4.37, and 3.69 out of 5, respectively, with substantial-to-near-perfect annotator agreement (78.9 to 95.0) measured by Gwet's AC1 (Gwet, 2008) (Appendix Table 6). Complementary metrics show that the generated speech is generally intelligible and preserves speaker characteristics: Fanar Aura ASR obtains a WER of 15.5, NISQA gives an overall score of 3.95 out of 5 and speaker cosine similarity with the reference speakers is 0.66.

3.2 Training Datasets

Datasets for language-centric alignment. For the model alignment experiments, we curate a high-quality ASR dataset of **≈15K hours** from multiple sources, as shown in Appendix Table 2. We select 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 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 on a subset. This process yields an additional ≈15,000 hours of synthetic high-variance audio. The final dataset consists of **≈27K 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 6.44% 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 **5,109** 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 2, in the Appendix, 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 Appendix Table 3.

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., 2017) 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 **KSU Emotions:** acted Arabic speech (Meftah et al., 2021)

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 comparisons, we include both open and proprietary models. As an open baseline, we evaluate Qwen2.5-Omni-3B. For proprietary baselines, we compare against GPT-audio, GPT-5-chat, and Gemini-2.5-Pro, all evaluated in zero-shot mode without any task-specific adaptation.

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.
  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 (SER and DID)**, we use weighted F1 and normalize label variants (e.g., mapping "KSA" to "Saudi Arabia") during post-processing.

5 Results and Discussion

5.1 Results Across Tasks

In Table 1, we report results, averaged over test-sets, across adaptation phases. Phase 2 multi-task training improves both models over the base and LA settings. The gains are clearest for DID, SER, and summarization.

Appendix E provides the full per-dataset breakdown. Tables 7–10 report detailed results for ASR, SSUM and TSUM, SER, and DID. Table 9 includes LLM-as-judge results for summary tasks.

Figure 2 summarizes the same trends across task families. Panel (a) shows that adaptation sharply reduces ASR WER for Qwen2.5-Omni-7B. UM, TPC, and TPC→ADS reach simi-

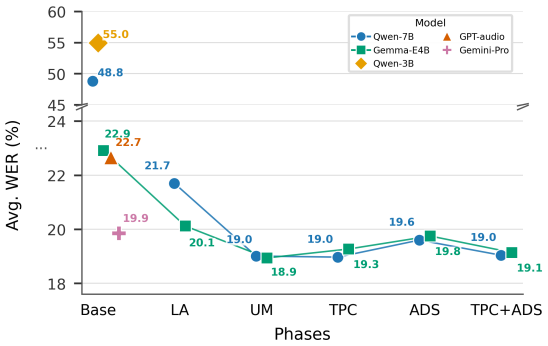
Phase / Model	Generative		Discriminative	
	ASR ↓	Summ. ↑	DID ↑	SER ↑
Qwen2.5-Omni-7B				
Base	48.79	55.0	10.3	13.71
LA	21.70	54.3	9.8	24.43
UM	19.01	**66.9**	87.8	59.99
TPC	**18.97**	**66.9**	87.8	59.39
ADS	19.60	64.9	86.7	72.01
TPC→ADS	19.04	66.6	**88.8**	**75.94**
Δ TPC vs. UM	**−0.04**	0.0	0.0	−0.6
Δ TPC→ADS vs. TPC	+0.07	−0.3	**+1.0**	**+16.6**
Gemma-4-E4B-it				
Base	22.91	54.3	21.2	–
LA	20.12	46.2	7.8	–
UM	**18.94**	67.1	78.8	–
TPC	19.27	**67.2**	78.6	–
ADS	19.75	66.0	75.9	–
TPC→ADS	19.14	66.9	**80.0**	–
Δ TPC vs. UM	+0.33	**+0.1**	−0.2	–
Δ TPC→ADS vs. TPC	**−0.13**	−0.3	**+1.4**	–
Reference Models (no fine-tuning)				
Qwen2.5-Omni-3B	54.95	46.4	7.7	34.4
GPT-audio	22.66	50.2	38.4	46.2
Gemini-2.5-Pro	19.85	58.0	63.4	59.3

Table 1: Results across generative and discriminative tasks by phases. ASR reports WER (↓); summarization: average BERTScore F1; DID and SER report weighted-F1 (↑). SER is evaluated on Qwen2.5-Omni-7B only. The shaded Δ rows decompose scheduling effects in two steps: Δ TPC vs. UM isolates the gain from curriculum ordering over the standard baseline, and Δ TPC→ADS vs. TPC isolates the additional effect of diversity-oriented sampling. **Teal**: improvement; orange: degradation.

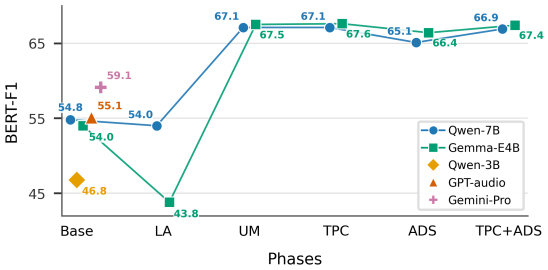
lar WER values near 19. Panel (b) shows that UM and TPC give the strongest summarization performance. ADS improves over the base models, but it remains below UM and TPC. Panel (c) shows a similar pattern for DID and SER. TPC→ADS gives the strongest Qwen2.5-Omni-7B results on both tasks. These trends support a clear trade-off. UM and TPC provide stronger generative stability, while TPC→ADS provides stronger discriminative robustness.

5.2 Effect of Language-Centric Alignment

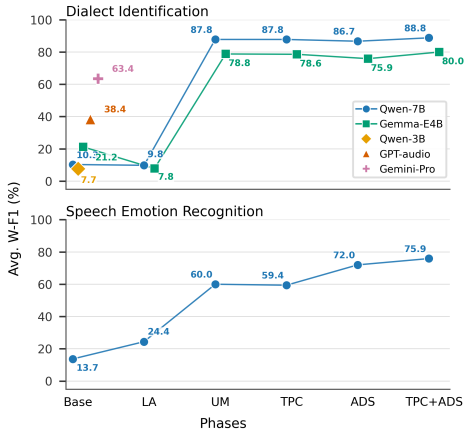
LA gives a large ASR gain. It reduces Qwen2.5-Omni-7B WER from 48.79 to 21.70. It also reduces Gemma-4-E4B-it WER from 22.91 to 20.12. This shows that Phase 1 improves language-dependent acoustic and lexical alignment. However, LA alone does not solve higher-level speech understanding. DID and SER remain weak after LA, and summarization shows limited or negative gains. We therefore treat LA as a necessary alignment step, not as a complete adaptation strategy. Phase 2 multi-task instruction tuning is needed for semantic and paralinguistic behavior.



(a) ASR WER



(b) Summarization BERTScore F1



(c) DID and SER macro-F1

Figure 2: Phase-level results across task families. Panel (a) reports average ASR WER across MSA, dialectal Arabic, code-switched speech, native English, and L2 English. Panel (b) reports average summarization BERTScore F1 over SSUM and TSUM. Panel (c) reports DID and SER macro-F1. Lower is better for WER. Higher is better for all other metrics.

5.3 Effect of Phase 2 Data Scheduling

We compare UM, TPC, ADS, and TPC→ADS under the same model, optimization settings, and training budget. This design isolates the effect of data scheduling and batch construction.

Generative tasks. UM and TPC provide the strongest results for ASR and summarization. UM benefits from stable exposure to the natural task distribution. TPC starts with ASR and gives the model a strong audio-to-text mapping before adding higher-level tasks. This helps both ASR

and summarization. ADS also improves over the base setting, but it usually trails UM and TPC on generative tasks. This suggests that ASR and summarization benefit from repeated exposure to common acoustic and lexical patterns.

Discriminative tasks. DID and SER benefit more from label coverage and acoustic diversity. TPC→ADS gives the best DID and SER results for Qwen2.5-Omni-7B. For Gemma-4-E4B-it, the TPC→ADS gives the best DID score. We keep SER for Gemma as a future task. These results show that diversity helps discriminative tasks, but its effect depends on the model and metric. They also show why ADS works better after TPC than as a standalone schedule. TPC first stabilizes the shared representation. ADS then improves coverage of labels, speakers, and acoustic conditions.

Overall trade-off. Scheduling choices have a measurable impact under limited resources. UM and TPC are preferable when ASR or summarization is the primary objective, as both strategies provide stable generative performance with minimal tuning. TPC→ADS is the most reliable choice when discriminative robustness matters, combining early stabilization with later diversity-aware sampling. Notably, TPC→ADS outperforms large proprietary reference models on discriminative tasks despite operating at a fraction of their scale, surpassing Gemini-2.5-Pro on both DID (88.8 vs. 63.4) and SER (75.94 vs. 59.3), while remaining competitive on ASR and summarization.

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., 2023; Zhang et al., 2023a; 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 (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 and KSUEmotions (Meftah et al., 2018). Unified instruction-following audio models for Arabic remain largely unexplored.

7 Conclusion

We presented a controlled study of data scheduling and batch construction for multi-task instruction tuning of Arabic-centric omni audio LLMs, covering ASR, speech summarization, dialect identification, and emotion recognition. We also introduced **AraMega-SSum**, the first Arabic speech summarization benchmark. Our results confirm a clear efficiency–robustness trade-off: TPC provides the strongest generative performance, ADS improves paralinguistic robustness but hurts generative stability when used alone, and TPC→ADS delivers the best overall balance across tasks. Notably, TPC→ADS outperforms large proprietary models including Gemini-2.5-Pro on discriminative tasks, demonstrating that targeted scheduling can compensate for scale in low-resource, dialect-rich settings. Future work will extend this study across diverse AudioLLM architectures, compute budgets, and fine-tuning configurations.

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.

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. 2025. OASIS: A multilingual and multimodal framework for culturally grounded spoken visual qa. *arXiv preprint arXiv:2510.06371*.

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, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve

Renals, and Khalid Choukri. 2019. The MGB challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026--1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. 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.

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 *Interspeech 2024*, pages 4978--4982.

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, Huayang Liu, Zhenfei Yin, Si Liu, Lyu Sheng, Wanli Ouyang, Jing Shao, and 1 others. 2024. Octavius: Mitigating task interference in mllms via lora-moe. In *International Conference on Learning Representations*, volume 2024, pages 24652--24673.

Shammur A. 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 2020*, pages 2382--2386.

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*.

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*.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui

Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

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.

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.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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, Barcelona, Spain. Association for Computational Linguistics.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Masato Mimura, Takatomo Kano, Atsunori Ogawa, and Marc Delcroix. 2024. Sentence-wise Speech Summarization: Task, Datasets, and End-to-End Modeling with LM Knowledge Distillation. In *Interspeech 2024*, pages 1945--1949.

Ali Hamid Meftah, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. 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.

Patrick K. ONeill, 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. In *Interspeech 2021*, pages 1434--1438.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206--5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 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, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. AudioPaLM: A large language model that can speak and listen. *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. *arXiv 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 *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244--8248.

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

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. *arXiv 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. *arXiv preprint arXiv:2509.17765*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. 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*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, and 1 others. 2023b. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

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 *Interspeech 2018*, pages 2783--2787.

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.

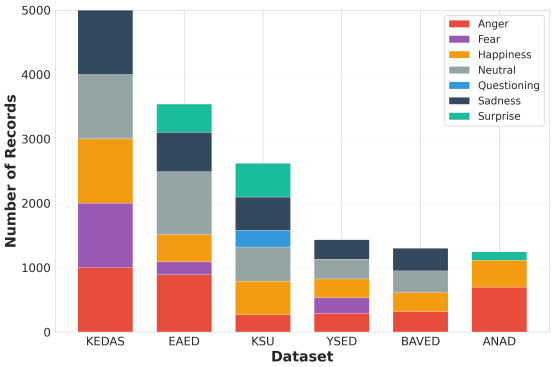


Figure 3: Emotion label distribution across training datasets.

Appendix

A Dataset Details

Figure 3 presents the distribution of emotion labels across multiple Arabic emotion datasets used for

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., 2017, 2019), 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 (ONeill et al., 2021), + in-house	12,000h
ASR	ASR	Augmentation	≈15,000h
Total			**≈27,000h**
		Phase 2: Multi-task Training	
Gen.	ASR	∼15% subset of Phase 1	≈1,740h
Gen.	SSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	374h
Gen.	TSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	99,995#
Disc.	DID	ADI-17 (Shon et al., 2020) + ADI-5 (MSA only) (Ali et al., 2017)	≈2,983h
Disc.	Emo.	ANAD (Klaylat et al., 2018) + EAED (Safwat et al., 2023) + KEDAS (Belhadj et al., 2022) + KSU (Meftah et al., 2021)	12.17h
Total			**5,109h + 99,995#**

Table 2: 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., 2017), 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)	w-F1
DID	ADI-17 (Shon et al., 2020)	w-F1

Table 3: 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.

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 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.

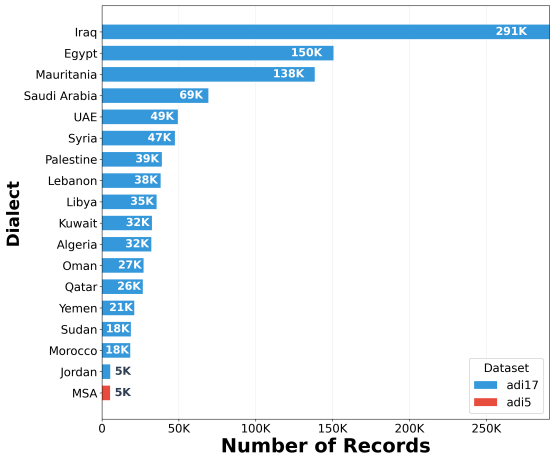


Figure 4: Dialect distribution in the training set.

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.

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-

lingual and cross-modal speech summarization training and evaluation.

B AraMega-SSum Validation

AraMega-SSum is the only newly developed semi-synthetic dataset introduced in this work. The other datasets used for ASR, emotion recognition, and dialect identification are curated from publicly available real-speech benchmarks. We therefore provide additional validation for AraMega-SSum to clarify its construction process, intended use, and quality-control procedures.

AraMega-SSum is designed to support Arabic speech summarization, a setting in which publicly available datasets with aligned speech, transcripts, and abstractive summaries remain scarce. The dataset serves as a quality-controlled semi-synthetic resource that enables controlled experimentation in this under-resourced setting.

It is derived from MegaSSum. We translate the source transcripts and abstractive summaries into Arabic using Gemini-2.5-Flash, followed by manual checks to improve semantic consistency, information preservation, contextual accuracy, completeness, and coherence. For speech generation, we use neural voice cloning conditioned on real speaker references. The generated speech is then evaluated through both human assessment and automatic quality checks.

B.1 Translation Quality

In Table 4, 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 5.

Metric (EN→AR transfer)	Score (/10)
Semantic Equivalence	9.96
Information Preservation	9.96
Contextual Accuracy	9.96
Completeness	9.96
Coherence	9.98

Table 4: LLM-as-a-judge (GPT-4.1) scores for English→Arabic summary translation by Gemini. Scores are out of 10.

B.2 Audio Quality

For speech generation, we use neural voice cloning conditioned on real speaker references. To validate the generated Arabic speech, we conduct both human and automatic quality checks. For the human evaluation, three paid annotators evaluate 100 randomly sampled audio clips from the test set. The annotators rate each clip on a 5-point Likert scale, where 5 means excellent, 4 good, 3 fair, 2 poor, and 1 bad. We evaluate three standard speech-quality dimensions: naturalness, audio quality, and intelligibility. As shown in Table 6, the generated speech receives good average scores across the three dimensions. We also report Gwet's AC1 (Gwet, 2008) to measure annotator agreement, which shows substantial agreement for naturalness and intelligibility and very high agreement for audio quality.

Metric (EN→AR transfer)	Score (/10)
Semantic Equivalence	9.93
Information Preservation	9.95
Contextual Accuracy	9.99
Completeness	9.95
Coherence	10.0

Table 5: Human scores for English→Arabic translation quality. Scores are out of 10.

Metric	Avg.	Gwet's AC1
Naturalness	3.975	78.9%
Audio Quality	4.373	95.0%
Intelligibility	3.692	78.9%

Table 6: Human speech-quality evaluation of 100 randomly sampled AraMega-SSum test-set audio clips.

In addition to the human evaluation, we report automatic speech-quality and speaker-consistency indicators. Specifically, we compute ASR word error rate using Fanar Aura ASR, overall speech quality using NISQA, and speaker cosine similarity between the generated audio and the corresponding speaker reference. The generated speech obtains a Fanar Aura ASR WER of 15.5, a NISQA overall score of 3.95 out of 5, and a speaker cosine similarity of 0.664. These values fall within acceptable ranges and are consistent with prior studies (Alam et al., 2025).

Overall, these results show that AraMega-SSum is a quality-controlled semi-synthetic resource for Arabic speech summarization. The dataset fills an important resource gap while keeping its semi-synthetic nature explicit. We therefore use it as a controlled benchmark for Arabic speech summarization experiments.

C Prompts

In Figures 6 and 7 and Listing 8, we present the prompts used for LLM-as-judge summarization

```
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.**

evaluation, LLM-as-judge translation evaluation, and multitask training.

D Training Loss

Figures 9 and 10 show the smoothed training-loss curves for the four multi-task fine-tunes, which all start from the same ASR-aligned LoRA checkpoint and differ only in how the multi-task data is presented. Uniform Multitask (UM) randomly shuffles all tasks together. Task Progressive Curriculum (TPC) reorders the same data so that each epoch progresses from easier to harder tasks, restarting the curriculum every epoch—visible as the intra-epoch U-shapes on Gemma TPC. Aligner Diverse Sampling (ADS), shown in blue, weights samples by aligner diversity, producing a higher early loss but the lowest final training loss on both backbones. TPC $\rightarrow$ ADS is our proposed hybrid: two epochs of curriculum to build a structured representation, followed by a third epoch of diversity-weighted sampling to broaden coverage. The step-6356 transition appears as a visible uptick where the distribution shifts, after which the loss re-stabilizes. The tail-zoom insets highlight that UM and TPC converge to comparable final loss (approximately 0.19 on Qwen and 0.23 on Gemma), confirming that the curriculum remains competitive with random shuffling at matched compute.

E Additional Results

In Table 7, we provide a detailed ASR breakdown across language varieties and speech conditions. The results show that adaptation consistently improves over the base Qwen2.5-Omni-7B model across MSA, dialectal Arabic, code-switched speech, native English, and L2 English. LA gives large gains over the base model, especially for English ASR, but the multi-task phases further improve Arabic and code-switched speech. For Qwen2.5-Omni-7B, TPC and TPC$\rightarrow$ADS achieve the strongest Arabic and code-switching results, while UM remains highly competitive. For Gemma-4-E4B-it, UM and TPC$\rightarrow$ADS perform best across most Arabic and code-switched settings, with TPC$\rightarrow$ADS giving the lowest WER on several dialectal and code-switched benchmarks. The closed models perform strongly on some subsets, but the adapted open models remain competitive, especially on MSA, English, and several dialectal or code-switched conditions. Overall, the detailed breakdown supports the aggregate ASR trends in the main results and shows that adaptation improves robustness across diverse speech conditions.

Table 8 reports detailed SSUM and TSUM results across English and Arabic using ROUGE-1, ROUGE-L, and BERTScore. Multi-task adaptation improves both Qwen2.5-Omni-7B and Gemma-4-E4B-it over their base and LA settings across most metrics. For Qwen2.5-Omni-7B, UM and TPC achieve the strongest SSUM results, while TPC and TPC$\rightarrow$ADS perform best on TSUM. For Gemma-4-E4B-it, UM gives the best SSUM scores, whereas TPC and TPC$\rightarrow$ADS lead on TSUM. These results support the main finding that UM and TPC provide strong summarization performance, while TPC$\rightarrow$ADS gives targeted gains in selected TSUM settings.

Table 9 reports LLM-as-a-judge scores for SSUM across English and Arabic. The judge scores are broadly consistent with the automatic metrics: UM and TPC remain competitive across both models, ADS scores lowest among adapted

```
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.**

phases, and Gemini-2.5-Pro leads overall.

Table 10 reports detailed dialect identification results on ADI17. Multi-task adaptation strongly improves both open models over their base and LA settings. For Qwen2.5-Omni-7B, TPC→ADS achieves the best accuracy, macro-F1, and weighted-F1, while TPC gives the second-best macro-F1. For Gemma-4-E4B-it, TPC→ADS achieves the best accuracy and weighted-F1, while UM gives the best macro-F1. These results show that adaptation is especially important for dialect identification. They also support the main finding that TPC→ADS is highly effective for discriminative speech understanding, although the strongest phase can still vary by model and metric.

```
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.**

Model	Phase	MSA	DA		CS		En L1		En L2
		MGB2	MGB3	SADA	ESCWA	DACS	Libri-clean	Libri-other	L2-Arctic
Qwen2.5-Omni-7B	Base	33.48	54.59	76.62	61.91	45.53	41.51	39.20	37.47
Qwen2.5-Omni-7B	LA	13.51	31.53	54.87	42.53	22.72	1.46	2.96	4.02
Qwen2.5-Omni-7B	UM	12.98	27.49	43.61	40.70	20.09	1.48	2.82	2.90
Qwen2.5-Omni-7B	TPC	12.98	27.46	43.66	40.31	19.94	1.45	2.86	3.10
Qwen2.5-Omni-7B	ADS	13.50	27.98	44.02	43.32	20.18	1.47	2.93	3.39
Qwen2.5-Omni-7B	TPC→ADS	13.21	27.28	43.79	40.64	20.06	1.48	2.90	2.95
Gemma-4-E4B-it	Base	16.70	30.97	42.60	43.81	22.70	3.86	9.34	13.30
Gemma-4-E4B-it	LA	13.77	26.87	44.32	34.37	19.79	2.57	6.59	12.64
Gemma-4-E4B-it	UM	12.42	24.05	40.06	41.95	17.57	2.27	5.85	7.31
Gemma-4-E4B-it	TPC	12.88	23.95	40.27	43.15	17.98	2.27	5.90	7.74
Gemma-4-E4B-it	ADS	13.36	24.48	43.39	40.90	18.52	2.58	6.22	8.53
Gemma-4-E4B-it	TPC→ADS	13.00	23.81	39.65	39.99	17.97	2.28	5.99	10.45
Qwen2.5-Omni-3B	Base	45.79	62.43	85.41	71.50	57.16	33.74	35.45	48.13
GPT-audio	Base	14.44	30.38	46.88	33.52	22.97	7.16	10.70	15.25
Gemini-2.5-Pro	Base	18.98	26.55	34.91	38.92	20.18	3.70	6.84	8.70

Table 7: Detailed ASR WER results across language varieties and speech conditions. We report WER for MSA, dialectal Arabic, Arabic–English code-switching, dialect–MSA code-switching, native English, and L2 English. Lower scores indicate better performance.

Model	Phase	SSUM						TSUM					
		English			Arabic			English			Arabic		
		R-1↑	R-L↑	B-F1↑	R-1↑	R-L↑	B-F1↑	R-1↑	R-L↑	B-F1↑	R-1↑	R-L↑	B-F1↑
Qwen2.5-Omni-7B	Base	29.37	26.48	57.37	23.70	21.53	53.75	25.67	23.30	51.13	29.84	27.56	57.86
Qwen2.5-Omni-7B	LA	25.75	23.05	56.61	23.84	21.69	53.59	24.35	21.87	52.87	25.34	23.11	54.22
Qwen2.5-Omni-7B	UM	46.23	**43.62**	**69.79**	36.74	34.71	63.14	48.14	45.41	70.78	38.20	36.06	**63.97**
Qwen2.5-Omni-7B	TPC	**46.36**	43.61	69.75	**36.75**	**34.74**	**63.21**	**48.26**	**45.55**	**70.85**	38.16	36.09	63.89
Qwen2.5-Omni-7B	ADS	41.43	38.79	67.32	33.66	31.65	61.23	43.56	40.86	68.41	35.77	33.66	62.52
Qwen2.5-Omni-7B	TPC→ADS	45.16	42.40	69.20	36.14	34.14	62.92	47.78	45.13	70.60	**38.28**	**36.16**	63.85
Gemma-4-E4B-it	Base	25.98	23.20	56.02	24.39	22.21	54.04	23.19	20.78	49.47	29.75	27.35	57.51
Gemma-4-E4B-it	LA	20.53	18.32	50.31	24.00	21.85	53.70	1.86	1.71	25.60	26.24	23.96	55.02
Gemma-4-E4B-it	UM	**45.32**	**42.57**	**69.05**	**37.49**	**35.46**	**63.53**	49.46	46.71	71.19	39.89	37.76	64.80
Gemma-4-E4B-it	TPC	45.03	42.31	68.87	36.76	34.81	63.37	**50.09**	**47.20**	**71.46**	39.76	37.69	64.90
Gemma-4-E4B-it	ADS	42.48	40.01	67.71	34.46	32.55	61.97	48.22	45.30	70.52	38.01	35.78	63.70
Gemma-4-E4B-it	TPC→ADS	43.98	41.24	68.34	36.34	34.27	63.07	49.78	46.81	71.30	**40.04**	**37.99**	**64.99**
Qwen2.5-Omni-3B	Base	13.13	11.79	38.90	22.05	19.95	52.49	14.36	12.95	38.33	27.68	25.31	56.34
GPT-audio	Base	5.48	5.21	29.62	28.63	26.24	57.42	–	–	–	–	–	–
GPT-5-chat	Base	–	–	–	–	–	–	34.08	30.87	60.83	31.07	28.56	59.41
Gemini-2.5-Pro	Base	28.48	25.75	57.01	27.70	25.40	56.15	33.20	30.35	59.74	34.13	31.59	60.55

Table 8: Detailed summarization results for SSUM and TSUM across English and Arabic. We report ROUGE-1, ROUGE-L, and BERTScore F1 (B-F1). Higher scores indicate better performance. Bold marks the best score within each adapted open-model block.

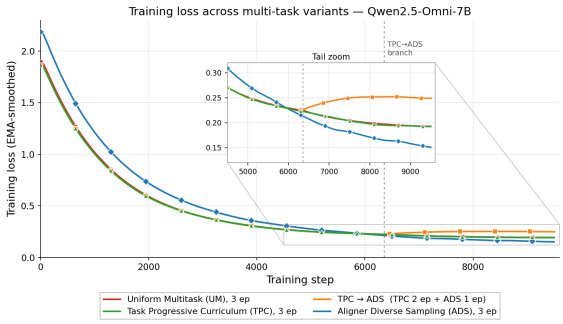


Figure 9: Training loss across multi-task variants for Qwen2.5-Omni-7B.

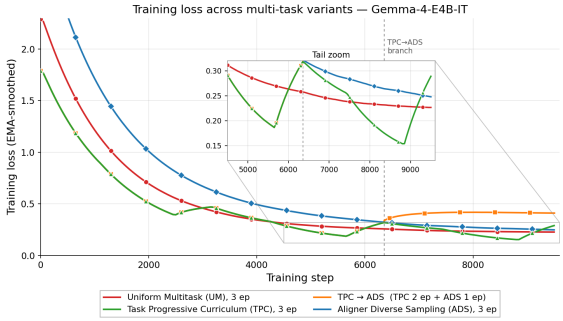


Figure 10: Training loss across multi-task variants for Gemma-4-E4B-IT.

Model	Phase	EN↑	AR↑
Qwen2.5-Omni-7B	Base	8.82	7.26
Qwen2.5-Omni-7B	LA	6.87	7.55
Qwen2.5-Omni-7B	UM	8.66	7.66
Qwen2.5-Omni-7B	TPC	8.63	7.65
Qwen2.5-Omni-7B	ADS	8.13	7.23
Qwen2.5-Omni-7B	TPC→ADS	8.58	7.61
Gemma-4-E4B-it	Base	8.13	7.95
Gemma-4-E4B-it	LA	6.45	7.79
Gemma-4-E4B-it	UM	**8.37**	7.71
Gemma-4-E4B-it	TPC	8.31	**7.74**
Gemma-4-E4B-it	ADS	8.11	7.34
Gemma-4-E4B-it	TPC→ADS	8.29	7.65
Qwen2.5-Omni-3B	Base	6.32	6.59
GPT-audio	Base	6.60	8.97
Gemini-2.5-Pro	Base	**9.03**	**9.01**

Table 9: LLM-as-a-judge scores for SSUM across English and Arabic (out of 10). Higher is better. Bold marks the best score within each adapted open-model block and among baselines respectively.

Model	Phase	Acc. ↑	M-F1 ↑	W-F1 ↑
Qwen2.5-Omni-7B	Base	13.77	7.56	10.31
Qwen2.5-Omni-7B	LA	5.08	0.01	8.58
Qwen2.5-Omni-7B	UM	87.91	82.80	87.76
Qwen2.5-Omni-7B	TPC	87.93	87.71	87.78
Qwen2.5-Omni-7B	ADS	86.77	77.49	86.72
Qwen2.5-Omni-7B	TPC→ADS	**88.85**	**88.80**	**88.80**
Gemma-4-E4B-it	Base	26.62	17.76	21.17
Gemma-4-E4B-it	LA	3.31	0.01	6.17
Gemma-4-E4B-it	UM	79.28	**78.89**	78.83
Gemma-4-E4B-it	TPC	79.03	78.57	78.55
Gemma-4-E4B-it	ADS	76.03	71.79	75.93
Gemma-4-E4B-it	TPC→ADS	**80.12**	75.68	**79.99**
Qwen2.5-Omni-3B	Base	11.14	7.14	7.66
GPT-audio	Base	40.96	25.04	38.44
Gemini-2.5-Pro	Base	65.09	59.85	63.37

Table 10: Detailed dialect identification results on ADI17. We report accuracy, macro-F1, and weighted-F1. Higher scores indicate better performance. Bold marks the best score within each adapted open-model block, and underlining marks the second best.