Harmonizing the Arabic Audio Space with Data Scheduling

Anonymous ACL submission

Abstract

Audio large language models (LLMs) enable unified speech understanding and generation, yet their adaptation to linguistically complex, dialect-rich settings remains underexplored. This paper presents the first systematic study of multi-task instruction tuning for an Arabic-centric audio LLM, covering a hierarchy of generative tasks (ASR, speech summarization) and discriminative tasks (dialect and emotion identification). To support this study, we introduce **AraMega-SSum**, a novel dataset for Arabic speech summarization. We fine-tune Qwen2.5-Omni (7B) and propose **Task-Progressive Curriculum (TPC)** along with **Aligner-Based Diverse Sampling (ADS)**, a strategy that constructs information-dense batches by selecting task- and label-balanced examples. Our results reveal a critical efficiency–robustness trade-off: while ADS accelerates initial convergence and boosts paralinguistic F1-scores, its inherent gradient volatility can destabilize generative decoding under prolonged training. Furthermore, while the TPC stabilizes core acoustic mapping, it often induces negative transfer in downstream tasks. We demonstrate that a **Hybrid TPC+ADS Strategy** provides an optimal training "recipe", first establishing a robust representative foundation before employing diversity-aware refinement to capture fine-grained nuances. These findings offer practical guidance for the efficient adaptation of Omni-models in complex, low-resource multimodal environments.

1 Introduction

Large Language Models (LLMs) are rapidly evolving into native multimodal systems capable of unified speech understanding and generation (Xu et al., 2025a). While these "Omni" models exhibit impressive zero-shot capabilities, their deployment in linguistically complex, low-resource settings – such as the Arabic-English bilingual landscape – remains a significant challenge. Mastering this domain requires instruction tuning that accounts for extreme dialectal variation, frequent code-switching, and paralinguistic nuances like emotion, which often confound general-purpose audio backbones.

Effective multi-task audio instruction tuning must harmonize *generative* objectives, such as Automatic Speech Recognition (ASR) and summarization (SSUM), with *discriminative* tasks like Dialect Identification (DID) and Emotion Recognition (SER). However, the optimal path to this unification is non-trivial. Naive data mixing often induces negative transfer or "gradient interference", where high-resource tasks like ASR overwhelm minority classes, leading to catastrophic forgetting or poor generalization in paralinguistic reasoning (Chen et al., 2024).

In this paper, we systematically investigate how **data scheduling and sampling strategies** influence the adaptation of Qwen2.5-Omni (7B) to a diverse Arabic-English audio task suite. To facilitate this study, we address the prevailing scarcity of Arabic multimodal resources by introducing a new benchmark: **AraMega-SSum**, a spoken short summarization dataset generated via translation and high-fidelity voice cloning.

Using a fixed compute budget[1] and Parameter-Efficient Fine-Tuning (PEFT) via LoRA, we evaluate three competing training regimes: (A) **Stochastic Mixing (SM)**, the standard baseline of uniform sampling; (B) **Task-Progressive Curriculum (TPC)**, which introduces tasks in increasing order of semantic abstraction; and (C) **Aligner-Based Diverse Sampling (ADS)**, a novel approach that utilizes K-means clustering in the aligner's embedding space to construct "information-dense" effective batches. (D) **Hybrid Strategy with**

[1]We define full-budget training as an unconstrained optimization process continued until convergence, rather where as fixed compute adhere to a fixed number of steps irrespective of its convergence or not.

TPC+ADS, to mitigate challenges of pure TPS and ADS.

Our findings reveal a critical trade-off: while TPC strengthens core acoustic mapping (ASR), it often compromises paralinguistic robustness as higher-level tasks are introduced. Conversely, our proposed ADS method significantly accelerates initial convergence but risks earlier saturation compared to stochastic mixing. We hypothesize that aggressive within-batch diversification, while efficient, may exacerbate objective conflict in shared representational spaces (Guo et al., 2023).

Our main contributions are as follows: (a) We develop a specialized Arabic-English Audio LLM by fine-tuning Qwen2.5-Omni across a hierarchy of five generative and discriminative tasks. (b) We introduce **AraMega-SSum**, providing a new benchmark for high-level semantic compression in Arabic speech. (c) We propose **Aligner-Based Diverse Sampling (ADS)** and demonstrate that a **Hybrid TPC+ADS** approach optimizes the trade-off between convergence speed and long-term task robustness. (d) We provide a comprehensive analysis of multi-task scheduling, offering a technical "recipe" for adapting Omni-models to complex, dialectal variant audio environments. To the best of our knowledge, this is the first work to investigate specialized training regimes for an Arabic-centric Omni-model across the full spectrum of acoustic, linguistic, and paralinguistic tasks.

2 Related Work

Audio and multimodal instruction-tuned LLMs. LLMs have been extended from text to speech and other modalities by aligning pretrained encoders with language decoders via pretraining and instruction tuning. Early unified speech models such as SpeechT5 demonstrated multi-task speech learning in a shared encoder–decoder framework (Ao et al., 2022). Later, AudioPaLM and SpeechGPT integrated strong LLM backbones with speech tokenization or multi-stage instruction training to enable speech understanding and generation (Rubenstein et al., 2023; Zhang et al., 2023). Qwen-Audio scaled this paradigm to many audio tasks under a unified instruction interface (Chu et al., 2023), and recent "omni" models aim to unify perception and generation across modalities at scale (Xu et al., 2025b,c; Rouditchenko et al., 2025). However, these works largely under-analyze how *task mixtures* and *scheduling* shape stability and cross-task generalization in end-to-end audio instruction tuning.

Multi-task interference and scheduling. Multi-objective audio instruction tuning often suffers from negative transfer, where dominant tasks suppress minority or higher-level skills. Octavius addresses this via LoRA-based mixture-of-experts to reduce "tug-of-war" interference (Chen et al., 2024), while MINT shows that indiscriminate task aggregation can hurt generalization and that structured grouping mitigates interference (Shan et al., 2025). Curriculum learning can further improve optimization when task order is compatible (Bengio et al., 2009), motivating systematic evaluation of curricula versus diversity-driven sampling.

Arabic speech understanding beyond ASR. Arabic benefits from multilingual speech foundation models such as Whisper and XLS-R and from LLM-based Arabic ASR/DID systems like Octopus (Radford et al., 2022; Babu et al., 2022; Althubaiti et al., 2025). In contrast, higher-level Arabic speech understanding remains limited: summarization work is mostly text-based (e.g., AraBART) (Eddine et al., 2022), and SER typically relies on small, task-specific corpora such as BAVED and KSUEmotions (Aouf, 2020; Meftah et al., 2018), with little integration into unified instruction-following audio models.

Positioning of this work. We study an Arabic–English audio-capable LLM that unifies *generative* (ASR, summarization) and *discriminative* (dialect, emotion) tasks, and we empirically analyze *data scheduling and sampling* – stochastic mixing, task-progressive curricula, and aligner-space diversity sampling to characterize trade-offs in convergence, interference, and robustness for Arabic-centric omni-model adaptation.

3 Methodology

We propose a multi-stage framework to adapt a multimodal LLM for complex Arabic-English audio tasks. Our approach transitions from language-centric alignment to a comparative study of task-scheduling regimes. We utilized the **Qwen2.5-7B-Omni** architecture (Xu et al., 2025a), which integrates a **Whisper-v3 (Radford et al., 2023)** based audio encoder with a 7B-parameter transformer backbone via a linear projection aligner.

3.1 Task Formalization

We define a unified instruction-tuning space, $\mathcal{T} = \{\mathcal{T}_g, \mathcal{T}_c\}$, with both generative ($\mathcal{G}$) and discriminative classification ($\mathcal{C}$) tasks. For any audio input x and task prompt p, the model is then jointly optimized via cross-entropy loss.

Generative tasks. $\mathcal{G}$ tasks includes: (i) *ASR*: generate a verbatim transcription of the input utterance; (ii) *Text Summarization (TSUM)*: given transcription, generate a short written summary in Arabic or English; (iii) *Speech Summarization (SSUM)*: given speech, generate a short summary in Arabic or English.

Classification tasks. $\mathcal{C}$ tasks includes: (iv) *Dialect Identification (DID)*: predict the Arabic dialect label from speech; and (v) *Emotion Recognition (SER)*: predict one of seven emotion categories from speech.

3.2 Phase 1: Language-centric Alignment

To bridge the modality gap, we first perform full-parameter fine-tuning of the audio encoder and the linear aligner using a large-scale ASR objective with LoRA for LLM. This "anchors" the latent space to phonetic and lexical representations of both Arabic and English before introducing higher-level reasoning tasks.

3.3 Phase 2: Multi-task Training Regimes

To investigate the optimal scheduling of heterogeneous tasks, we then compare four regimes under a fixed compute budget using *LoRA* for the LLM backbone. In this phase the audio encoder is frozen.

Regime A: Stochastic Mixing for Multi-tasking (SM)

We employ uniform random sampling across the union of all datasets $\mathcal{D} = \bigcup_{t \in \mathcal{T}} \mathcal{D}_t$. This represents the standard approach to instruction tuning but is prone to gradient dominance by high-resource tasks.

Regime B: Task-Progressive Curriculum (TPC)

We implement a curriculum based on semantic abstraction. The model is exposed to tasks in five sequential stages $S_1 \ldots S_5$:

> **Acoustic**: ASR $\rightarrow$ **Paralinguistic**: DID, SER $\rightarrow$ **Reasoning**: TSUM, SSUM.

where we also include a portion of previous stage tasks in the current one.

Regime C: Aligner-Based Diverse Sampling (ADS)

To mitigate the challenges of task interference and label imbalance – particularly the dominance of Modern Standard Arabic (MSA) over regional dialects, we propose Aligner-Based Diverse Sampling (ADS). This method optimizes the "information density" of each training batch by sampling from a discretized latent space created by K-means clustering of aligner's output representation.

Latent Codebook Construction 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.

Regime C: TPC followed by ADS

Following, we employs a staged optimization strategy. We first conduct initial steps on using TPC strategy to stabilize multi-task learning through structured task complexity. This is immediately followed by ADS for the remaining steps, utilizing dynamic weighting to maximize linguistic variety and mitigate task interference.

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

4 Datasets

To facilitate our language-centric alignment and multi-task objectives, we compile a diverse training corpus encompassing Modern Standard Arabic (MSA), Dialectal Arabic (DA), and English. However, the development of Arabic-centric audio LLMs is severely hindered by a significant scarcity of high-level semantic and paralinguistic benchmarks.

Specifically, **Arabic Spoken Summarization** data is virtually non-existent in the public domain, as existing speech corpora are largely restricted to transcription (ASR) task.To address these gaps, we introduce and publicly release **AraMega-SSum**: A novel, large-scale Arabic Speech Summarization dataset consisting of 50,618 training (222 hours) and 4,000 test samples (16.98 hours), serves as the first end-to-end benchmark for semantic compression in Arabic audio This contributions provide the necessary foundation for evaluating high-level reasoning in Arabic-centric multimodal systems.

4.1 AraMega-SSum: Arabic Short Speech Summarization

A primary contribution of this work is the introduction of **AraMega-SSum**, the first large-scale, high-quality dataset for end-to-end Arabic Spoken Short Summarization. Given the extreme scarcity of paired audio-to-summary data in Arabic, we developed a sophisticated semi-synthetic pipeline designed to preserve semantic integrity while maximizing acoustic diversity.

Translation and Semantic Alignment 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 speechtext 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 to evaluate the quality of human summary translations with respect to semantic equivalence, information preservation, contextual accuracy, completeness, and coherence. The scores on the test set are presented in Table 3.

Audio Synthesis via Neural Voice Cloning To reconstruct the speech modality, we utilize **XTTS-v2** (Casanova et al., 2024), a state-of-the-art multi-speaker latent diffusion model for speech synthesis. To mitigate the risk of the model overfitting to static synthetic acoustic profiles, we leverage the zero-shot **voice-cloning** capabilities of the model. We utilize a curated bank of reference speaker embeddings derived from **47 distinct speakers** representing **10 Arab countries**. This geographic breadth ensures that the generated speech encapsulates a wide spectrum of regional prosody, pitch variations, and glottal characteristics inherent to the Arab world.

For each reference speaker, we maintain at least 10 high-quality voice clips; during the generation of the *AraMega-SSum* corpus, a reference clip is randomly sampled to ensure high intra-speaker acoustic variability. This approach provides the requisite "acoustic diversity" for the LLM to generalize across heterogeneous recording setups. For quality control, we exploit a pre-trained Arabic-centric state-of-the-art *Fanar-STT-v1*[3] model (Team et al., 2025) to select the generated audio for each utterance. We only took one audio candidates for (Transcription, Summary) pair where the ASR output exhibited a Word Error Rate (WER) $< 15\%$.

[3] https://api.fanar.qa/docs

Phase 1: Language-centric Alignment			
Task Obj.	**Task**	**Dataset(s)**	**Total (h / #)**
ASR	ASR	QASR (Mubarak et al., 2021), MASC (Al-Fetyani et al., 2023), MGB-3 (Ali et al., 2017a), MGB-5 (Ali et al., 2017b), GigaSpeech (Chen et al., 2021), GALE,[2] TED-LIUM 3 (Hernandez et al., 2018), LibriSpeech (Panayotov et al., 2015), Common Voice (ar/en) (Ardila et al., 2020), SPGISpeech (O'Neill et al., 2021), + in-house data	10,000h
Phase 2: Multi-task Training			
Task Obj.	**Task**	**Dataset(s)**	**Total (h / #)**
Generative	ASR	20% subset of Phase 1	≈1,740h
	TSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	101,242#
	SSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	378h
Discriminative	Dialect ID	ADI-17 (Shon et al., 2020) + ADI-5 (only MSA subset) (Ali et al., 2017a)	2,983.36h
	Emotion	ANAD (Klaylat et al., 2018) + EAED (Safwat et al., 2023) + KEDAS (Belhadj et al., 2022) + KSU (Alotaibi, 2020) + BAVED (Aouf, 2020) + YSED (Derhem et al., 2025)	12.17h

Table 1: Datasets used in each training phase. "h" denotes audio duration; "#" denotes the number of instruction instances/samples.

Task	**Benchmarks**	**Metrics**
ASR (Native)	LibriSpeech (Panayotov et al., 2015), MGB-2 (Ali et al., 2019)	WER
ASR (Dialect)	MGB-3 (Ali et al., 2017a), 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
Summarization	MegaSUM-SSum (Matsuura et al., 2024), AraMega-SSum (**ours**)	RL, BScore, Lj
Emotion	KSU (Alotaibi, 2020), MELD (Poria et al., 2019)	w-F1
Dialect ID	ADI-17 (Shon et al., 2020)	w-F1

Table 2: Evaluation benchmarks and metrics used in this work. RL: ROUGE-L, BScore: BERTScore, Lj: LLM-judge, w-F1: Weighted-F1

Metric (EN→AR summary transfer)	**Score (/10)**
Semantic Equivalence	9.96
Information Preservation	9.96
Contextual Accuracy	9.96
Completeness	9.96
Coherence	9.99
Average	**9.97**

Table 3: Translation quality scores for English→Arabic summary transfer produced by Gemini, evaluated by an LLM-judge (GPT-5). Scores are out of 10.

The resulting dataset, **AraMega-SSum**, comprises ≈**50,000** validated pairs, providing a robust foundation for training the model's semantic compression capabilities.

4.2 Training Datasets

Phase 1: Language-centric Modal Alignment. To anchor the multimodal latent space, we first conduct a language-centric alignment phase focused exclusively on ASR. While Arabic-English speech data is abundant, we curate a high-quality subset of approximately **10.1K hours** (comprising 5,448,405 samples) to ensure a balanced linguistic distribution and maintain computational efficiency.

To enhance the model's acoustic robustness against real-world variability, we implement a targeted data augmentation strategy. We select a 100-hour representative subset and apply **speed perturbation** (factors of 0.9x and 1.1x) alongside **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. During this phase, we perform full-parameter updates on the encoder and linear aligner to specialize the audio perception stack for the bilingual phonology of Arabic and English before proceeding to multi-task instruction tuning. See Table 1 for details.

Phase 2: Multi-task Training.

Our final multi-task corpus aggregates approximately **5113** hours of speech, strategically partitioned into generative and discriminative task suites. To ensure training stability and optimize memory utilization, we constrained the duration of all speech instances to $t \leq 180$ seconds.

Task Balancing and Paralinguistic Data: Given the high-resource nature of ASR compared to downstream semantic tasks, we performed a controlled **subsampling of the ASR dataset**. This downsampling minimizes the distributional shift and prevents the generative ASR objective from dominating the gradient signal during multi-task

SFT. For **SER**, we curated publicly available Arabic affective corpora, ensuring a broader coverage of spontaneous emotional prosody. For **DID**, we unified the ADI-17 (Shon et al., 2020) corpus with the MSA partitions from the ADI-5 dataset, providing a robust taxonomy for six regional variants. For per-task label-wise distribution see Figure 3-**??** in Appendix.

Multilingual Speech Summarization: For English speech summarization, we utilize the standard training partitions of the Mega-SSum dataset. For Arabic, we used the designed aforementioned **AraMega-SSum** datasets. Examples of the dataset are presented in Appendix Figure 5.

4.3 Evaluation Benchmarks

To assess the multi-task capabilities of our adapted Qwen2.5-Omni model, we curate a comprehensive evaluation suite spanning generative and discriminative tasks from publicly available test sets.

Automatic Speech Recognition (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., 2019) for Modern Standard Arabic (MSA). We further evaluate the models robustness to dialectal variation and code-switching through a suite of challenging datasets: (a) **Dialectal Diversity:** We use the **MGB-3** Egyptian test set (Ali et al., 2017a) and the **SADA** (Alharbi et al., 2024) Saudi dialectal corpus to measure regional linguistic adaptation; (b)**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).

This multi-faceted evaluation strategy provides a comprehensive assessment of the model's performance across MSA, diverse DA, and standard English in both spontaneous and code-switched contexts.

Emotion Recognition (ER) We evaluate affective reasoning across two distinct benchmarks to measure cross-domain generalization: (a) **KSU Emotions:** A benchmark of acted Arabic speech (Alotaibi, 2020), used to test the model's sensitivity to overt vs. spontaneous emotional cues; (b) **MELD (Cross-Lingual):** The Multimodal Emotion Lines Dataset (Poria et al., 2019) is used to evaluate the model's zero-shot emotion recognition capabilities on English conversational audio.

Dialect Identification (DID) We employ the **ADI17** test set (Shon et al., 2020), which requires the classification of 17 distinct 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.

Spoken and Textual Summarization (SSum and TSum) To evaluate semantic compression and abstractive reasoning, we utilize the test partitions of **Mega-SSum** for English (Matsuura et al., 2024) and our proposed **AraMega-SSum** for Arabic. These benchmarks assess the model's ability to synthesize concise, high-fidelity summaries from short-form audio. To isolate the impact of acoustic modeling from textual reasoning, we introduce a **cross-modal diagnostic**: we evaluate a *Text-Only Version* of these test sets by bypassing the audio encoder and feeding gold transcripts directly to the LLM backbone. This establish an Oracle upper-bound for the summarization logic.

5 Experimental Setup

5.1 Model Configuration

We initialize our experiments with the **Qwen2.5-7B-Omni** backbone (Xu et al., 2025a). The model architecture consists of a Whisper-v3 based audio encoder and a linear aligner that maps acoustic features into the 4096-dimensional LLM embedding space.

Training Procedure and Parameters Training is conducted in two distinct phases: **Phase 1 (Alignment):** 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 SFT):** We execute the four comparative training regimes (SM, TPC, ADS and TPC+ADS) for approximately 10K steps. Here, the encoder and aligner are frozen, and only the LoRA adapters in the LLM are updated. For all experiments, we used the learning rate 3×10^{-5}, with a linear warmup for the first 30% of epoch 1,

followed by a cosine annealing schedule. The training is distributed across 24 (H100) GPUs with a per-device batch size of 2. By utilizing 16 gradient accumulation steps, we achieve an effective global batch size of 768.

ADS Clustering Details For the ADS regime, the global codebook is $K = 500$ centroids based on a 3% subset of the data ($\approx$ 75 hours). The high value of K is chosen to ensure that fine-grained paralinguistic variationssuch as emotional prosody and regional dialectal phonemesare captured as distinct clusters.

Evaluation Metrics We employ a multi-faceted evaluation suite: For ASR, we used Word Error Rate (WER); for Arabic ASR specifically, we normalize both hypothesis and reference strings by unifying Alef/Hamza variants and removing all diacritics to ensure consistency. For summarization, we reported ROUGE-L and BERTScore (Zhang et al., 2019) to measure semantic overlap, and supplemented by an LLM-as-a-Judge framework using GPT-4.1. This judge evaluates the model on a scale of 1–10 based on Clarity, Conciseness, Coherence, Completeness, Semantic alignment, Accuracy, Relevance, and Information density (see prompts used in Figure 6). For classification tasks, we used weighted-F1 to account for inherent label imbalance. Finally, for discriminative outputs, we utilize a post-processing script to map variants to canonical labels (e.g., mapping "KSA" to "Saudi Arabia").

6 Results and Discussion

Base Model vs. Language-Centric Alignment

Our results show that *language-centric alignment* (LA) successfully shifts the base model's acoustic space toward Arabic without compromising much of the English performance (see Table 4). Notably, LA benefits *unseen* tasks like SSUM, SER, and DID (see Tables 5–6). Improved TSUM scores – evaluated by our GPT-4.1 judge – further indicate that LA provides higher-fidelity representations that enhance the LLM decoder's reasoning. Note that for the study, we have not compared with SOTA results present for any specialized model (for e.g., (Team et al., 2025)). We only put Multimodal/Audio LLM-based models available such as Gemini and for DID we compare with (Althubaiti et al., 2025). The objective of the paper is not to push SOTA but compare data scheduling strategy with fixed computation budget.

Multi-Task Training Regimes

Stochastic Multi-tasking (SM) SM remains the most stable for resource-rich generative tasks (ASR, SSUM). While convergence is slower (Appendix Figure 1), sampling from the natural distribution provides a consistent gradient that reaches a balanced global minimum. However, SM fails to capture niche paralinguistic nuances, yielding lower F1-scores (in Table 6).

Task-Progressive Curriculum (TPC) TPC stabilizes early ASR learning but induces significant paralinguistic interference. By postponing non-transcription tasks, the model develops a "phonetic-lexical anchor" that makes later adaptation to affective or dialectal signals difficult which is a clear case of negative transfer (Table 6).

Aligner-Based Diverse Sampling (ADS) ADS prioritizes information-dense outliers, accelerating early gains. However, this creates high gradient noise and task interference.[4] Under prolonged training, ADS overfits to the "tails" of the distribution, memorizing outliers rather than generalizable features. This is particularly damaging for generative tasks (ASR, SSUM) where the decoder requires stable, redundant patterns to master linguistic structures.

Hybrid Strategy (TPC+ADS) TPC+ADS mitigates the weaknesses of both TPC and ADS. Initial TPC stages establish a stable representative foundation, while the subsequent ADS phase acts as a "diversity-refinement" period. This strategy boosts F1-scores without the catastrophic saturation seen in pure ADS, proving that diversity is most effective when refining a pre-stabilized mapping.

Key Takeaways

Data scheduling matters: How data is ordered impacts dynamics and final performance significantly under fixed computation budgets.
When Diversity Hurts vs. Helps: ADS helps discriminative tasks find class boundaries but hurts generative tasks. 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"

[4] Also can be seen in the switching point between TPC and ADS (Figure 1 (Appendix))

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
SM	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 4: 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.

Model	Arabic			English		
	R-L	**BERT**	**Judge**	**R-L**	**BERT**	**Judge**
SSUM (Speech Summary)						
Base	8.51	40.89	6.12	28.94	57.63	8.48
LA	22.88	54.25	7.43	23.23	55.86	7.89
SM	**35.75**	**63.47**	**7.79**	42.12	66.74	**8.54**
TPC	35.46	63.45	7.75	**44.49**	**68.83**	8.51
ADS	34.35	62.73	7.48	42.92	68.01	8.40
TPC + ADS	35.16	63.24	7.61	42.31	67.80	8.41
Gemini	24.17	55.22	8.00	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
SM	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 5: 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
SM	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 6: 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).

knowledge. The low-resource tasks then become highly sensitive to the gradients of the dominant tasks (like ASR) causing the signal washout. Furthermore, since we rely on LoRA, high-variance diverse gradients risk causing representation collapse in the decoder.

Hybrid Approach as a Solution: TPC builds the "bridge" (rules) and stabilize the acoustic to semantic learning process. Whereas, ADS provides the "nuance" (diverse set). Introducing ADS after TPC reverses curriculum-induced negative transfer, forcing the model to look beyond words to paralinguistic features.

7 Conclusion

We studied how data scheduling affects multi-task instruction tuning of an Arabic-centric audio LLM. We benchmarked Arabic AudioLLM, with Qwen2.5-Omni as backbone, on both generative tasks (ASR, SSUM) and discriminative tasks (SER, DID), and introduced AraMega-SSum to support Arabic speech summarization. Our investigation of the proposed Task-Progressive Curriculum (TPC) and Aligner-Based Diverse Sampling (ADS) uncovered a critical efficiency–robustness trade-off: while ADS accelerates convergence for discriminative tasks, its inherent gradient volatility can destabilize generative decoding. We demonstrated that a Hybrid staged strategy (TPC+ADS) successfully mitigates these effects by stabilizing acoustic-semantic mappings before introducing diversity-aware refinement. Our future work will investigate these effects with interpretability perspective and test architectural changes that better balance generative and discriminative speech understanding.

Limitations

Our study focuses on multi-task audio instruction tuning of a single audioLLM (Qwen2.5-Omni, 7B) in an Arabic-English setting. While this choice reflects a realistic and challenging deployment scenario, our findings may not directly generalize to models with substantially different architectures, capacity, or training objectives. In particular, mod-

els with explicit task-specific heads or decoupled audio-text pathways may exhibit different sensitivity to data scheduling strategies.

We evaluate data scheduling strategies under a fixed training budget and parameter-efficient fine-tuning setup. Although this constraint is intentional and reflects practical compute limitations, different conclusions may emerge under significantly larger training budgets or full-parameter fine-tuning.

Acknowledgments

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. 2019. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. *Preprint*, arXiv:1609.05625.

Ahmed Ali, Najim Dehak, , and 1 others. 2017a. The MGB-3 challenge: Arabic multi-dialect broadcast media recognition. In *SLT*.

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

Yousef Alotaibi. 2020. The ksu arabic speech emotion corpus. *Journal of King Saud University - Computer and Information Sciences*.

Sara Althubaiti, Vasista Sai Lodagala, Tjad Clark, Youssef 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, and 1 others. 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.

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.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*, pages 2278–2282.

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. *Preprint*, arXiv:2406.04904.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Zeyu Chen, Yuhui Zhang, Xiaolong Li, and Hang Zhao. 2024. Octavius: Mitigating task interference in multimodal instruction tuning with lora mixture-of-experts. *arXiv preprint arXiv:2404.0XXXX*.

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. *arXiv 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.

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 7th Arabic Natural Language Processing Workshop (WANLP)*. Association for Computational Linguistics.

Han Guo, Rui Wang, Yiming Li, and Zhou Yu. 2023. Curriculum learning for multimodal reasoning. In *Proceedings of ICLR*.

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.

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.

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. *Preprint*, arXiv:2408.00205.

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.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource – a large scale annotated arabic speech corpus. *Preprint*, arXiv:2106.13000.

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. *Preprint*, arXiv:1810.02508.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. 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, 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 12 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

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

Yifan Shan, Ziming Liu, Zeyu Chen, and Hang Zhao. 2025. Mint: Multimodal interaction grouping for reducing task interference in multimodal instruction tuning. *arXiv preprint arXiv:2501.0XXXX*.

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, 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. 2025b. 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. 2025c. 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. *arXiv preprint arXiv:2305.11000*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Guoguo Zhao, Süleyman Sonsaat, Anuphap Silpachai, Igor Lucic, Ekaterina Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech 2018*, pages 2783–2787.

8 Appendix

This section includes the training loss comparision on fixed bugdet computation with max steps 10K in Fingure 1 along with data distributions and prompts used for the study.

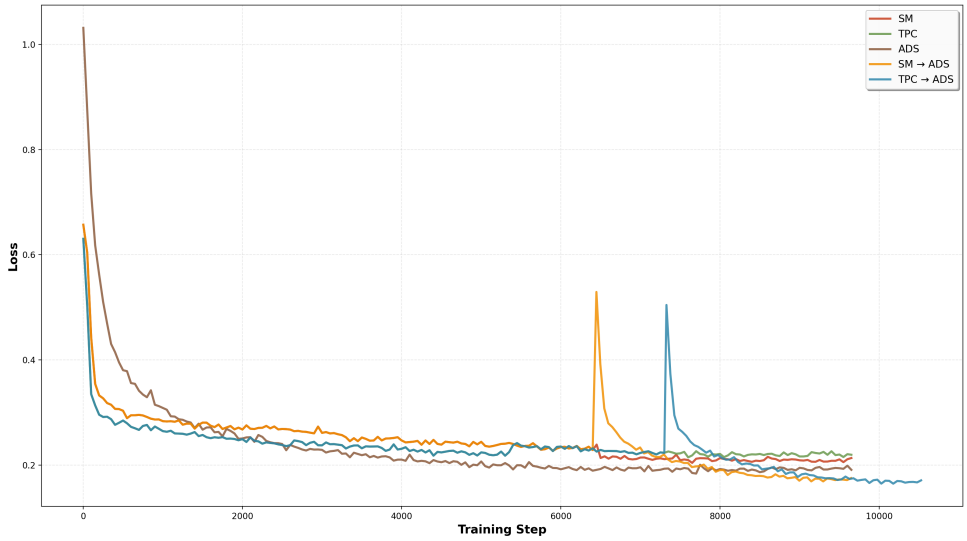

Figure 1: **Training loss comparison across multi-task training regimes.** This figure compares training loss trajectories for different multi-task training regimes over three epochs. Stochastic Multi-tasking (SM, red) and Task-Progressive Curriculum (TPC, green) exhibit smooth and stable convergence. 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.

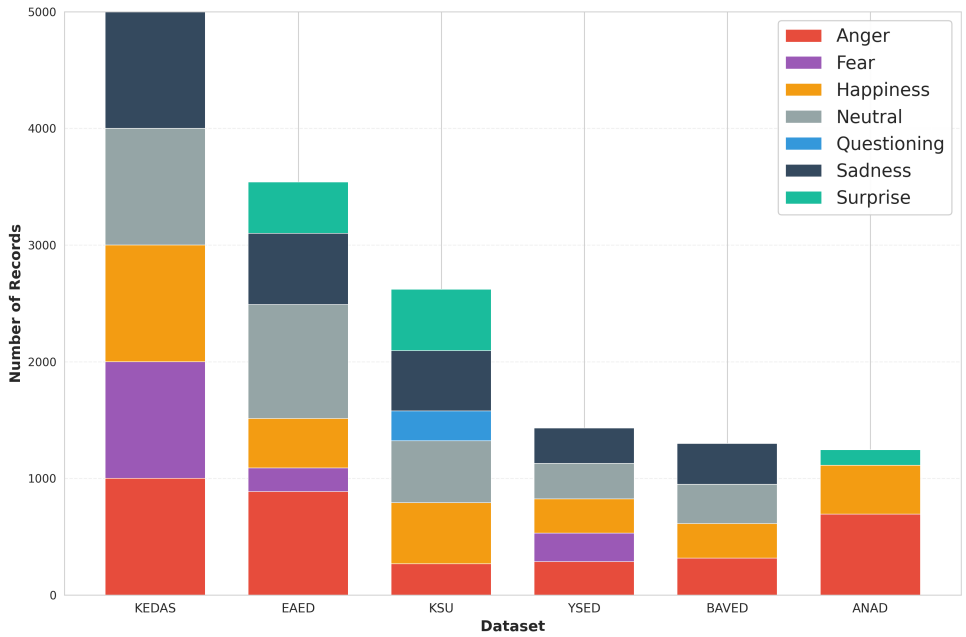

Figure 2: **Emotion label distribution across training datasets.** This figure presents the distribution of emotion labels across multiple Arabic emotion datasets used for training, including KEDAS, EAED, KSU, YSED, BAVED, and ANAD. Each dataset exhibits a distinct class composition, with emotions such 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.

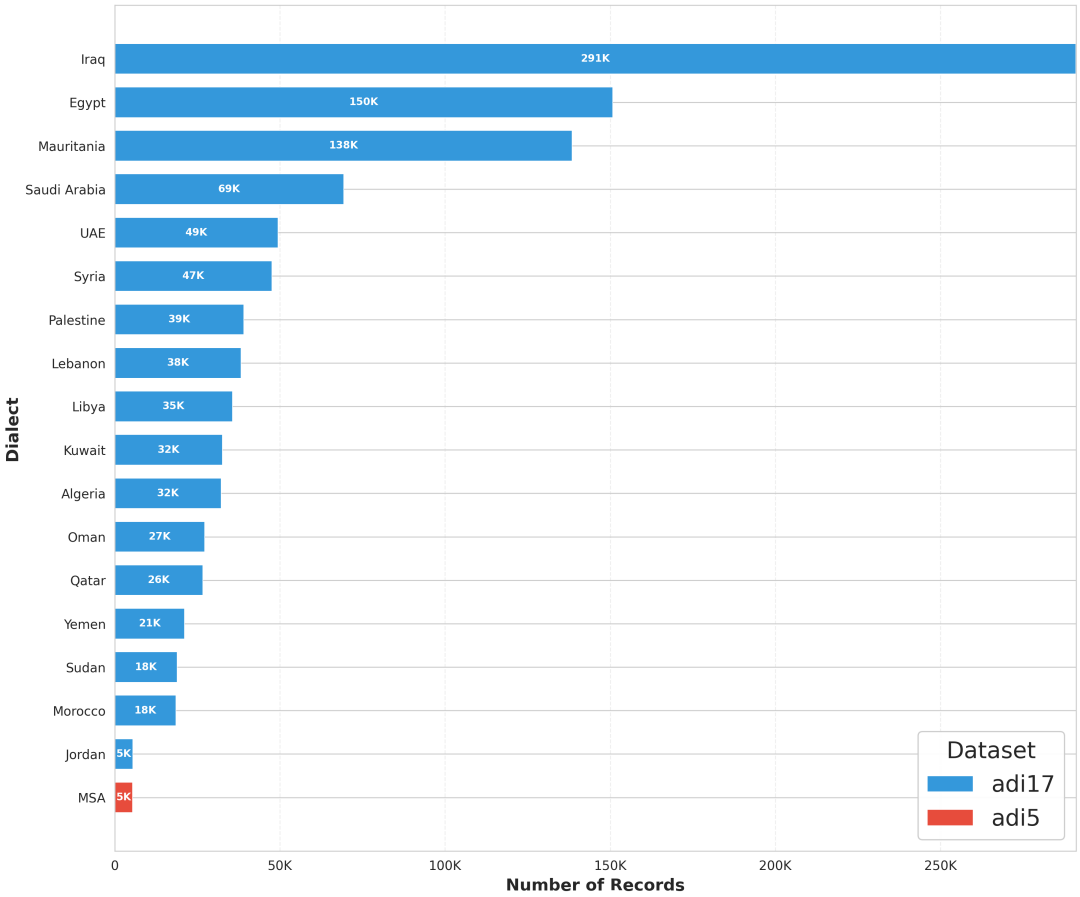

Figure 3: **Dialect distribution in the training set.** The training data covers dialects across 17 countries using the **ADI–17** dataset, along with MSA from **ADI–5**. The distribution is dominated by Egyptian, Iraqi, and Mauritanian dialects, while the remaining dialects are represented with varying but comparatively smaller sample sizes.

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 4: **System and user prompts used for multi-task training.** This figure shows the prompt templates employed during multi-task training. The system prompt is kept fixed across all tasks to enforce consistent model behavior, while task-specific user prompts are varied to cover different formulations of each objective. Arabic prompt variations are included for each task to increase linguistic diversity and improve robustness across dialects and use cases.

English Transcription	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	yugoslav court issues subpoenas to nato leaders
Arabic Transcription	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	محكمة يوغوسلافية تصدر أوامر استدعاء بحق قادة ناتو.

English Transcription	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	a man suicides in vatican 's basilica of st. peter
Arabic Transcription	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	رجل ينتحر في كاتدرائية القديس بطرس بالفاتيكان

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

Listing 1: Task and label-level distribution per super-batch for **Aligner-Based Diverse Sampling**. The batch exhibits a balanced and diverse composition across ASR, dialect identification, summarization, and emotion recognition tasks, ensuring effective multi-task training.


```
{
  "task_distribution": {
    "asr": 274,
    "did": 258,
    "speech_summarization": 80,
    "text_summarization": 80,
    "emotion": 31
  },
  "label_distribution": {
    "did": {
      "Palestine": 13,
      "United Arab Emirates": 12,
      "Egypt": 6,
      "Iraq": 5,
      "Libya": 14,
      "Jordan": 25,
      "Sudan": 19,
      "Syria": 12,
      "Algeria": 14,
      "Yemen": 18,
      "Lebanon": 13,
      "Morocco": 19,
      "Oman": 16,
      "Qatar": 16,
      "Mauritania": 7,
      "Kuwait": 14,
      "Saudi Arabia": 10,
      "Modern Standard Arabic": 25
    },
    "emotion": {
      "Surprise": 5,
      "Anger": 3,
      "Sadness": 3,
      "Neutral": 3,
      "Happiness": 3,
      "Fear": 4,
      "Questioning": 10
    }
  }
}
```

```
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.** The prompt enforces reference-grounded comparison between predicted and gold summaries in Arabic or English, scoring multiple quality dimensions and producing a structured JSON output for automatic evaluation.