

Closing the Modality Reasoning Gap for Speech Large Language Models

Chaoren Wang^{1*}, Heng Lu², Xueyao Zhang¹, Shujie Liu^{2†},
Yan Lu², Jinyu Li², Zhizheng Wu^{1†}

¹The Chinese University of Hong Kong, Shenzhen,

²Microsoft Corporation

Abstract

Although Speech Large Language Models have achieved notable progress, a substantial modality reasoning gap remains: their reasoning performance on speech inputs is markedly weaker than on text. This gap could be associated with representational drift across Transformer layers and behavior deviations in long-chain reasoning. To address this issue, we introduce **TARS**, a reinforcement-learning framework that aligns text-conditioned and speech-conditioned trajectories through an asymmetric reward design. The framework employs two dense and complementary signals: representation alignment, which measures layer-wise hidden-state similarity between speech- and text-conditioned trajectories, and behavior alignment, which evaluates semantic consistency between generated outputs and reference text completions. Experiments on challenging reasoning benchmarks, including MMSU and OBQA, show that our approach significantly narrows the modality reasoning gap and achieves state-of-the-art performance among 7B-scale Speech LLMs.

1 Introduction

Recent advances in Speech Large Language Models (Speech LLMs) enable a unified framework for spoken language processing tasks such as automatic speech recognition (ASR), speech translation, and speech QA. Most Speech LLMs follow a three-stage architecture, consisting of a pre-trained speech encoder, lightweight adapters, and a decoder-only text LLM (Peng et al., 2025a; Cui et al., 2025). The encoder transforms raw speech into high-resolution acoustic representations, which are then projected into the text-embedding space through learned adapters, enabling the downstream LLM to process speech inputs using its text-native reasoning stack. Through shared representations across modalities, this architecture allows speech inputs to leverage the generation and reasoning capabilities of text-based LLMs.

However, Speech LLMs exhibit a persistent and critical challenge: the modality reasoning gap, denoting a substantial decline in reasoning performance on speech inputs compared to text, as evidenced by empirical analyses (Xiang et al., 2025) on VoiceBench and SpeechMMLU benchmarks (Chen et al., 2024; Xiaomi, 2025).

To close modality reasoning gap, prior works have primarily focused on input-side fusion and output-side supervision. Input-level modality fusion methods aim to reduce the discrepancy between speech representations and text embeddings at input stage, by freezing the LLM backbone and training lightweight adapters (Chen et al., 2026; Lu et al., 2025a; Xu et al., 2025c). As speech naturally contains paralinguistic cues absent in text, strict input equivalence may not be desirable. However, for complex reasoning tasks, the underlying logical progression should remain invariant regardless of the input modality (Mousavi et al., 2025). Relying solely on inputs can cause subtle discrepancies to propagate and amplify through Transformer layers, leading to significant representational drift. Another line of work provides output-level supervision. They focus on knowledge distillation or prompt-switching training to encourage speech-conditioned generations to mimic text-conditioned behaviors (Wang et al., 2025; KimiTeam et al., 2025). However, these methods enforce strict token-level supervision in an off-policy manner. Since the speech-conditioned distribution differs from the text one, forcing the model to generate exact text-conditioned tokens targets an unreachable objective. Furthermore, this rigid supervision suffers from exposure bias: a single token error shifts the model into an unsupervised state, causing it to fall into behavioral divergence.

To address these limitations, we introduce **TARS** (Trajectory Alignment for Reasoning in Speech), which combines representation alignment to mitigate drift and behavior alignment with more flexible objective. By leveraging on-policy explo-

*Work done during an MSRA internship.

†Corresponding authors.

ration, our method enables the model to mitigate exposure bias and maintain reasoning stability, preventing the model from drifting into out-of-distribution states. Specifically, representation alignment is computed from the cosine similarity of layer-wise hidden states, providing coarse-grained representation-level feedback. Complementarily, behavior alignment is derived from the semantic consistency of the final outputs, providing token-level but comparatively sparser feedback. These signals effectively steer the speech modality toward the text reasoning trajectory.

Under an asymmetric reward design with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), we jointly optimize task accuracy and these two alignment rewards. Even when all samples in the generated group exhibit zero task accuracy, a common outcome for speech-conditioned reasoning given its greater difficulty than text-based inference, the alignment signals remain informative, enabling direct trajectory alignment between speech and text. Empirical results on complex reasoning benchmarks, such as MMSU and OBQA (Chen et al., 2024), demonstrate that our method outperforms existing baselines, achieving state-of-the-art performance among 7B models. The contributions of our work are summarized as follows:

- We propose an on-policy reinforcement-learning framework for trajectory alignment that aligns speech-conditioned reasoning trajectory with its text-conditioned counterpart, closing the modality reasoning gap without architectural modifications.
- We introduce an asymmetric dense reward with two complementary alignment signals: representation alignment that reduces layer-wise hidden-state drift, and behavior alignment that enforces semantic consistency.
- Our method achieves the state-of-the-art performance on reasoning benchmarks (MMSU, OBQA) among 7B-scale models. We release datasets and code to facilitate reproducibility.¹

2 Related Works

Speech LLMs. Speech LLMs have progressed from cascaded pipelines (ASR + text LLM + optional TTS) to end-to-end architectures that couple speech perception with LLM-style generation,

enabling spoken dialogue and spoken QA while better leveraging paralinguistic cues beyond transcripts (Peng et al., 2025a; Cui et al., 2025). A dominant design follows a three-stage paradigm: a pretrained speech encoder extracts acoustic features, which are mapped into the text embedding space via lightweight projectors to condition a decoder-only LLM, preserving text-pretrained reasoning priors while extending to speech tasks (Tang et al., 2023; Chu et al., 2023; Gong et al., 2023). Recent open and proprietary omni systems further target low-latency and multi-modal interaction. Open-weight examples such as Qwen2.5-Omni and Qwen3-Omni integrate unified perception and generation, and introduce modality-specialized capacity (e.g., MoE routing) to improve scalability (Xu et al., 2025a,b). Audio-centric models like Kimi-Audio and MiniCPM-o emphasize practical voice interaction and general audio understanding (KimiTeam et al., 2025; Yao et al., 2024).

Modality Alignment and Reasoning Gap. Despite unified architectures, a performance disparity between speech and text modalities persists, termed the modality reasoning gap. Empirical studies (Chen et al., 2024; Xiaomi, 2025; Mousavi et al., 2025) and representational analyses (Xiang et al., 2025) reveal that speech-conditioned hidden states often drift from their text counterparts, leading to degraded reasoning. Existing efforts to bridge this gap generally fall into two categories: **(1) Architectural Adaptations.** To preserve text-based capabilities, a prominent line of work adopts a frozen-backbone strategy. Methods such as AlignChat (Chen et al., 2026), DeSTA (Lu et al., 2025a), OTReg (Xu et al., 2025c), and MTBI (Xie et al., 2025) keep Large Language Model (LLM) parameters fixed, focusing exclusively on refining the input-side projector. However, this yields only surface-level alignment. By freezing the backbone, the model cannot adapt to speech-specific dynamics, causing reasoning trajectories to diverge even when inputs are closely projected. **(2) Supervised Alignment Strategies.** Other works employ static matching during pre-training or supervised fine-tuning. Early methods explored manifold mixup (Fang et al., 2022) and contrastive objectives (Cappellazzo et al., 2024) to align representation spaces. Recent LLM adaptations employ joint speech-text supervised fine-tuning (Peng et al., 2025b), task-independent contrastive pretraining (Züfle and Niehues, 2025), or

¹<https://github.com/AmphionTeam/TARS>

interleaving tokens directly (Nguyen et al., 2025). To align output behaviors, approaches like Kimi-Audio (KimiTeam et al., 2025) utilize prompt-switching, while others apply cross-modal knowledge distillation (Wang et al., 2025) or data selection (Cuervo et al., 2025) to match text-based teachers. Works like SSR (Tan et al., 2025) combine both, aiming to align representation and output behavior simultaneously. However, these methods rely on off-policy supervision (e.g., cross-entropy or KL divergence) on static targets. This forces the model to mimic final answers but does not teach it how to dynamically correct its own reasoning trajectory, leading to compounding errors in complex tasks. In contrast, our reinforcement learning (RL) based framework aligns the reasoning trajectory itself via on-policy exploration.

Reinforcement Learning for Reasoning. RL has proven essential for enhancing the reasoning capabilities of LLMs beyond standard supervised fine-tuning (Liu et al., 2025a). Techniques like GRPO enable models to learn from sparse rewards and self-exploration, significantly improving performance on math and logic tasks (Shao et al., 2024; Yu et al., 2025b). In the speech domain, however, RL application remains nascent, primarily limited to aligning paralinguistic attributes or general helpfulness rather than reasoning logic (Li et al., 2025; Liu et al., 2025b). A concurrent work, SoundMind-RL (Diao et al., 2025), also applies RL to Qwen2.5-Omni for speech reasoning. Despite this, they rely on sparse, rule-based rewards (format and correctness constraints) to shape outputs. Most importantly, standard binary rewards are sparse and insufficient for modality alignment. Our work bridges this gap by adapting RL with dense alignment signals—leveraging the text modality as a stable reference to guide the speech reasoning trajectory, explicitly aligning latent spaces and mitigating representation drift.

3 Method

3.1 Problem Formulation

We define a Speech LLM π_θ as a composite architecture consisting of an audio encoder, a modality projector, and a decoder-only LLM initialized from a text-pretrained LLM π_{base} . For a given query $q \in \mathcal{D}$, the model accepts either its speech representation q_{speech} or text representation q_{text} as input to generate a text response y . Despite extensive alignment training on large-scale speech-text pairs,

a significant modality reasoning gap persists, where the model’s performance on speech inputs lags behind its text capabilities.

We quantify this gap using Modality Recovery Rate (MRR). Let $y_{\text{speech}} = \pi_\theta(q_{\text{speech}})$ be the completion generated by the current model, and $y_{\text{text}}^{\text{base}} = \pi_{\text{base}}(q_{\text{text}})$ be the reference completion from the base model. Given a reasoning metric \mathcal{S} (e.g., Question Answering (QA) accuracy), MRR measures the extent to which the Speech LLM retains the original reasoning capability:

$$\text{MRR}(\pi_\theta) = \frac{\mathbb{E}_{q \in \mathcal{D}}[\mathcal{S}(y_{\text{speech}})]}{\mathbb{E}_{q \in \mathcal{D}}[\mathcal{S}(y_{\text{text}}^{\text{base}})]} \times 100\%. \quad (1)$$

Our objective is to optimize parameters θ such that $\text{MRR} \geq 100\%$, effectively closing the gap.

3.2 Reward Modeling

We propose an asymmetric reward design to align reasoning trajectories across modalities. During training, we use text-conditioned completions y_{text} generated by the current policy π_θ as a moving reference. We optimize the policy on both text-conditioned and speech-conditioned completions, allowing the text branch to continue improving under base reward while providing an increasingly strong reference for aligning speech trajectories. As a result, the speech modality co-evolves with the model’s improving text reasoning capability.

For a speech-conditioned completion y_{speech} , the total reward is defined as:

$$R_{\text{total}} = R_{\text{base}} + \alpha \cdot R_{\text{rep}} + \beta \cdot R_{\text{beh}}, \quad (2)$$

where R_{rep} and R_{beh} correspond to representation alignment and behavior alignment signals, respectively. We apply R_{total} to speech-conditioned completions, while text-conditioned completions are optimized using R_{base} . We set $\alpha = 1.0$ and $\beta = 1.0$ in our experiments to simultaneously align internal representations and external behaviors.

Base Reward. Following the formulation in DeepSeek-R1 (DeepSeek-AI et al., 2025), we design the base reward to optimize task accuracy and output format. This configuration serves as the Standard GRPO baseline in our ablation studies. It is computed as:

$$R_{\text{base}} = R_{\text{acc}} + \lambda R_{\text{fmt}}, \quad (3)$$

where $R_{\text{acc}} \in \{0, 1\}$ indicates whether the answer extracted by xFinder (Yu et al., 2025a) matches the

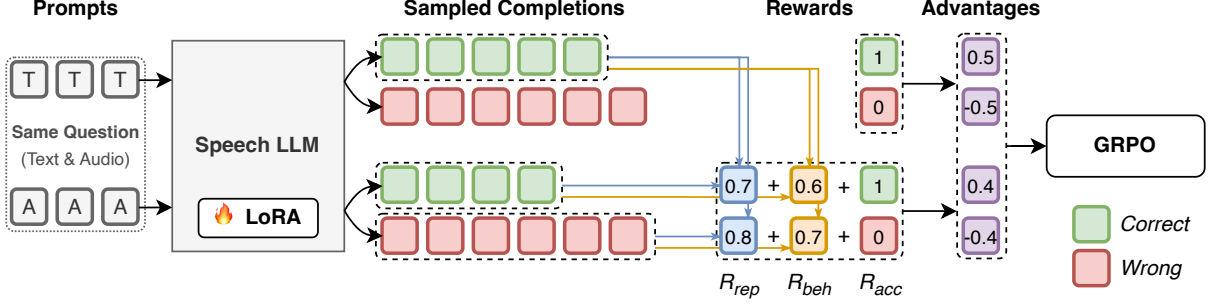


Figure 1: Overview of our framework. We introduce a reinforcement learning approach for trajectory alignment by optimizing an asymmetric reward function composed of representation alignment and behavior alignment.

ground truth, and $R_{\text{fmt}} \in \{0, 1\}$ rewards format compliance (see Appendix A). We set $\lambda = 0.5$.

Representation Alignment Reward. To align internal representations, we compute the layer-wise geometric similarity between speech completions and text references. Let $\mathbf{H}^{(l)} \in \mathbb{R}^{T \times d}$ denote the hidden states at layer l . To handle the variable sequence lengths of different modalities, we mean-pool the generated reasoning tokens (excluding n prompt tokens) into a fixed-size vector $\bar{\mathbf{h}}^{(l)}$:

$$\bar{\mathbf{h}}^{(l)} = \frac{1}{T - n} \sum_{t=n+1}^T \mathbf{H}_{t,:}^{(l)}. \quad (4)$$

For each speech completion, we randomly sample a correct text completion ($R_{\text{acc}} = 1$) from the same group as a reference, computing the average cosine similarity across L layers:

$$R_{\text{rep}} = \frac{1}{L} \sum_{l=1}^L \text{CosSim}(\bar{\mathbf{h}}_{\text{speech}}^{(l)}, \bar{\mathbf{h}}_{\text{text}}^{(l)}). \quad (5)$$

If no correct text completion exists, we set $R_{\text{rep}} = 0$. This fallback is rare (5.9%–10.7% of cases), ensuring most groups provide robust gradient signals. Ultimately, this dense reward encourages the speech modality to emulate text-based internal reasoning.

Behavior Alignment Reward. To ensure behavior consistency at the output level, we employ an external embedding model \mathcal{E} (e.g., Qwen3-Embedding-0.6B (Zhang et al., 2025)) to measure the semantic equivalence between the final generated speech completion y_{speech} and text reference y_{text}^* :

$$R_{\text{beh}} = \text{CosSim}(\mathcal{E}(y_{\text{speech}}), \mathcal{E}(y_{\text{text}}^*)). \quad (6)$$

Similarly, if no correct text completion is available, we set $R_{\text{beh}} = 0$. This objective allows the model to learn from diverse valid reasoning trajectories, provided the final semantic behavior remains consistent with the teacher.

3.3 Reinforcement Learning Framework

As illustrated in Figure 1, we employ GRPO to optimize our proposed reward. For a given prompt q , we generate a group of G completions, composed of equal numbers of speech-conditioned and text-conditioned completions. Following GRPO Done Right (Dr. GRPO) (Liu et al., 2025c), we define the advantage \hat{A}_i for the i -th completion as normalizing its reward against the group’s mean. The model’s parameters θ are updated using the DAPO loss (Yu et al., 2025b). Compared to using only the base reward, our alignment reward provides richer guidance by supplying a continuous similarity-based signal that remains effective even when task accuracy rewards are uniformly zero for speech-conditioned reasoning trajectories.

Modality Specific Normalization. A naive implementation of GRPO normalizes reward across the entire group. However, text-conditioned completions inherently achieve higher base rewards than speech-conditioned, which would cause speech-conditioned completions to consistently receive negative advantages, suppressing learning. To address this, we introduce modality-specific normalization, calculating advantages for text and speech completions in separate groups:

$$\hat{A}_{i,m} = r_{i,m} - \mu_m, \quad m \in \{\text{speech}, \text{text}\}, \quad (7)$$

where μ_m is the mean of rewards within modality m . This ensures that each modality is optimized relative to its own baseline, allowing continuous improvement in modality alignment.

4 Experiments

We evaluate whether our approach reduces the speech-text modality reasoning gap. Performance is reported on Audio (A) and Text (T) inputs, using MRR (Equation 1) to quantify reasoning recovery in speech modality. Our method centers on an alignment-aware RL framework with asymmetric rewards and modality-specific normalization to optimize reasoning.

4.1 Experimental Setup

Foundation Models. We evaluate two ~ 7 B composite Speech LLMs (Section 3.1): the officially released Qwen2.5-Omni and an internal Phi-4-Multimodal-Instruct (Phi-4-MM)² built on the Phi-4-7B backbone with a lightweight speech encoder. All post-training experiments share these initial checkpoints to ensure fair comparison.

Training Data. We use the UnifiedQA training set (Khashabi et al., 2020)³ as our primary dataset. To construct paired speech-text inputs, we synthesize speech using two high-quality TTS systems: CosyVoice2 (Du et al., 2024) and openaudio-s1-mini⁴, with reference speakers sampled from the Emilia-YODAS EN subset (He et al., 2025). After filtering the synthesized speech with a 10% WER threshold via whisper-medium (Radford et al., 2023), our final training set comprises 9,953 samples (203 hours).

Evaluation Benchmarks. We evaluate on two spoken multiple-choice QA benchmarks from VoiceBench (Chen et al., 2024): MMSU (3,074 examples derived from MMLU-Pro (Wang et al., 2024)) for multi-domain knowledge and reasoning, and OBQA (455 examples) for elementary science facts and commonsense reasoning. We also report WER on LibriSpeech (Panayotov et al., 2015) (test-clean & test-other) to diagnose ASR capabilities, obtained by prompting the model to transcribe the speech and computing WER with greedy decoding.

Baselines. All post-training baselines share the same foundation models, data, and LoRA configurations. We compare against three categories (details in Appendix B): (1) **Cascaded & Proprietary**

²Same architecture as the official 5.6B release, but scaled to 7B.

³Specifically, the auxiliary_train split from the MMLU repository (<https://huggingface.co/datasets/cais/mmlu>).

⁴<https://huggingface.co/fishaudio/openaudio-s1-mini>

systems: whisper-large-v3 paired with corresponding text backbones, and GPT-4o-mini-Audio. (2) **Cross-modal alignment:** SALAD (Cuervo et al., 2025), DeSTA2.5-Audio (Lu et al., 2025b), AlignChat (Chen et al., 2026), and Knowledge Distillation (KD) (Wang et al., 2025), which explicitly align speech and text via distillation or adapter tuning. (3) **General post-training:** Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Standard GRPO, which optimize for task accuracy without explicitly addressing the modality gap.

RL Training Protocol. We follow the GRPO training protocol with a group size of $G = 8$, comprising equal speech- and text-conditioned completions per prompt. During RL (Section 3.2), text completions from the current policy serve as alignment references (teacher signals) and also participate in gradient updates. Finally, we compute advantages with modality-specific normalization to prevent consistently negative advantages for speech completions (Section 3.3).

Implementation Details. Training is conducted via ms-swift⁵ for Qwen2.5-Omni and HuggingFace TRL⁶ for Phi-4-MM. We apply parameter-efficient fine-tuning using LoRA to all linear layers while freezing the audio encoder and projector. For RL, we use the DAPO loss estimator (Yu et al., 2025b) (Section 3.3) and sampling decoding to encourage exploration. Greedy decoding is used for deterministic evaluation, with xFinder (Yu et al., 2025a) extracting predicted options to reduce formatting noise. Training requires ~ 55 hours for Qwen2.5-Omni and ~ 35 hours for Phi-4-MM on $4 \times A100$ or $8 \times H200$ GPUs (details in Appendices C and D).

4.2 Main Results

Table 1 presents the performance of our framework against a suit of baselines. As shown by the base Speech LLMs, performance under speech inputs consistently lags behind text, revealing a clear modality reasoning gap. Existing cross-modal alignment baselines can narrow this gap, yet most still fall short of full recovery with $MRR < 100\%$. Our proposed framework demonstrates state-of-the-art performance among 7B models. For the Qwen2.5-Omni with Qwen2.5-7B as backbone, our approach achieves an average audio accuracy of

⁵<https://github.com/modelscope/ms-swift>

⁶<https://github.com/huggingface/trl>

Model	Backbone	MMSU		OBQA		Average		MRR (%)
		A	T	A	T	A	T	
<i>Proprietary & Cascaded Systems</i>								
GPT-4o-mini-Audio	GPT-4o-mini	72.90	81.23	84.84	90.11	78.87	<u>85.67</u>	92.06
ASR [†] + Llama3.1-8B	Llama3.1-8B	58.78	65.65	72.53	80.88	65.66	<u>73.27</u>	89.61
ASR [†] + Qwen2.5-7B	Qwen2.5-7B	67.1	71.65	84.0	83.74	75.55	<u>77.70</u>	97.23
ASR [†] + Phi-4-7B	Phi-4-7B	69.00	74.92	77.80	83.96	73.40	<u>79.44</u>	92.40
<i>Existing Baselines</i>								
DeSTA2.5-Audio	Llama3.1-8B*	60.87	65.65	74.06	80.88	67.47	73.27	92.08
SALAD-7B	Qwen2.5-7B	57.5	71.6	75.1	90.1	66.30	80.85	85.33
MiniCPM-o 2.6	Qwen2.5-7B	54.78	59.42	78.02	82.86	66.40	71.14	85.46
Knowledge Distillation	Qwen2.5-7B	63.09	69.15	82.64	84.62	72.87	76.89	93.78
AlignChat	Qwen2.5-7B*	69.65	71.65	85.49	83.74	77.57	77.70	99.83
<i>Base & Aligned Models</i>								
Qwen2.5-Omni	Qwen2.5-7B	61.51	67.94	81.09	84.40	71.30	76.17	91.76
Phi-4-MM	Phi-4-7B	54.81	72.15	71.65	84.62	63.23	78.39	79.59
TARS (Qwen2.5-Omni)	Qwen2.5-7B	67.96	68.54	85.71	88.57	76.84	78.56	98.89
TARS (Phi-4-MM)	Phi-4-7B	70.14	75.76	89.45	91.87	79.80	83.82	100.45

Table 1: Reasoning Benchmarks Results. Accuracy (%) on MMSU and OBQA are reported using VoiceBench evaluator for Audio (A) and Text (T) input. Underlined scores denote the values used as the denominator when computing MRR. [†] cascaded systems; * frozen LLM backbone. Cited results keep original precision.

76.84%, substantially outperforming other end-to-end alignment methods such as SALAD (66.30%) and MiniCPM-o (66.40%). It also reaches an MRR of **98.89%**. Our RL-based method proves more effective than supervised mimicry, surpassing the Knowledge Distillation (KD) baseline (72.87%) by a large margin. For the Phi-4-MM with Phi-4-7B as backbone, our method achieves the best performance with an accuracy of **79.80%**, even surpassing the original text accuracy of 78.39% and achieving the MRR = **100.45%**.

These results show that our method not only narrows the modality reasoning gap, but also improves text performance, from 76.17% to 78.56% for Qwen2.5-Omni and from 78.39% to 83.82% for Phi-4-MM, indicating that gains in speech are not obtained at the expense of text reasoning, instead, the knowledge learned from speech can further strengthen text-based reasoning. Notably, the improved text accuracy remains higher than the corresponding audio accuracy (76.84% and 79.80% respectively), suggesting that residual differences are likely due to imperfect speech representations and cross-modal projection noise, making text inputs a natural upper bound.

Cascaded systems are often considered strong baselines and can outperform end-to-end models. However, our end-to-end models on Qwen2.5-

Omni (76.84%) and Phi-4-MM (79.80%) exceed the performance of the ASR + Qwen2.5-7B pipeline (75.55%) and ASR + Phi-4-7B (73.40%), respectively. This suggests that directly processing speech signals can avoid certain ASR-induced errors, leading to a more robust reasoning process.

4.3 Effectiveness of Training Strategies

Table 2 compares different training strategies on the same backbone (Phi-4-MM), including inference-time prompting, SFT, DPO, Standard GRPO, and our method. Chain-of-Thought (CoT) prompting yields a clear gain on speech inputs, improving the average audio accuracy from 63.16% to 70.06% and increasing MRR from 79.60% to 88.29%. However, it is unable to fully eliminate the gap, suggesting that prompting alone is insufficient to resolve the cross-modal misalignment.

Post-training with supervised or preference-based objectives further improves performance, yet still falls short of full recovery. SFT and DPO raise the average audio accuracy to 72.52% and 75.37%, respectively. This indicates that while supervision and preference optimization help, they do not explicitly align cross-modal reasoning behavior. Standard GRPO, trained with the base reward R_{base} , provides additional improvements (MRR = 92.21%) but still underperforms DPO, highlighting

Method	MMSU		OBQA		Average		MRR (%)	WER (%)
	A	T	A	T	A	T		
<i>Inference Baseline</i>								
Phi-4-MM	54.00	71.15	72.31	85.05	63.16	78.10	79.60	4.16
+ CoT Prompting	60.77	70.85	79.34	86.15	70.06	78.50	88.29	-
<i>Post-training Alignment</i>								
SFT	63.50	70.85	81.54	87.47	72.52	79.16	91.37	4.18
DPO	66.33	74.72	84.40	91.43	75.37	83.08	94.98	4.23
<i>Reinforcement Learning</i>								
Text-only GRPO	67.79	76.97	85.93	90.55	76.86	83.76	96.86	4.18
Standard GRPO (R_{base})	63.04	72.54	83.30	89.45	73.17	81.00	92.21	4.24
+ Representation (R_{rep})	66.82	76.09	84.84	88.35	75.83	82.22	95.56	4.18
+ Behavior (R_{beh})	69.55	76.19	87.91	90.99	78.73	83.59	99.22	4.20
TARS	69.90	75.47	89.23	91.65	79.57	83.56	100.28	4.20

Table 2: Analysis of Training Strategies and Reward Components. Comparisons on the Phi-4-MM backbone using xFinder evaluator. WER reports the average Word Error Rate (\downarrow) on Librispeech.

the limitation of sparse, outcome-centric rewards. In contrast, our approach achieves the highest performance, reaching **79.57%** average audio accuracy and $\text{MRR} = \mathbf{100.28\%}$, demonstrating the effectiveness of proposed asymmetric dense alignment reward. Finally, we monitor ASR-related capability via WER and observe it remains unchanged ($\approx 4.16\% - 4.24\%$), supporting the conclusion that gains primarily stem from reasoning alignment rather than improved speech recognition.

4.4 Reward Components

Table 2 presents an ablation study on Phi-4-MM backbone, starting from Standard GRPO trained with the base reward R_{base} only, and then adding R_{rep} , R_{beh} , or their combination. Incorporating representation alignment reward consistently improves performance, increasing MRR from 92.21% to 95.56%, suggesting that aligning layer-wise hidden-state representations provides a denser signal for RL. Alternatively, adding the behavior alignment reward pushes the model close to full recovery ($\text{MRR} = 99.22\%$), indicating that semantic-consistency supervision constrains speech outputs toward correct text-conditioned behaviors. Combining both rewards achieves the best result ($\text{MRR} = 100.28\%$), showing that representation and behavior signals are complementary: representation alignment mitigates representation drift, while behavior alignment enforces semantic target consistency. This validates our objective of jointly aligning internal representations and external behaviors.

5 Discussion and Analysis

5.1 Layer-wise Trajectory Alignment

To investigate the inner workings of our trajectory alignment, we analyze the cosine similarity between the internal hidden states of text and speech modalities (detailed in Appendix E). Specifically, we pass identical queries—one as text and one as audio—through both the baseline and aligned models. For each layer l , we compute the average cosine similarity between the sequence of hidden states $\mathbf{H}_{\text{text}}^{(l)}$ and $\mathbf{H}_{\text{speech}}^{(l)}$.

Figure 2 presents a layer-wise representation alignment analysis on the Phi-4-MM backbone, measuring the cosine similarity between the mean hidden states of correct text and speech trajectories across layers. Under this setting, similarity naturally decreases with depth due to compounding transformations, where minor modality-specific differences accumulate as they propagate through the network. The key metric is not the downward trend, but the relative separation between methods across depth. Adding R_{rep} consistently lifts the similarity trajectory across layers, indicating that representation alignment reduces the representational drift. Furthermore, our joint strategy achieves the highest similarity, suggesting that behavior alignment acts as a complementary constraint that guides the speech branch toward semantically consistent reasoning paths. These internal observations align with the external improvements in MRR, supporting that our method performs genuine reasoning

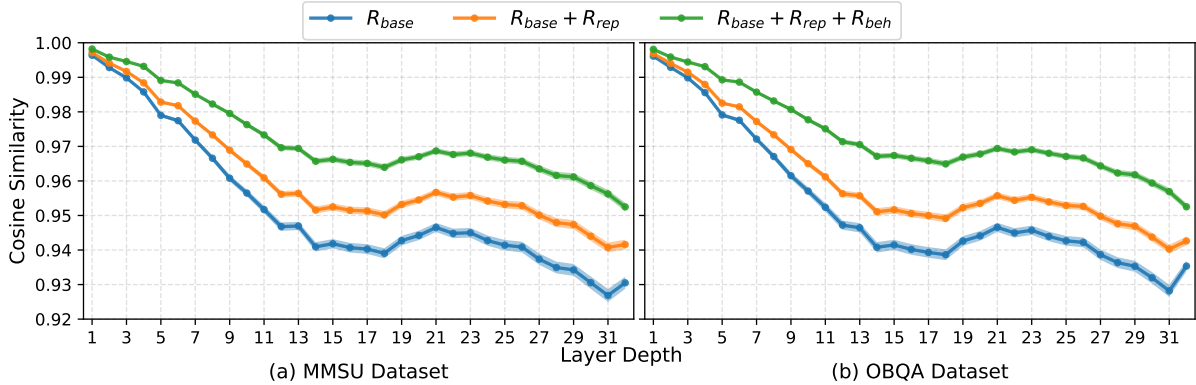


Figure 2: Layer-wise Representation Alignment Analysis. Shaded areas indicate 95% confidence intervals.

behavior transfer. The consistency of this effect across MMSU and OBQA further validates its robustness.

5.2 Layer Sensitivity

Figure 3 investigates the sensitivity of the representation-alignment reward to different depths within the 32-layer Phi-4-MM backbone. We partition the model into Shallow (layers 1–10), Middle (11–20), Deep (21–30), and Last (31–32) groups, comparing these against an All-layer baseline. Results indicate that the Middle layers are the most critical localized region, achieving **75.48%** accuracy. Conversely, applying the reward exclusively to Shallow or Last layers is less effective. This suggests that representation drift is primarily in the mid-to-late reasoning stages, rather than during early perceptual processing or final logit alignment. Selecting All layers yields the highest accuracy (**75.83%**), this confirms that global cross-layer similarity offers advantages over localized alignment.

5.3 Impact of Alignment Stage on Reward Design

Ablation on Qwen2.5-Omni (Table 3) reveals a stage-dependent regularization effect. As an internal model with minimal post-training (e.g., SFT only), Phi-4-MM is highly receptive to explicit layer-wise trajectory guidance (R_{rep}) to shape its reasoning pathways. In contrast, applying strict hidden-state constraints (R_{rep}) to a highly optimized model like Qwen2.5-Omni causes over-regularization. It forces the speech representations to rigidly mimic text representations at every layer, stripping the model of the necessary internal degrees of freedom to process modality-specific features. This phenomenon is broadly consistent with findings from knowledge distillation for genera-

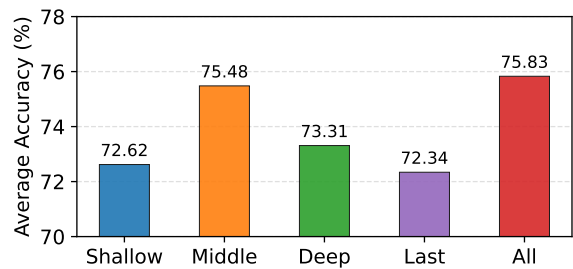


Figure 3: Sensitivity Analysis of Representation Reward Layers. Average audio accuracy on MMSU and OBQA across different layer groups.

tive models, which suggest that overly rigid imitation objectives can be suboptimal under capacity limitation, while more flexible output-level guidance better accommodates the model’s own representational constraints (Gu et al., 2024). Consequently, relying solely on the behavior reward (R_{beh}) yields the optimal guidance for Qwen2.5-Omni, as it enforces semantic consistency at the output while allowing flexible internal reasoning pathways, achieving the highest average audio accuracy (77.71%) and MRR (100.01%).

Method	Avg. A	MRR (%)
Baseline (w/ CoT)	74.50	95.88
Standard GRPO (R_{base})	77.10	99.23
+ R_{rep}	76.45	98.39
+ R_{beh}	77.71	100.01
+ Both ($R_{rep} + R_{beh}$)	76.84	98.89

Table 3: Ablation study on Qwen2.5-Omni.

5.4 Preservation of Text Reasoning Capability

To determine whether simultaneous speech alignment compromises text reasoning, we conducted

a “Text-only GRPO” ablation on the Phi-4-MM backbone, optimizing solely on text inputs with the base reward (R_{base}). As shown in Table 2, Text-only GRPO yields an average text accuracy of 83.76%, nearly identical to our joint alignment method (83.56%). This demonstrates that aligning speech trajectories does not interfere with the model’s text capabilities. Furthermore, our method achieves higher speech reasoning accuracy (79.57% vs. 76.86%) and modality recovery (MRR: 100.28% vs. 96.86%). These results confirm that our framework effectively bridges the modality gap while preserving text performance.

5.5 Generalization to Real Speech and Open-Ended Tasks

To verify that our alignment framework generalizes to real human speech and preserves the model’s ability to reason over rich paralinguistic features, we evaluate on two additional benchmarks. We select SD-QA (Faisal et al., 2021), a reference-based open-ended QA dataset featuring real spoken questions with diverse accents, and the MMSU benchmark (WANG et al., 2026)⁷, which assesses fine-grained perception of paralinguistic features like prosody. As shown in Table 4, our method consistently outperforms both the Phi-4-MM baseline and standard CoT prompting. Notably, while CoT prompting degrades fine-grained acoustic reasoning (dropping MMSU accuracy to 51.50%), our alignment framework avoids this regression, achieving 57.40% on MMSU Audio and 41.95% on SD-QA Audio. These results demonstrate that aligning reasoning trajectories preserves speech-specific attributes like prosody and accents, rather than overwriting them with text-based patterns, enabling robust generalization to open-ended, real-world spoken interactions.

Method	MMSU	SD-QA	
	A	A	T
Phi-4-MM	56.70	39.42	41.77
+ CoT Prompting	51.50	39.60	45.21
TARS	57.40	41.95	44.30

Table 4: Generalization performance on paralinguistic perception (MMSU) and real-speech QA (SD-QA).

⁷Note that this MMSU benchmark differs from the VoiceBench MMSU split used in our main experiments.

5.6 Domain-Specific Error Analysis

To understand how our alignment mitigates representational drift under a free-generation setting, we conducted a per-domain breakdown on the 12 domains of the MMSU dataset (see Appendix F for full score details). Our method yields consistent improvements across all domains. In high-performing tasks (e.g., Business, Biology, Law), the aligned model demonstrates enhanced arithmetic calculation, logical consistency, and domain knowledge. Conversely, in lower-performing domains (e.g., Economics, Engineering, History), errors often stem from subtle semantic traps or chronological retrieval failures. Detailed qualitative CoT traces and specific case studies are provided in Appendix H.

5.7 Computational Efficiency

Analysis of computational efficiency shows that our online RL approach introduces only a marginal 4.4% overhead compared to standard GRPO, while achieving better performance (Appendix G).

6 Conclusion

We introduced an on-policy trajectory alignment framework that mitigates representational drift and improves semantic consistency in speech reasoning. By combining dense representation and behavior alignment rewards under an asymmetric RL objective, our method substantially narrows the modality reasoning gap and achieves state-of-the-art performance on MMSU and OBQA among 7B-scale Speech LLMs.

Limitations

Despite its effectiveness, our trajectory alignment framework has several limitations. First, we evaluate alignment only at the 7B scale, and it remains unclear how the proposed reward design behaves for smaller or larger models. Second, our method focuses on single-turn reasoning, whereas multi-turn, interactive, or dialogue-driven speech reasoning may introduce additional dynamics not captured by our current formulation. Finally, while our alignment rewards mitigate modality drift, they still rely on text-only reference completions and may not fully account for paralinguistic cues, such as emotion, prosody, or intent, that do not have explicit textual counterparts.

References

- Umberto Cappellazzo, Enrico Fini, Muqiao Yang, Daniele Falavigna, Alessio Brutti, and Bhiksha Raj. 2024. [Continual contrastive spoken language understanding](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3727–3741, Bangkok, Thailand. Association for Computational Linguistics.
- Wei Chen, Yunzhong Zheng, Tu Zheng, Deng Cai, Wenxiao Wang, and Jieping Ye. 2026. [Alignchat: Endowing LLMs with end-to-end speech-to-text chat capability through token-level representation alignment](#).
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. [Voicebench: Benchmarking llm-based voice assistants](#). [arXiv preprint arXiv:2410.17196](#).
- 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](#).
- Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. 2025. [Closing the gap between text and speech understanding in llms](#). [arXiv preprint arXiv:2510.13632](#).
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. [Recent advances in speech language models: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Yan, Xikang Hao, Xin Kuang, Xinyu Yang, Xuan Dong, Xuehui Yu, Yajun Zheng, Yanjun Zou, Yanquan Zhu, Yashin Chen, Y. K. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yao-hui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). [arXiv preprint arXiv:2501.12948](#).
- Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiwei Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025. [SoundMind: RL-incentivized logic reasoning for audio-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 528–540, Suzhou, China. Association for Computational Linguistics.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. [Cosyvoice 2: Scalable streaming speech synthesis with large language models](#). [arXiv preprint arXiv:2412.10117](#).
- Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. [SD-QA: Spoken dialectal question answering for the real world](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. [STEMM: Self-learning with speech-text manifold mixup for speech translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

- Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. [arXiv preprint arXiv:2305.10790](#).
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. [MiniLLM: Knowledge distillation of large language models](#). In [The Twelfth International Conference on Learning Representations](#).
- Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2025. [Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation](#). [IEEE Transactions on Audio, Speech and Language Processing](#), 33:4044–4054.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. [arXiv preprint arXiv:2005.00700](#).
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. Kimi-audio technical report. [arXiv preprint arXiv:2504.18425](#).
- Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. 2025. Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. In [Findings of the Association for Computational Linguistics: EMNLP 2025](#), pages 18744–18754.
- Keliang Liu, Dingkan Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025a. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. [arXiv preprint arXiv:2509.16679](#).
- Yansong Liu, Jiateng Li, and Yuan Liu. 2025b. Enhancing speech large language models through reinforced behavior alignment. [arXiv preprint arXiv:2509.03526](#).
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. [Understanding r1-zero-like training: A critical perspective](#). In [Second Conference on Language Modeling](#).
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025a. Developing instruction-following speech language model without speech instruction-tuning data. In [ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing \(ICASSP\)](#), pages 1–5. IEEE.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, Yi-Cheng Lin, Yu-Xiang Lin, Chi-An Fu, Chun-Yi Kuan, Wenze Ren, Xuanjun Chen, Wei-Ping Huang, En-Pei Hu, Tzu-Quan Lin, Yuan-Kuei Wu, Kuan-Po Huang, Hsiao-Ying Huang, Huang-Cheng Chou, Kai-Wei Chang, Cheng-Han Chiang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025b. [Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment](#). [arXiv preprint arXiv:2507.02768](#).
- Pooneh Mousavi, Yingzhi Wang, Mirco Ravanelli, and Cem Subakan. 2025. [Alas: Measuring latent speech-text alignment for spoken language understanding in multimodal llms](#). [arXiv preprint arXiv:2505.19937](#).
- Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. [SpiRit-LM: Interleaved spoken and written language model](#). [Transactions of the Association for Computational Linguistics](#), 13:30–52.
- 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. IEEE.
- Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Han-kun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2025a. [A survey on speech large language models for understanding](#). [IEEE Journal of Selected Topics in Signal Processing](#), pages 1–32.
- Yifan Peng, Krishna C Puvvada, Zhehuai Chen, Piotr Zelasko, He Huang, Kunal Dhawan, Ke Hu, Shinji Watanabe, Jagadeesh Balam, and Boris Ginsburg. 2025b. [VoiceTextBlender: Augmenting large language models with speech capabilities via single-stage joint speech-text supervised fine-tuning](#). In [Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 5787–5802, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Weiting Tan, Hirofumi Inaguma, Ning Dong, Paden D. Tomasello, and Xutai Ma. 2025. *SSR: Alignment-aware modality connector for speech language models*. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 56–75, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmon: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.
- Dingdong WANG, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen M. Meng. 2026. *MMSU: A massive multi-task spoken language understanding and reasoning benchmark*. In *The Fourteenth International Conference on Learning Representations*.
- Enzhi Wang, Qicheng Li, Zhiyuan Tang, and Yuhang Jia. 2025. Cross-modal knowledge distillation for speech large language models. *arXiv preprint arXiv:2509.14930*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. *Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202, Suzhou, China. Association for Computational Linguistics.
- LLM-Core-Team Xiaomi. 2025. *Mimo-audio: Audio language models are few-shot learners*.
- Jingran Xie, Xiang Li, Hui Wang, Yue Yu, Yang Xiang, Xixin Wu, and Zhiyong Wu. 2025. Enhancing generalization of speech large language models with multi-task behavior imitation and speech-text interleaving. *arXiv preprint arXiv:2505.18644*.
- 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, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Wenze Xu, Chun Wang, Jiazhen Yu, Sheng Chen, Liang Gao, and Weihong Deng. 2025c. Optimal transport regularization for speech text alignment in spoken language models. In *Asian Conference on Pattern Recognition*, pages 280–294. Springer.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. 2025a. *xfinder: Large language models as automated evaluators for reliable evaluation*. In *International Conference on Representation Learning*, volume 2025, pages 59850–59892.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiase Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Maike Züfle and Jan Niehues. 2025. *Contrastive learning for task-independent SpeechLLM-pretraining*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8469–8490, Vienna, Austria. Association for Computational Linguistics.

A Format Reward Regular Expression

The format reward, R_{format} , is a binary reward assigned based on a regular expression match. The regular expression is:

```
^<think>.*?</think>\s*<answer>.*The
→ answer is [ABCD][\.,,].*</answer>$
```

B Baseline Details

We provide detailed descriptions of the baselines compared in our experiments.

Cascaded Systems. We construct pipeline baselines using whisper-large-v3 as the ASR front-end, paired with Llama3.1-8B, Qwen2.5-7B, and Phi-4-7B (the text backbone π_{base} of Phi-4-MM). For audio evaluation, we transcribe the speech using the ASR model and feed the transcripts into the corresponding text LLM, reflecting the impact of ASR errors on reasoning. For text evaluation, we feed the clean text to the LLM.

Cross-modal Alignment Methods. SALAD (Cuervo et al., 2025) focuses on sample-efficient distillation or targeted data selection to improve speech-text alignment. DeSTA2.5-Audio (Lu et al., 2025b) utilizes self-generated text completion as the alignment target for cross-modal consistency. AlignChat (Chen et al., 2026) represents a frozen-backbone method, focusing exclusively on the alignment of speech adapters without altering the backbone LLM parameters. Knowledge Distillation (KD) (Wang et al., 2025) serves as a distillation-based transfer baseline, mimicking the behavior from a text teacher to a speech student.

General Post-training Methods. For SFT and DPO, we construct preference data via reject sampling. SFT trains only on the chosen completions, while DPO leverages chosen-rejected pairs. Standard GRPO corresponds to the base-reward-only RL baseline (as described in Section 3.3), which optimizes task performance without explicitly modeling the modality gap.

C Hyperparameters

Table 5 lists the detailed hyperparameters used for the SFT and RL training stages across all experiments. We utilize LoRA for efficient fine-tuning to reduce computational overhead.

Hyperparameter	Value
<i>LoRA Configuration</i>	
Rank (r)	8
Alpha (α)	32
Dropout	0.05
Target Modules	All Linear Layers
<i>Training Configuration</i>	
Learning Rate	1e-5 (Qwen), 2e-5 (Phi)
LR Scheduler	Cosine
Warmup Ratio	0.01
Num Epochs	3
Batch Size (Global)	64
Gradient Accumulation	4
Optimizer	AdamW
Weight Decay	0.01
Max Grad Norm	1.0
Precision	bfloat16
<i>GRPO / DAPO Configuration</i>	
Generations per Prompt (G)	8
Temperature	1.0
Max Completion Length	1024
Epsilon High (ϵ_{high})	0.28
KL Coefficient (β)	0.0

Table 5: Detailed hyperparameters for training.

D Prompt Templates

We employ same prompt format across all experiments. This instruction enforces a Chain-of-Thought reasoning structure. Table 6 illustrates the specific templates used. The content of [QUESTION TEXT] (and the spoken content of <|audio_1|>) follows the multiple-choice format:

```
{question}
```

```
Option A: {option_a}
```

```
Option B: {option_b}
```

```
Option C: {option_c}
```

```
Option D: {option_d}
```

E Layer-wise Analysis Methodology

For the layer-wise representation alignment analysis presented in Section 5.1, we perform a teacher-forcing analysis to ensure identical token alignment. We feed the exact same text-conditioned generated CoT response tokens to both the text- and speech-conditioned branches. We align the response span starting from the last <|assistant|> token and compute the cosine similarity of hidden states at each corresponding token position for every layer. The reported curves in Figure 2 are obtained by averaging these similarities over all valid tokens and test samples, accompanied by 95% confidence intervals, evaluated on both the MMSU and OBQA datasets.

Component	Content
System	< system >A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think>[THINKING PROCESS]</think><answer>The answer is [CHOICE].</answer>< end >
User (Audio)	< user >< audio_1 >< end >
User (Text)	< user >[QUESTION TEXT]< end >
Assistant	< assistant >

Table 6: Prompt templates for training and inference.

F Per-domain Breakdown Analysis

Table 7 presents the detailed per-domain accuracy breakdown on the MMSU dataset, demonstrating that our trajectory alignment method consistently improves performance across all 12 categories.

Category	Count	Baseline (w/ CoT)	TARS	Delta
Business	236	61.02%	75.00%	+13.98%
Biology	172	69.77%	83.72%	+13.95%
Law	51	52.94%	66.67%	+13.73%
Physics	383	61.36%	73.89%	+12.53%
Health	406	56.65%	66.75%	+10.10%
Philosophy	305	46.89%	55.08%	+8.20%
Chemistry	167	65.87%	73.05%	+7.19%
Psychology	317	64.04%	70.66%	+6.62%
Economics	280	74.64%	80.71%	+6.07%
Other	546	60.99%	66.85%	+5.86%
History	104	51.92%	57.69%	+5.77%
Engineering	107	56.07%	61.68%	+5.61%

Table 7: Per-domain breakdown on MMSU Audio Accuracy.

G Training Efficiency

We provide a runtime and memory analysis of our method compared to offline strategies on $8 \times H200$ GPUs for 3 epochs. As shown in Table 8, standard SFT and DPO require approximately 1.6 hours for training. However, this excludes the substantial offline data generation cost, which exceeds 10.9 GPU-hours to generate 8 rollouts per prompt. In contrast, our method operates entirely online. Compared to the direct Standard GRPO baseline (27.5 hours), computing our alignment rewards introduces only a marginal computational overhead of 4.4%, totaling 28.7 hours. Furthermore, our method maintains an

identical peak memory footprint of 90% relative to the GRPO baseline.

Method	Peak Memory	Data Gen. Time	Training Time
SFT	98%	10.9h*	1.6h
DPO	95%	10.9h*	1.6h
Standard GRPO	90%	N/A	27.5h
TARS	90%	N/A	28.7h

Table 8: Efficiency comparison. *Excludes additional time required for reward scoring.

H Qualitative Analysis

Free-Generation Alignment Analysis. To investigate how our method performs under a free-generation setting without teacher-forcing, we analyze the reward curves during the online RL training process. Since the training rollouts are generated autoregressively by the policy, these curves intrinsically reflect the model’s performance under free generation. As shown in Figure 4, the upward trends of the base reward (R_{base}), representation alignment reward (R_{rep}), and behavior alignment reward (R_{beh}) demonstrate consistent improvements in base task accuracy, as well as in the representation similarity and semantic consistency between speech and text modalities.

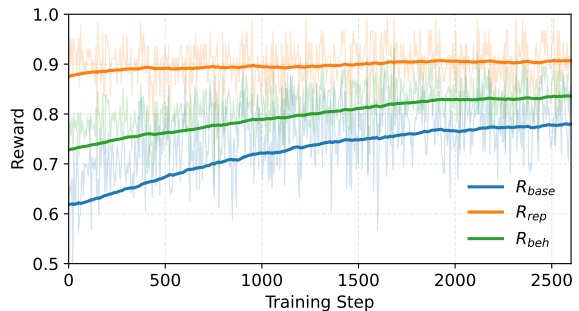


Figure 4: Reward curves during training.

Training Improvements. Table 9 shows specific qualitative improvements. The trained model exhibits better math calculation, logical reasoning, and domain knowledge, successfully correcting previous hallucinations and reasoning failures.

Persistent Errors. Table 10 analyzes cases where the model fails to reach the ground truth after training, reflecting the challenges in lower-performing domains. This reveals that failures often stem from subtle semantic traps, conflicting domain logic, or chronological retrieval failures.

Question	Before Training	After Training	Analysis
[Business] Paul took a loan of \$384.75 for 60 days at 6% interest. How much interest did he pay? (GT: B. \$3.85)	<think> ...we need to convert the time to years... plug in the values and solve for the interest. </think> <answer> The answer is (A) \$5.75. </answer>	<think> ... $I = \$384.75 \times 0.06 \times (60/365)$... $I \approx \$3.79$... The closest option to our calculated interest of \$3.79 is option B, which is <u>\$3.85</u> . </think> <answer> The answer is B: <u>\$3.85</u> . </answer>	Math Calculation: Model executed detailed arithmetic steps and correctly matched the closest option despite rounding variations.
[Biology] Which of the following about meiosis is <u>NOT</u> true? (GT: C. Sister chromatids separate during meiosis 1.)	<think> ...Option C says sister chromatids separate during meiosis I. <u>This is not true</u> ... I will choose option A as the final answer, because it is the most straightforward mistake. </think> <answer> <u>A</u> </answer>	<think> ...C. Sister chromatids separate during meiosis I - <u>This is not true</u> . Sister chromatids do not separate during meiosis I; instead, homologous chromosomes separate... </think> <answer> The answer is <u>C</u> . </answer>	Logical Reasoning: Model resolved the logical conflict in “NOT/EXCEPT” questions, consistently linking its evaluation to the final answer.
[Law] Hume’s attack on natural law is founded on his argument that: (GT: A. We cannot objectively know what is morally right or wrong.)	<think> ...Hume was a skeptic... did not focus on protection... best answer reflects his skepticism... </think> <answer> <u>B. Natural law is backward-looking.</u> </answer>	<think> ...He questioned the idea that there are objective moral truths that can be known a priori... representation of Hume’s attack... is his argument that we cannot objectively know what is morally right or wrong... </think> <answer> The answer is <u>A</u> . </answer>	Domain Knowledge: Model corrected its domain knowledge, replacing a generic guess with the specific epistemological argument of the philosopher.

Table 9: Qualitative analysis of training improvement. Bold text denotes the question category. GT denotes Ground Truth. Underlined text highlights improvements.

Question	Before Training	After Training	Analysis
[Economics] If the reserve ratio is 10 percent and a new customer deposits \$500, what is the maximum amount of money created? (GT: A. \$4,500)	<think> ... so the money multiplier is $1/0.10 = 10$. The maximum amount of money created is the initial deposit multiplied by the multiplier... $\$500 \times 10 = \5000 . </think> <answer> <u>B) \$5000</u> </answer>	<think> ...the money multiplier would be 1 divided by 0.10 ... equals 10. ...the maximum amount of money that can be created is the initial deposit multiplied by the money multiplier... $\$500 \times 10 = \5000 ... </think> <answer> The answer is <u>B: \$5000</u> . </answer>	Logical Trap: It correctly calculates the <i>total</i> money supply: \$5000, but fails by missing the nuance: “new money <i>created</i> ”, which requires subtracting the deposit: $\$5000 - \$500 = \$4500$.
[Engineering] The resistance in the circuit of the moving coil of a dynamometer wattmeter should be: (GT: C. High)	<think> ...the moving coil is used to <u>measure the current</u> ... The resistance should be low to minimize power loss... </think> <answer> <u>(B) very low</u> </answer>	<think> ...The moving coil should have a very low resistance ... the power consumed ... If the resistance were high, it would consume more power... <u>A resistance that is almost zero would be ideal</u> ... </think> <answer> The answer is <u>B: very low</u> . </answer>	Logical Error: The model wrongly assumes low resistance prevents power consumption. Because the moving (voltage) coil is connected in parallel, $P = V^2/R$ governs; thus, a low resistance would actually <i>increase</i> power draw and cause a short.
[History] Which of the following is the last to be occupied by human beings? (GT: D. New Zealand)	<think> ...Hawaii: ...arrived in Hawaii around <u>1,500 years ago</u> ... New Zealand: ...arrived in New Zealand around <u>1,000 years ago</u> ... the last place to be occupied... <u>is Hawaii</u> . </think> <answer> <u>C</u> </answer>	<think> ... <u>Hawaii</u> ... settled ... around <u>300-500 CE</u> ... <u>New Zealand</u> ... settled ... around <u>1300 CE</u> , which is ... <u>earlier than Hawaii</u> . Comparing these dates... <u>Hawaii was the last</u> to be settled... </think> <answer> The answer is <u>C: Hawaii</u> . </answer>	Chronological Error: The model retrieves the correct historical dates (New Zealand at 1300 CE vs. Hawaii at 300-500 CE) but spectacularly fails at basic temporal comparison, wrongly deducing that 1300 CE is “earlier” than 300 CE.

Table 10: Qualitative analysis of persistent errors. GT denotes Ground Truth. Underlined text highlights incorrect reasoning steps.