SICL-AT: Another way to adapt Auditory LLM to low-resource task

Anonymous ACL submission

Abstract

Auditory Large Language Models (LLMs) have demonstrated strong performance across a wide range of speech and audio understanding tasks. Nevertheless, they often struggle when applied to low-resource or unfamiliar tasks. In case of labeled in-domain data is scarce or mismatched to the true test distribution, direct fine-tuning can be brittle. In-Context Learning (ICL) provides a training-free, inference-time solution by adapting auditory LLMs through conditioning on a few in-domain demonstrations. In this work, we first show that *Vanilla ICL*, improves zero-shot performance across diverse speech and audio tasks for selected models which suggest this ICL adaptation capability can be generalized to multimodal setting. Building on this, we propose **Speech In-Context Learning Adaptation Training (SICL-AT)**, a post-training recipe utilizes only high resource speech data intending to strengthen model's in-context learning capability. The enhancement can generalize to audio understanding/reasoning task. Experiments indicate our proposed method consistently outperforms direct fine-tuning in low-resource scenario.

1 Introduction

In-domain data is difficult to collect at scale for low-resource settings such as child's speech and audio reasoning tasks, where labels often require careful curation and expert review. As a result, target-domain supervision is typically limited and may be under-representative of the true test distribution, making direct fine-tuning brittle and sometimes harmful under domain shift. However, large-scale data from out-of-domain sources is often readily available. In the speech domain, for instance, adult English ASR data is abundant and easy to collect, whereas ASR data for child's speech remains scarce. This contrast raises an important question:

Can low-resource tasks benefit from high-resource but out-of-domain data?

In-Context Learning (ICL) (Brown et al., 2020) offers a promising paradigm for leveraging large-scale out-of-domain data. Through ICL, LLMs can be adapted to new tasks by conditioning on a small set of labeled in-domain examples, without requiring gradient updates. This approach has been shown to be effective across a wide range of modalities (Huang et al., 2023; Kong et al., 2024; Xiaomi, 2025). Within the speech domain specifically, ICL has demonstrated notable gains on various tasks including automatic speech recognition (ASR) for children's speech and unseen languages or dialects (Wang et al., 2024b,a; Zhou et al., 2025; Zheng et al., 2025b,a), speech translation (ST) (Pan et al., 2023; Chen et al., 2024), and speech emotion recognition (Yang et al., 2024; Ihori et al., 2025). Fundamentally, ICL enables models to explicitly exploit contextual information to guide generation. This enables us to improve performance on low-resource tasks using only a small number of in-domain examples. In this paper, we first demonstrate that vanilla ICL consistently improves performance across diverse speech and audio tasks.

Moreover, although high-resource out-of-domain data often suffers from domain mismatch, is it possible to perform ICL-style fine-tuning on such data to teach models *how* to utilize contextual cues, rather than merely memorizing domain-specific knowledge? To this end, we propose **Speech In-Context Learning Adaptation-Tuning (SICL-AT)**, a post-training strategy that explicitly trains models to perform inference conditioned on audio demonstrations, thereby strengthening the models' abilities to utilize contextual information through ICL. Notably, by applying SICL-AT using only high resource speech data, we achieve consistent performance gains across low-resource ASR and AU/AR on two different model backbones. Furthermore, we run a case study to show our proposed method is more stable than directly finetuning in low resource scenario.

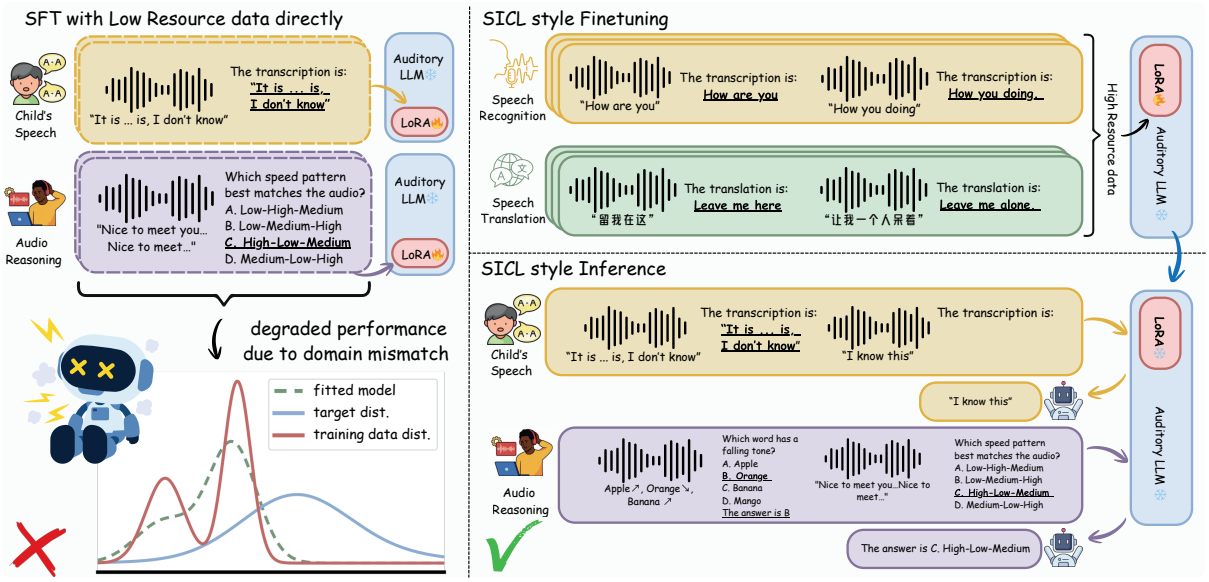

Figure 1: Motivation and overview of SICL-AT for low-resource audio tasks. Left: Direct supervised fine-tuning (SFT) on limited in-domain data often degrades performance under distribution shift due to domain mismatch between scarce training samples and the target test distribution. Right: We instead perform SICL-style fine-tuning on abundant, high-resource speech tasks using an ICL formatted objective, then apply SICL-style inference by providing task demonstrations at test time to adapt the same LoRA-augmented auditory LLM to low-resource domains, improving robustness and downstream performance.

2 Methodology

We propose **Speech In-Context Learning Adaptation-Tuning (SICL-AT)**, a post-training recipe that explicitly teaches an auditory LLM to inference conditioned on a small set of in-context audio demonstrations. Figure 1 illustrates the motivation and overall framework. Algorithm 1 summarizes the training procedure, and Table 1 lists the data used in each stage.

2.1 Training Instance

SICL-AT uses an episodic training format that mirrors inference-time in-context learning. For each task c, a *query set* $\mathcal{D}^{(c)}_{\text{query}}$ and a *demonstration pool* $\mathcal{D}^{(c)}_{\text{pool}}$ are maintained. A query instance $(x_{\text{query}}, y_{\text{query}}) \sim \mathcal{D}^{(c)}_{\text{query}}$ and retrieve k in-context demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$ are sampled from $\mathcal{D}^{(c)}_{\text{pool}}$ at each step. The prompt is then constructed using the concatenation of the k demonstrations followed by the query. Conditioned on this full context, the model generates the response y_{query} according to $P\big(y_{\text{query}} \mid x_1, y_1, \ldots, x_k, y_k, x_{\text{query}}\big)$.

2.2 Training Data

We include three types of training data: Speech Recognition (ASR), Speech Translation (ST), and Speech Question Answering (SQA) (Table 1).

Cfg.	Dataset	Task / pair	#samples
SICL-AT1	Common Voice	ASR (en)	16,368
SICL-AT2	CoVoST2	+ST (total)	37,087
		en→zh	15,427
		de→en	13,500
		zh→en	4,842
		pt→en	3,318
SICL-AT3*	MMSU	+SQA	5,000

Table 1: Training data breakdown.

We utilized the English subset of CommonVoice (Ardila et al., 2020) and the en→zh, de→en, zh→en, pt→en subsets of CoVoST2 (Wang et al., 2021) for the ASR and ST tasks, respectively. For both tasks, demonstrations are retrieved from the training split using TICL (Zheng et al., 2025b). The data is then organized into three configurations. In SICL-AT1, only ASR data is included while ST sets are further involved in SICL-AT2. Additionally, SICL-AT3 includes MMSU dataset (Wang et al., 2025) as SQA task for ablation study. Since MMSU lacks official splits and is relatively small, only the query instance is excluded from the demonstration pool in a "leave-one-out" manner.

2.3 Models

SICL-AT is applied to both *Qwen2.5-Omni* (Jin et al., 2025) and *MiMo-Audio* (Xiaomi, 2025).

Only LoRA adapters with rank of 8 and alpha of 32 are updated to avoid overfitting.

2.4 Evaluation Data & Metrics

Both Child's ASR and Audio Understanding/Reasoning tasks are utilized to evaluate the ICL capability under low resource scenarios. Non-overlap tasks spanning multilingual ASR and speech translation are further incorporated to validate the generalization of ICL.

Child's ASR. Performance of child ASR is evaluated on two corpora with distinct distributions: My Science Tutor (MyST) (Pradhan et al., 2024) and Redmond Sentence Recall (RSR) (Redmond et al., 2019). MyST contains conversational speech from Grade 3–5 students, whereas RSR comprises scripted speech from children aged 5–9. Following (Zheng et al., 2025c), the utterance-level word error rate (WER) is computed, capped at 1, and then averaged across utterances to mitigate the impact of severe hallucinations on the aggregate metric.

Audio Understanding/Reasoning. Performance of general audio understanding and reasoning is evaluated on MMAU and MMAR, which encompass speech, ambient sound, and music comprehension tasks. MMAU includes a diverse set of tasks covering 27 skills, with a focus on perception and domain-specific reasoning. MMAR comprises 16 subcategory tasks spanning speaker, environment, and content reasoning; audio quality and difference assessment; music and aesthetics; anomaly, spatial, and temporal analysis; and general reasoning. For both datasets, accuracy is reported on the public test split using the official evaluation scripts.

Multilingual ASR & Speech Translation. To verify that the improvements generalize beyond the training tasks, multilingual ASR and ST are additionally evaluated on CommonVoice (Ardila et al., 2020) and CoVoST2 (Wang et al., 2021), respectively. Evaluation is conducted on unseen language pairs and languages (de, fr, zh, en→ja, and ja→en). Word error rate (WER) is reported for alphabetic languages, character error rate (CER) for Chinese, and BLEU with up to 4-gram precision for ST performance.

3 Experiments

3.1 Vanilla In-Context Learning

Across both models, adding retrieved demonstrations at inference time (*Vanilla SICL*) improves performance on most benchmarks, indicating that auditory LLMs can leverage in-context examples as a lightweight test-time adaptation signal. Gains are especially consistent on child's ASR and AU/AR, with only a small number of exceptions.

3.2 Effect of SICL Fine-tuning

SICL-AT1 (ASR-only). SICL-AT1 yields the strongest improvements on child's ASR, supporting the idea that SICL-style post-training enhances the model's ability to leverage in-domain demonstrations for inference-time adaptation. The fact that performance also improves on multilingual ASR suggests the benefit extends beyond English recognition. However, limited gains on speech translation and inconsistent improvement on AU/AR indicate the enhanced adaptation capability is mainly concentrated in ASR.

SICL-AT2 (ASR + ST). After incorporating more training data from speech translation(SICL-AT2), model's ICL adaptation ability for ST increase as expected according to the rise of BLEU score on non-overlap ST evaluation. We notice AU/AR performance further increases for both models which is interesting because neither ASR nor ST task are overlap task as AU/AR, but we manage to enhance model's ICL adaptation capability on those tasks by training on those relatively high resource data. Overall, these results suggest that strengthening a model's ICL adaptation does not necessarily require large-scale data from the exact same downstream task.

SICL-AT3 (ASR + ST + SQA). We further add SQA to the post-training data (SICL-AT3), whose prompt–answer structure more closely matches AU/AR. This yields additional gains on AU/AR, but comes with slight degradations on ASR/ST. Together, these results offer a practical hint for SICL-AT: post-training is most effective when the training tasks resemble the intended downstream tasks in supervision and prompt–answer format. This observation motivates more principled mixture design when targeting specific capabilities.

4 Compared to Direct Fine-tuning

To highlight the advantage of our approach in low-resource settings, we run a small direct fine-tuning baseline on a representative target task: child's ASR on RSR. Concretely, we fine-tune Qwen2.5-Omni on the official RSR training split using supervised training, while keeping the setup comparable

Table 2: Summary of experiments. The de, zh and fr subset of CommonVoice is used for evaluating Multilingual ASR. Speech Translation (ST) chooses a corresponding subset from CoVoST2. Detailed breakdown of MMAU and MMAR is in the Appendix B.

		Child's ASR		AU/AR		Multilingual ASR			ST	
	Fewshot	↓WER		↑Acc.		↓WER			↑BLEU	
Tasks		MyST	RSR	MMAU	MMAR	de	zh	fr	en→ja	ja→en
MiMo-Audio	✗	14.25	31.39	66.90%	54.70%	69.74	14.43	90.68	5.25	4.56
+Vanilla SICL	✓	11.55	16.84	72.60%	58.20%	37.46	11.05	45.88	26.56	13.95
+SICL-AT1	✓	**11.49**	**16.59**	71.90%	57.70%	34.11	6.59	45.50	1.40	12.43
+SICL-AT2	✓	11.51	16.89	72.90%	61.00%	**30.49**	**6.51**	**39.47**	**36.92**	**16.76**
+SICL-AT3	✓	**11.49**	16.95	**73.40%**	**61.40%**	31.22	6.62	40.46	36.84	15.24
Qwen2.5-Omni	✗	23.05	35.65	65.80%	49.20%	5.75	5.15	7.80	41.70	22.84
+Vanilla SICL	✓	22.72	27.86	67.30%	53.80%	5.07	5.39	7.50	42.58	24.27
+SICL-AT1	✓	**14.76**	**20.96**	69.60%	54.20%	**4.42**	5.12	**6.39**	43.16	23.58
+SICL-AT2	✓	17.03	21.95	71.10%	54.40%	4.75	**4.71**	6.51	**47.57**	**26.46**
+SICL-AT3	✓	17.42	22.16	**72.10%**	**54.50%**	4.58	4.73	6.57	47.19	26.34
Fine-tuned on CV-en	✗	19.83	30.61	63.90%	50.50%	8.64	7.37	10.58	31.80	11.07
+Vanilla SICL	✓	18.05	23.62	64.10%	50.40%	8.16	8.29	10.53	33.28	18.05
Fine-tuned on RSR	✗	29.47	31.09	65.30%	44.90%	8.43	7.84	11.42	43.72	16.27

to SICL-AT by updating only LoRA adapters. On the RSR test split, where the available training data is scarce and likely under representative, direct fine-tuning improves over the zero-shot baseline. However, it does not surpass Vanilla SICL and remains clearly worse than SICL-AT. More importantly, despite both datasets contain children's speech, the degradation on the MyST test split further confirm the harmfulness of distribution mismatch between the fine-tuning and evaluation data.

We further examine whether fine-tuning on high-resource data can help low-resource data for the same task. Specifically, we fine-tune Qwen2.5-Omni on the Common Voice English subset. Strengthening general English ASR yields consistent gains on children's ASR, improving over the original model in both zero-shot and few-shot settings. Nevertheless, SICL-AT still delivers stronger adaptation ability, underscoring the necessity of explicitly training for ICL behavior rather than relying on supervised fine-tuning alone.

Overall, this case study suggests that in low resource scenario directly fine-tuning on narrowly matched (or domain-shifted) data can overspecialize and hurt generalization. In contrast, leveraging limited in-domain data as demonstrations enables more robust adaptation at inference time. When high-resource data is available, SICL-AT remains a more reliable strategy than direct fine-tuning, as it more effectively strengthens gradient-free, demonstration-conditioned adaptation that transfers to low-resource scenarios.

5 Conclusion

This work studies speech in-context learning (SICL) as an inference-time adaptation mechanism for large auditory LLMs which can be applied to a broad range of speech and audio tasks by simply conditioning on a small set of audio demonstrations (Vanilla SICL), including child's ASR, multilingual ASR, speech translation, and general audio understanding/reasoning. Building on this, we propose Speech In-Context Learning Adaptation-Training (SICL-AT), a post-training recipe that explicitly trains models in the same demonstration-conditioned format used at inference. Across two model families, SICL-AT strengthens and stabilizes in-context learning behavior, and improvements transfer beyond the training skill types. Our ablations suggest that aligning post-training episodes with downstream task format can further boost targeted capabilities. Finally, we use a case study to show that our proposed method is preferable when in-domain data is limited but high resource data is largely available.

Limitations

Our experiments cover two model families and a fixed set of benchmarks and retrieval choices, and we do not fully characterize the inference-cost

scaling with longer contexts or perform extensive qualitative failure analysis. Researchers should also notice SICL performance depends on retrieval quality and the availability of representative examples, which may be limited in truly data-scarce deployments.

Ethical Statement

This work studies *speech in-context learning* and post-training strategies for large multimodal speech models, with experiments on automatic speech recognition (including child speech), multilingual ASR, speech translation, and audio understanding/reasoning.

Data use and privacy. All experiments use existing datasets released by their respective creators under their original licenses and access conditions. Several benchmarks include recordings of minors (child speech). We do not collect new human-subject data, and we rely on the dataset providers' consent procedures and de-identification/anonymization practices. In our processing and evaluation, we treat audio as sensitive data: we do not attempt speaker identification, attribute inference, or any linkage to real identities, and we report only aggregate metrics. If we release code, we will avoid distributing any audio or metadata that could re-identify participants.

Potential risks and mitigations. Improved speech recognition and understanding can be beneficial (e.g., accessibility and education), but also carries risks, including privacy-invasive surveillance, profiling, or harmful deployment in high-stakes settings. In addition, ASR errors are not uniformly distributed across speakers; child speech, accented speech, and low-resource languages are historically more error-prone. To mitigate these concerns, we (i) explicitly evaluate on diverse settings (children speech and multilingual/translation tasks) to surface performance gaps, (ii) emphasize that reported improvements do not imply suitability for safety-critical or rights-impacting uses, and (iii) recommend that any deployment include informed consent, security controls, and continuous monitoring for differential error rates across groups.

Misuse considerations. Our methods could be used to adapt general models to new domains with limited data, which might lower the barrier for misuse. We therefore focus on research settings, report limitations under domain shift, and encourage responsible release practices (e.g., documentation of training data, intended use, and known failure modes). We do not claim that the models provide clinical or diagnostic judgments, and our results should not be used as a substitute for professional assessment.

Environmental impact. Training and evaluation require non-trivial compute. We reduce cost by using parameter-efficient adaptation (LoRA) and by keeping most experiments comparable and bounded in scale. Where possible, we will report key training configurations to support reproducibility and to help others estimate compute needs.

AI usage statement. Generative AI tools (large language models) were used in a limited way to assist with writing and editing (e.g., improving clarity, grammar, and LaTeX formatting). All technical content—including experimental design, implementation, results, and claims—was produced and verified by the authors, who take full responsibility for the paper. No private, restricted, or unpublished dataset content was provided to these tools beyond text intended for the manuscript, and the tools were not used to generate or manipulate evaluation data, labels, or reported metrics.

Software and packages. Our experiments were implemented using standard open-source toolchains (e.g., PyTorch, Hugging Face Transformers/Datasets, and common evaluation libraries for ASR/ST), along with publicly available scripts provided by dataset/benchmark authors when applicable.

Descriptive statistics. We report corpus-level WER/BLEU/accuracy on the full evaluation sets. Unless otherwise noted, each result corresponds to a single evaluation run of a fixed checkpoint with fixed decoding/retrieval settings (we do not report mean/std over multiple random seeds).

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 4211–4215.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation. In *ICASSP*, pages 13521–13525. IEEE.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language Is Not All You Need: Aligning Perception with Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 72096–72109.

Mana Ihori, Taiga Yamane, Naotaka Kawata, Naoki Makishima, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, and Ryo Masumura. 2025. Few-shot Personalization via In-Context Learning for Speech Emotion Recognition based on Speech-Language Model. In *ASRU*, pages 1–6.

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

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 25125–25148.

Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248*.

Sameer Pradhan, Ronald Cole, and Wayne Ward. 2024. My Science Tutor (MyST) – A Large Corpus of Children's Conversational Speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12040–12045.

Sean M Redmond, Andrea C Ash, Tyler T Christopulos, and Theresa Pfaff. 2019. Diagnostic Accuracy of Sentence Recall and Past Tense Measures for Identifying Children's Language Impairments. *Journal of Speech, Language, and Hearing Research*, 62(7):2438–2454.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *INTERSPEECH*, pages 2247–2251.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. *arXiv preprint arXiv:2506.04779*.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024a. Bayesian example selection improves in-context learning for speech, text and visual modalities. In *EMNLP*, pages 20812–20828.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024b. Can Whisper Perform Speech-Based In-Context Learning? In *ICASSP*, pages 13421–13425. IEEE.

LLM-Core-Team Xiaomi. 2025. MiMo-Audio: Audio Language Models are Few-Shot Learners.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *NeurIPS*, pages 56802–56827.

Haolong Zheng, Yekaterina Yegorova, and Mark Hasegawa-Johnson. 2025a. TICL+: A Case Study On Speech In-Context Learning for Children's Speech Recognition. *Preprint*, arXiv:2512.18263.

Haolong Zheng, Yekaterina Yegorova, and Mark Hasegawa-Johnson. 2025b. TICL: Text-Embedding KNN For Speech In-Context Learning Unlocks Speech Recognition Abilities of Large Multimodal Models. *Preprint*, arXiv:2509.13395.

Xiuwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu J. Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrik Kuila, Colin Lea, Bob MacDonald, Gautam Mantena, Venkatesh Ravichandran, Leda Sari, Katrin Tomanek, Chang D. Yoo, and Chris Zwilling. 2025c. The Interspeech 2025 Speech Accessibility Project Challenge. In *INTERSPEECH*, pages 3269–3273.

Jiaming Zhou, Shiwan Zhao, Jiabei He, Hui Wang, Wenjia Zeng, Yong Chen, Haoqin Sun, Aobo Kong, and Yong Qin. 2025. M2R-Whisper: Multi-stage and Multi-scale Retrieval Augmentation for Enhancing Whisper. In *ICASSP*, pages 1–5. IEEE.

A SICL-AT Algorithm

Algorithm 1: SICL-AT

Input: C training tasks,
each has query set
$\mathcal{D}_{query}^{(c)} = \{(x_c, y_c)^{(j)}\}_{j=1}^{N_{c,query}}$, and demo
pool $\mathcal{D}_{pool}^{(c)} = \{(x_c, y_c)^{(j)}\}_{j=1}^{N_{c,pool}}$

for $step \leftarrow 1$ **to** *Total_Step* **do**

- **(1)** Sample task index $c \sim \{1, \ldots, C\}$
- **(2)** Sample $(x_{\text{query}}, y_{\text{query}}) \sim \mathcal{D}_{\text{query}}^{(c)}$
- **(3)** Retrieve $\{(x_i, y_i)\}_{i=1}^{k}$ from $\mathcal{D}_{\text{pool}}^{(c)}$
- **(4)** Update parameters to maximize $P\big(y_{\text{query}} \mid x_1, y_1, \ldots, x_k, y_k, x_{\text{query}}\big)$

B General Audio Understanding and Reasoning Performance Breakdowns

Table 3: MMAU accuracy breakdown by group/item for **MiMo-Audio**.

Group	**Item**	**n**	**0shot**	**ICL**	**SICL-AT1**	**SICL-AT2**	**SICL-AT3**
Task	sound	333	71.77%	75.98%	78.98%	75.98%	78.08%
Task	music	334	65.27%	66.77%	65.57%	68.86%	68.26%
Task	speech	333	63.66%	75.08%	71.17%	73.87%	73.87%
Difficulty	easy	224	55.80%	66.52%	59.82%	64.29%	65.18%
Difficulty	hard	236	64.41%	71.19%	72.88%	71.61%	69.92%
Difficulty	medium	540	72.59%	75.74%	76.48%	77.04%	78.33%
Sub-category	Acoustic Source Inference	48	72.92%	81.25%	85.42%	70.83%	79.17%
Sub-category	Temporal Event Reasoning	48	66.67%	62.50%	75.00%	62.50%	66.67%
Sub-category	Dissonant Emotion Interpretation	35	80.00%	85.71%	74.29%	82.86%	80.00%
Sub-category	Event-Based Knowledge Retrieval	33	75.76%	90.91%	81.82%	93.94%	87.88%
Sub-category	Counting	29	48.28%	55.17%	55.17%	62.07%	55.17%
Sub-category	Phonemic Stress Pattern Analysis	53	39.62%	62.26%	52.83%	50.94%	60.38%
Sub-category	Emotion State summarisation	44	56.82%	56.82%	61.36%	61.36%	63.64%
Sub-category	Conversational Fact Retrieval	22	86.36%	95.45%	95.45%	90.91%	100.00%
Sub-category	Key highlight Extraction	21	80.95%	90.48%	95.24%	90.48%	90.48%
Sub-category	Multi Speaker Role Mapping	27	100.00%	100.00%	100.00%	100.00%	100.00%
Sub-category	Phonological Sequence Decoding	49	59.18%	81.63%	69.39%	75.51%	71.43%
Sub-category	Emotion Flip Detection	20	35.00%	45.00%	55.00%	55.00%	50.00%
Sub-category	Instrumentation	35	60.00%	71.43%	68.57%	77.14%	68.57%
Sub-category	Temporal Reasoning	56	41.07%	37.50%	39.29%	41.07%	39.29%
Sub-category	Lyrical Reasoning	10	90.00%	90.00%	90.00%	90.00%	90.00%
Sub-category	Socio-cultural Interpretation	20	65.00%	75.00%	70.00%	70.00%	80.00%
Sub-category	Rhythm and Tempo Understanding	46	69.57%	63.04%	60.87%	65.22%	71.74%
Sub-category	Musical Texture Interpretation	34	73.53%	76.47%	70.59%	76.47%	76.47%
Sub-category	Melodic Structure Interpretation	33	66.67%	66.67%	54.55%	63.64%	54.55%
Sub-category	Harmony and Chord Progressions	33	63.64%	63.64%	66.67%	69.70%	63.64%
Sub-category	Musical Genre Reasoning	34	70.59%	79.41%	88.24%	85.29%	88.24%
Sub-category	Event-Based Sound Reasoning	48	79.17%	83.33%	81.25%	89.58%	81.25%
Sub-category	Emotional Tone Interpretation	33	84.85%	84.85%	84.85%	84.85%	87.88%
Sub-category	Eco-Acoustic Knowledge	47	68.09%	72.34%	76.60%	82.98%	78.72%
Sub-category	Ambient Sound Interpretation	48	62.50%	72.92%	77.08%	72.92%	75.00%
Sub-category	Acoustic Scene Reasoning	48	70.83%	70.83%	64.58%	64.58%	75.00%
Sub-category	Sound-Based Event Recognition	46	82.61%	89.13%	93.48%	89.13%	91.30%
Overall	Total Accuracy	1000	66.90%	72.60%	71.90%	72.90%	73.40%

Table 4: MMAU accuracy breakdown by group/item for **Qwen2.5-Omni**.

Group	Item	n	0shot	ICL	SICL-AT1	SICL-AT2	SICL-AT3
Task	sound	333	68.17%	73.87%	77.18%	75.08%	78.08%
Task	music	334	62.28%	62.28%	61.98%	65.27%	66.77%
Task	speech	333	66.97%	65.77%	69.67%	72.97%	71.47%
Difficulty	easy	224	63.84%	61.16%	63.39%	68.30%	68.30%
Difficulty	hard	236	61.44%	60.17%	64.83%	63.56%	69.07%
Difficulty	medium	540	68.52%	72.96%	74.26%	75.56%	75.00%
Sub-category	Acoustic Source Inference	48	81.25%	81.25%	93.75%	89.58%	87.50%
Sub-category	Temporal Event Reasoning	48	41.67%	64.58%	52.08%	58.33%	60.42%
Sub-category	Dissonant Emotion Interpretation	35	71.43%	88.57%	82.86%	94.29%	82.86%
Sub-category	Event-Based Knowledge Retrieval	33	75.76%	75.76%	78.79%	87.88%	78.79%
Sub-category	Counting	29	68.97%	58.62%	58.62%	62.07%	55.17%
Sub-category	Phonemic Stress Pattern Analysis	53	43.40%	41.51%	43.40%	49.06%	47.17%
Sub-category	Emotion State summarisation	44	50.00%	45.45%	56.82%	45.45%	54.55%
Sub-category	Conversational Fact Retrieval	22	95.45%	86.36%	90.91%	95.45%	90.91%
Sub-category	Key highlight Extraction	21	76.19%	85.71%	80.95%	85.71%	85.71%
Sub-category	Multi Speaker Role Mapping	27	100.00%	100.00%	100.00%	100.00%	100.00%
Sub-category	Phonological Sequence Decoding	49	79.59%	77.55%	89.80%	91.84%	89.80%
Sub-category	Emotion Flip Detection	20	25.00%	10.00%	20.00%	30.00%	45.00%
Sub-category	Instrumentation	35	74.29%	62.86%	74.29%	77.14%	77.14%
Sub-category	Temporal Reasoning	56	42.86%	35.71%	35.71%	41.07%	37.50%
Sub-category	Lyrical Reasoning	10	60.00%	90.00%	90.00%	90.00%	100.00%
Sub-category	Socio-cultural Interpretation	20	70.00%	65.00%	70.00%	65.00%	70.00%
Sub-category	Rhythm and Tempo Understanding	46	50.00%	54.35%	56.52%	58.70%	45.65%
Sub-category	Musical Texture Interpretation	34	70.59%	76.47%	64.71%	73.53%	67.65%
Sub-category	Melodic Structure Interpretation	33	57.58%	63.64%	57.58%	57.58%	72.73%
Sub-category	Harmony and Chord Progressions	33	51.52%	60.61%	63.64%	63.64%	69.70%
Sub-category	Musical Genre Reasoning	34	88.24%	82.35%	70.59%	79.41%	88.24%
Sub-category	Event-Based Sound Reasoning	48	66.67%	77.08%	83.33%	79.17%	81.25%
Sub-category	Emotional Tone Interpretation	33	75.76%	72.73%	78.79%	81.82%	90.91%
Sub-category	Eco-Acoustic Knowledge	47	74.47%	78.72%	80.85%	85.11%	82.98%
Sub-category	Ambient Sound Interpretation	48	77.08%	70.83%	81.25%	68.75%	79.17%
Sub-category	Acoustic Scene Reasoning	48	58.33%	62.50%	66.67%	64.58%	75.00%
Sub-category	Sound-Based Event Recognition	46	78.26%	82.61%	82.61%	80.43%	80.43%
Overall	Total Accuracy	1000	65.80%	67.30%	69.60%	71.10%	72.10%

Table 5: MMAR Accuracy breakdown for **MiMo-Audio**.

Group	Item	n	0shot	Vanilla SICL	SICL-AT1	SICL-AT2	SICL-AT3
Modality	sound	165	53.33%	52.73%	55.15%	60.61%	62.42%
Modality	music	206	39.32%	43.20%	44.17%	46.60%	46.12%
Modality	speech	294	58.84%	63.61%	63.61%	66.33%	66.67%
Modality	mix-sound-music	11	45.45%	27.27%	27.27%	27.27%	45.45%
Modality	mix-sound-speech	218	63.30%	68.35%	65.60%	69.27%	67.43%
Modality	mix-music-speech	82	58.54%	57.32%	56.10%	58.54%	60.98%
Modality	mix-sound-music-speech	24	58.33%	83.33%	66.67%	70.83%	75.00%
Category	Signal Layer	43	53.49%	55.81%	48.84%	62.79%	60.47%
Category	Perception Layer	404	52.72%	53.71%	56.19%	57.92%	58.42%
Category	Semantic Layer	412	58.98%	65.05%	62.38%	66.50%	66.26%
Category	Cultural Layer	141	48.23%	51.77%	51.06%	53.19%	56.03%
Sub-category	Speaker Analysis	48	62.50%	62.50%	54.17%	64.58%	60.42%
Sub-category	Environmental Perception and Reasoning	149	59.06%	61.07%	66.44%	71.14%	73.15%
Sub-category	Content Analysis	304	60.20%	67.76%	64.47%	67.43%	68.42%
Sub-category	Correlation Analysis	50	62.00%	58.00%	60.00%	58.00%	62.00%
Sub-category	Counting and Statistics	99	42.42%	44.44%	40.40%	39.39%	39.39%
Sub-category	Professional Knowledge and Reasoning	71	47.89%	54.93%	53.52%	52.11%	56.34%
Sub-category	Culture of Speaker	52	50.00%	46.15%	48.08%	57.69%	59.62%
Sub-category	Aesthetic Evaluation	8	37.50%	62.50%	37.50%	50.00%	50.00%
Sub-category	Emotion and Intention	60	50.00%	53.33%	58.33%	63.33%	60.00%
Sub-category	Anomaly Detection	17	58.82%	52.94%	47.06%	64.71%	64.71%
Sub-category	Spatial Analysis	15	60.00%	53.33%	73.33%	73.33%	66.67%
Sub-category	Temporal Analysis	28	53.57%	57.14%	53.57%	57.14%	57.14%
Sub-category	Acoustic Quality Analysis	18	33.33%	44.44%	33.33%	44.44%	44.44%
Sub-category	Music Theory	63	44.44%	46.03%	50.79%	52.38%	49.21%
Sub-category	Audio Difference Analysis	8	87.50%	87.50%	87.50%	100.00%	87.50%
Sub-category	Imagination	10	50.00%	50.00%	60.00%	40.00%	40.00%
Overall	Total Accuracy	1000	54.70%	58.20%	57.70%	61.00%	61.40%

Table 6: MMAR Accuracy breakdown for **Qwen2.5-Omni**.

Group	Item	n	0shot	Vanilla SICL	SICL-AT1	SICL-AT2	SICL-AT3
Modality	sound	165	47.27%	52.12%	55.76%	59.39%	60.00%
Modality	music	206	36.89%	39.81%	42.23%	44.66%	41.75%
Modality	speech	294	51.70%	56.46%	56.12%	55.44%	57.48%
Modality	mix-sound-music	11	27.27%	63.64%	45.45%	36.36%	54.55%
Modality	mix-sound-speech	218	56.42%	58.72%	57.34%	55.96%	56.42%
Modality	mix-music-speech	82	54.88%	63.41%	63.41%	60.98%	58.54%
Modality	mix-sound-music-speech	24	62.50%	70.83%	66.67%	62.50%	58.33%
Category	Signal Layer	43	25.58%	46.51%	51.16%	51.16%	60.47%
Category	Perception Layer	404	46.29%	48.76%	52.23%	52.72%	50.50%
Category	Semantic Layer	412	55.10%	59.22%	57.04%	58.74%	60.19%
Category	Cultural Layer	141	47.52%	54.61%	52.48%	47.52%	47.52%
Sub-category	Speaker Analysis	48	56.25%	45.83%	52.08%	60.42%	64.58%
Sub-category	Environmental Perception and Reasoning	149	64.43%	64.43%	63.76%	65.77%	66.44%
Sub-category	Content Analysis	304	55.26%	62.17%	57.89%	59.87%	60.53%
Sub-category	Correlation Analysis	50	46.00%	60.00%	54.00%	56.00%	62.00%
Sub-category	Counting and Statistics	99	31.31%	32.32%	36.36%	35.35%	30.30%
Sub-category	Professional Knowledge and Reasoning	71	47.89%	56.34%	50.70%	56.34%	47.89%
Sub-category	Culture of Speaker	52	44.23%	51.92%	53.85%	36.54%	48.08%
Sub-category	Aesthetic Evaluation	8	75.00%	50.00%	37.50%	37.50%	50.00%
Sub-category	Emotion and Intention	60	53.33%	55.00%	56.67%	51.67%	55.00%
Sub-category	Anomaly Detection	17	35.29%	47.06%	47.06%	47.06%	76.47%
Sub-category	Spatial Analysis	15	53.33%	40.00%	66.67%	53.33%	53.33%
Sub-category	Temporal Analysis	28	28.57%	42.86%	50.00%	46.43%	42.86%
Sub-category	Acoustic Quality Analysis	18	11.11%	44.44%	50.00%	50.00%	55.56%
Sub-category	Music Theory	63	33.33%	33.33%	46.03%	49.21%	38.10%
Sub-category	Audio Difference Analysis	8	37.50%	50.00%	62.50%	62.50%	37.50%
Sub-category	Imagination	10	40.00%	60.00%	70.00%	50.00%	40.00%
Overall	Total Accuracy	1000	49.20%	53.80%	54.20%	54.40%	54.50%