Adapter-Only Bridging of Frozen Speech Encoder and Frozen LLM for ASR

Junseok Oh[1] and Ji-Hwan Kim[1⋆]

Department of Computer Science and Engineering, Sogang University, Seoul, Republic of Korea
{ohjs, kimjihwan}@sogang.ac.kr

Abstract. Integrating Large Language Models (LLMs) with speech encoders can improve the performance of Automatic Speech Recognition (ASR) by leveraging enhanced linguistic knowledge. Prior work faces two key limitations. Fine-tuning the LLM may cause forgetting of pre-trained domain knowledge. Methods that freeze both components but use only a single linear projection suffer catastrophic domain-transfer failures. We propose adapter-only bridging that keeps both the speech encoder and the LLM frozen while training lightweight adapters with causal convolutions for temporal modeling (24.6M parameters, 0.44% of total parameters). Our adapters comprise a convolutional downsampler and a multi-layer MLP projection. Trained on 2.9k hours of general-domain speech, our approach achieves competitive performance on LibriSpeech and significantly outperforms Whisper-large-v2 on academic domains (26.8% relative WER reduction). Unlike prior work with frozen components that fails on out-of-domain data, our causal convolutional adapters enable robust cross-domain performance while preserving the LLM's text-trained knowledge for inference-time domain prompting. On four academic domains, domain prompting achieves additional WER reductions in three domains (Engineering 2.10%, Social Sciences 1.30%, Medical 0.73%) and improves domain-specific terminology recognition by 7.2%p F1 and 5.8%p recall. Our approach enables a single frozen model to maintain general-domain accuracy while adapting across domains via prompts alone.

Keywords: Automatic Speech Recognition · Adapter-Based Speech-LLM · Parameter-Efficient Learning · Inference-Time Domain Prompting

1 Introduction

Automatic Speech Recognition (ASR) has achieved strong performance on general-domain benchmarks through end-to-end neural architectures [14, 3, 28]. However, academic domains with technical terminology and domain-specific discourse patterns remain difficult [9]. The language modeling in end-to-end ASR [28, 9] is learned from paired speech-text data. In contrast, modern Large Language Models (LLMs) are trained on massive text corpora and thus contain broader domain

[⋆] Corresponding author: kimjihwan@sogang.ac.kr knowledge [2]. This motivates integrating LLMs with speech encoders: the LLM provides linguistic knowledge for ASR, and a domain description at inference time can activate relevant knowledge for academic domains.

When integrated with speech encoders for ASR, LLMs can contribute linguistic knowledge to recognition [24, 12, 5, 16]. However, prior work faces critical limitations that create a fundamental dilemma. Methods that fine-tune the LLM risk catastrophic forgetting of pre-trained domain knowledge essential for adaptation, effectively destroying the very knowledge we seek to leverage. Conversely, approaches that freeze both components but rely on simple linear projection suffer severe domain-transfer failures due to insufficient cross-modal alignment, failing to effectively bridge the representational gap between speech and text modalities.

Therefore, we require a third approach that preserves the LLM's knowledge while achieving robust cross-modal alignment. We propose adapter-only bridging between a frozen speech encoder and a frozen LLM for ASR. We explicitly freeze both components and train only lightweight modality adapters as a bridge to connect speech and text representations. This design preserves the LLM's pre-trained domain knowledge and enables simple domain prompts for inference-time domain adaptation without any domain-specific training. With the LLM's text-trained knowledge preserved, providing a domain description improves recognition. In contrast, speech-only models [28, 9] are limited to knowledge learned from paired speech-text data.

Our approach builds upon recent work in speech–LLM integration [24, 30, 33] but differs in a key design choice: rather than adapting the LLM, we keep both the speech encoder and LLM frozen and learn only cross-modal alignment. We employ Whisper [28] as the speech encoder and Gemma [13] as the LLM, training only lightweight adapters (convolutional downsampling and MLP projection layers) on general-domain speech while keeping both components frozen.

We evaluate on four academic domains from AI Hub's international conference interpretation data[1]: Engineering Area (EA, 34.86h), Medical Area (MA, 23.24h), Natural Sciences Area (NA, 18.78h), and Social Sciences Area (SA, 15.62h). Our adapter-only architecture significantly outperforms Whisper-large-v2 across all domains with average 26.8% relative WER reduction. With only domain prompts and no domain-specific training, we achieve additional improvements over the generic prompt setting in three domains (Engineering 2.10%, Social Sciences 1.30%, Medical 0.73%).

Our main contributions are: (1) Adapter-only bridging of frozen speech encoder and LLM that preserves pre-trained knowledge while enabling inference-time domain prompting. (2) Parameter-efficient design (24.6M trainable; 0.44% of 5.6B) that significantly outperforms baselines. (3) Experimental validation on four academic domains showing 26.8% average improvement over Whisper. (4) Practical solution replacing manual per-domain vocabulary creation with simple inference-time domain prompting.

[1] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71693

2 Related Work

Speech-LLM Integration. Recent work has explored various approaches to integrate speech encoders with LLMs for ASR [8]. SLAM-ASR [24] uses simple linear projection to connect frozen HuBERT [17] encoder with Vicuna-7B [6], achieving strong performance on LibriSpeech [25] but suffering severe domain transfer issues. SALMONN [30] uses dual encoders (Whisper [28] and BEATs [4]) connected to Vicuna-13B [6] through Q-Former from BLIP-2 [21], with LoRA [18] for modality adaptation. Other approaches include Qwen-Audio [7], WavLLM [19], SALM [5], AudioChatLlama [12], and various adapter-based methods [16, 23, 32, 33].

Domain Adaptation for ASR. Domain adaptation in ASR has traditionally relied on external language models or domain-specific vocabulary expansion [9], requiring manual curation of word lists and separate models for each domain. Recent work explores contextualized ASR [26] and multi-domain training [9]. Prompt-based methods have shown success in NLP [2] but remain underexplored for speech domain adaptation.

Multimodal Alignment. Cross-modal alignment is key for speech-language integration [22]. Vision-language models like CLIP [27] and CLAP [10] demonstrate effective contrastive learning. For speech-text alignment, various techniques address the modality gap [11, 26].

Parameter-Efficient Fine-Tuning. Methods like LoRA [18] and adapter modules [16] reduce trainable parameters but still modify LLM's internal representations. Recent work shows such modifications can distort pre-trained representation space [20].

3 Method

3.1 Architecture Overview

We present our adapter-only architecture design in Figure 1(a). Unlike prior speech-LLM integration methods that either fine-tune the LLM [30] or use minimal projection layers [24], we explicitly freeze both the speech encoder and LLM (5.64B frozen vs. 24.6M trainable, 0.44% of total parameters) while training only temporal modeling adapters for robust cross-modal alignment. We integrate three core components in our architecture: (1) a frozen Whisper-large-v2 encoder [28] for audio feature extraction (636.8M parameters), (2) trainable adapters including a convolutional downsampler (14.8M parameters) and MLP projection layer (9.8M parameters) for cross-modal alignment (total 24.6M trainable), and (3) a frozen Gemma-3-4B-Instruct LLM [13][2] for text generation (5.0B parameters).

Frozen Audio Encoder. We employ Whisper-large-v2 as the audio encoder, which processes 80-dimensional log-mel spectrograms and produces representations $\mathbf{S} \in \mathbb{R}^{T_s \times 1280}$ where T_s is the temporal sequence length and 1,280

[2] `https://huggingface.co/google/gemma-3-4b-it`

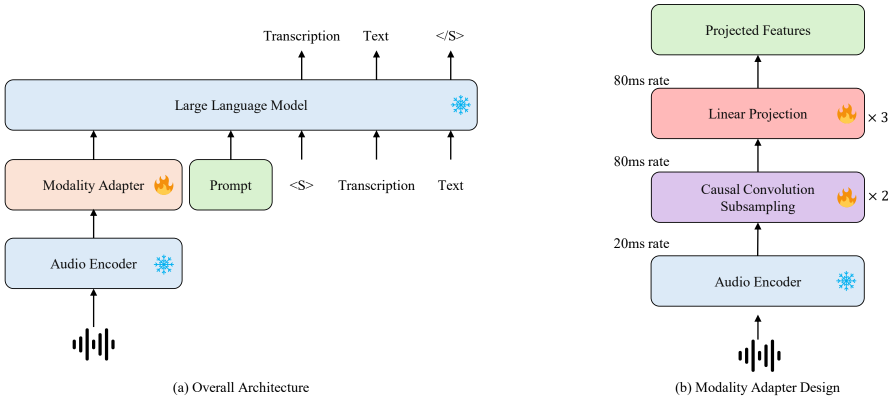

Fig. 1. Adapter-only architecture for domain-adaptive ASR. (a) Overall architecture: end-to-end pipeline with frozen encoder and LLM connected by a trainable modality adapter. (b) Modality adapter design: internal structure with causal convolutions for temporal modeling and cross-modal alignment.

is the hidden dimension. We maintain the encoder frozen during training to preserve its acoustic modeling capabilities learned from 680,000 hours of multilingual data.

Trainable Modality Adapters (24.6M parameters). As Figure 1(b) details, our modality adapter network bridges the representation gap between audio and text modalities [11, 26, 22] through two components: (1) *Causal Convolutional Downsampler* (14.8M): We employ a 2-layer causal convolutional architecture [31] with kernel size $k = 4$, stride $s = 2$ for 2× temporal reduction per layer, achieving overall 4× reduction. We design this with causal convolutions to enable future streaming inference. Each layer applies layer normalization, GELU (Gaussian Error Linear Unit) activation [15], and residual connections via adaptive pooling. (2) *Multi-Layer Projection MLP* (9.8M): A 3-layer MLP with residual connections projecting from Whisper dimension (1,280) to Gemma embedding dimension (2,560), learning to align audio and text representations.

By freezing 99.56% of parameters, we reduce memory requirements and enable training in resource-constrained environments, advancing practical deployment of speech processing.

Frozen LLM. Gemma-3-4B-Instruct (5.0B parameters) serves as the frozen language model. We use the instruction-tuned variant to leverage its conversational capabilities for prompt-based domain adaptation.

3.2 Training Strategy

The model is trained on combined general-domain English speech datasets: LibriSpeech (960h, 281,241 samples), Common Voice 17 (1,765.88h, 1,117,563 samples), and TED-LIUM Release 2 (208.48h, 93,564 samples), totaling approximately 2,934 hours across 1.5M utterances. The instruction-tuning format is used with Gemma's chat template. For each training sample with audio $\mathbf{X}$ and transcription text $\mathbf{Y} = [y_1, \ldots, y_N]$, audio embeddings $\mathbf{E}_{\text{audio}}$ are inserted immediately after the context marker, followed by the instruction and transcription tokens. The cross-entropy loss is calculated only over the model's response tokens (transcription $\mathbf{Y}$), with instruction tokens masked to focus learning on transcription generation. AdamW optimizer is used with learning rate 5×10^{-4}, weight decay 0.01, gradient clipping at norm 1.0, cosine learning rate schedule with 2,000 warmup steps over 5 epochs, batch size 8 per GPU with gradient accumulation over 2 steps, and mixed precision (bfloat16) training across 6 NVIDIA RTX A6000 GPUs.

3.3 Training Objective

Let $\mathbf{X} \in \mathbb{R}^{F\times T}$ denote the 80-dimensional log-mel spectrogram for an utterance (F=80, T frames). The frozen Whisper encoder produces acoustic representations $\mathbf{S} = E_{\text{enc}}(\mathbf{X}) \in \mathbb{R}^{T_s\times d_s}$ with d_s=1,280. The trainable causal convolutional downsampler D reduces the temporal resolution to $\mathbf{A} = D(\mathbf{S}) \in \mathbb{R}^{T_a\times d_s}$ (overall 4× reduction), and the trainable projection MLP M maps to the LLM embedding dimension $\mathbf{Z} = M(\mathbf{A}) \in \mathbb{R}^{T_a\times d_\ell}$, where d_ℓ is the Gemma embedding size.

During training, inputs are formatted using the instruction-tuning chat template. Following Gemma's format, the full sequence includes special tokens, audio embeddings $\mathbf{Z}$, instruction tokens $\mathbf{c} = (c_1, \ldots, c_{N_c})$ (e.g., "Transcribe this audio:"), and transcription tokens $\mathbf{y} = (y_1, \ldots, y_N)$ in the model response. The LLM (frozen) performs autoregressive next-token prediction over this sequence. Loss is computed only on the transcription tokens $\mathbf{y}$ by masking all other positions. Only the adapter parameters $\theta = \{\text{downsampler}, \text{projection}\}$ are optimized via masked cross-entropy:

$$\mathcal{L}(\theta) \;=\; \frac{1}{N}\sum_{t=1}^{N}\Big[-\log p_\theta\big(y_t \;\big|\; \mathbf{y}_{<t},\, \mathbf{Z},\, \mathbf{c}\big)\Big],$$

where the loss is calculated only over the N transcription tokens. The label -100 is assigned to all audio embeddings, instruction tokens, and padding to exclude them from loss computation. The Whisper encoder and Gemma LLM are kept frozen; all gradients flow only through D and M.

3.4 Inference-Time Domain Prompting

At inference time, we facilitate inference-time domain prompting via natural language prompts without parameter updates or retraining. We use a fixed template across domains: "This audio is from a [DOMAIN] conference. Transcribe this audio accurately, including all technical terms." For medical conferences, the [DOMAIN] slot is set to "medical" and "technical terms" is extended to "technical and medical terms." The frozen LLM processes this description with audio embeddings, activating its pre-trained domain vocabulary [2]. This reduces workload by replacing manual term identification and retraining with simple text prompts.

4 Experiments

4.1 Datasets

Experimental Design. To evaluate zero-shot domain adaptation, the model is trained on general-domain speech and evaluated on academic domains without domain-specific training. This tests whether inference-time prompting can activate relevant knowledge in the frozen LLM.

Training Data. Following this design, the model is trained on general-domain English speech: LibriSpeech (960h, 281,241 samples) [25], Common Voice 17 (1,765.88h, 1,117,563 samples) [1], and TED-LIUM Release 2 (208.48h, 93,564 samples) [29], totaling approximately 2,934 hours across 1.5M utterances.

Evaluation Data. The model is evaluated on academic conference speech from AI Hub[3]. Table 1 shows composition across four domains. These test sets contain highly domain-specific vocabulary and terminology not seen during training, making them an ideal out-of-distribution (OOD) testbed for evaluating zero-shot domain adaptation.

Table 1. Evaluation dataset composition by domain. All data is from English speech test sets of AI Hub international academic conference data.

Domain	Samples	Duration (h)	Total Words
Engineering Area (EA)	17,245	34.86	316,408
Medical Area (MA)	10,417	23.24	182,565
Natural Sciences Area (NA)	8,061	18.78	147,908
Social Sciences Area (SA)	6,860	15.62	129,567
Total	**42,583**	**92.50**	**776,448**

Comparison Methods. The following methods are compared: (1) **Whisper-large-v2** baseline, (2) **Generic Prompt (GP)**: the proposed architecture with generic prompt, and (3) **Domain Prompt (DP)**: the proposed method with domain-specific prompts.

Implementation and Reproducibility. We implement the model using PyTorch 2.4.0, PyTorch Lightning 2.5.0, and Hugging Face Transformers 4.52.4 (CUDA 12.1, bfloat16 training). We fix the random seed to 42 (via `pl.seed_everything`). Due to the high computational cost of full training on 2.9k hours, each setting is trained once; the reported scores correspond to that run. To facilitate reproducibility, we provide key implementation details in this paper, including training hyperparameters and prompt templates.

[3] `https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71693`

5 Results

5.1 General-Domain Performance

Table 2 shows performance on general-domain benchmarks. The proposed model achieves similar performance to Whisper-large-v2 on LibriSpeech (clean: 2.44% vs. 2.70%, other: 5.28% vs. 5.20%) and Common Voice 17 (14.94% vs. 14.19%), while significantly outperforming on TEDLIUM2 (5.26% vs. 12.24%), showing successful cross-modal alignment without reducing acoustic modeling performance.

Compared to other speech-LLM methods, SLAM-ASR achieves strong LibriSpeech results but suffers on conversational datasets (CV17: 60.36%, TEDLIUM2: 13.61%). SALMONN shows competitive performance on LibriSpeech and TEDLIUM2 but degrades on CV17 (34.29%). The proposed adapter-only design achieves robust performance across all datasets while requiring only 24.6M trainable parameters (0.44% of total parameters).

Table 2. Word Error Rate (%) on general-domain test sets. LS: LibriSpeech, CV17: Common Voice 17. Lower is better. Best results in each column are shown in **bold**. For TEDLIUM2, Whisper-large-v2 exhibits excessive insertion errors; the value in parentheses shows Whisper-medium performance.

Method	LS-clean	LS-other	TEDLIUM2	CV17
Whisper-large-v2 [28]	2.70	5.20	12.24 (7.7)	**14.19**
SLAM-ASR [24]	**1.90**	**3.60**	13.61	60.36
SALMONN [30]	2.20	5.70	6.47	34.29
Ours (Generic Prompt)	2.44	5.28	**5.26**	14.94

5.2 Domain Prompting Results

Table 3 presents WER results across four academic domains. The proposed adapter-only architecture without domain prompts already significantly outperforms Whisper-large-v2 across all domains, achieving average relative WER reduction of 26.8% (EA: 25.7%; MA: 27.0%; NA: 25.2%; SA: 34.8%).

Comparison with Speech-LLM Methods. SLAM-ASR shows severe degradation on academic domains (EA: 41.24%, SA: 62.63%, NA: 88.64%, MA: ∞%), contrasting with its strong LibriSpeech performance. This reveals a fundamental limitation: single linear projection lacks capacity to bridge the modality gap [11]. When encountering unfamiliar inputs, misaligned embeddings cause excessive insertion errors, resulting in WER exceeding 100%. SALMONN maintains stable performance (EA: 17.61%, MA: 24.00%, NA: 24.61%, SA: 17.77%) but with limited improvement. The frozen LLM with causal convolutional adapters achieves robust cross-domain performance by capturing temporal dependencies in speech representations through causal convolutions, enabling effective feature extraction while preserving temporal structure.

Table 3. Word Error Rate (%) on academic domains. All methods trained only on general-domain data. Lower is better. Best results in each column are shown in **bold**. †SLAM-ASR achieves ∞ WER on Medical Area (MA) due to excessive insertion errors (actual value: 136.46%).

Method	**EA**	**MA**	**NA**	**SA**
Whisper-large-v2 [28]	19.25	24.40	24.31	19.98
SLAM-ASR [24]	41.24	$\infty^{\dagger}$	88.64	62.63
SALMONN [30]	17.61	24.00	24.61	17.77
Ours (Generic Prompt)	14.29	17.80	**18.18**	13.03
Ours (Domain Prompt)	**13.99**	**17.67**	18.37	**12.86**

Domain Prompting Effectiveness. Domain-specific prompts further improve performance in three domains: Engineering (2.10% relative improvement), Social Sciences (1.30%), and Medical Sciences (0.73%). Natural Sciences shows slight decrease (-1.04%), suggesting prompt sensitivity when domain descriptions are overly broad or when the LLM's pre-training already covers the domain extensively. Empirically, prompts are more effective when they include both an explicit domain label and an instruction to preserve domain terminology. In contrast, broad descriptors without domain-specific lexical cues tend to yield smaller gains. Despite this, average improvement across three successful domains is 1.38%.

5.3 Domain-Specific Terminology Recognition

Recognition of domain-specific words unseen during training is evaluated by identifying words in test sets but not in training data, ranking by frequency, and selecting the top 20 words after removing abbreviations and hyphenated terms. Table 4 shows domain prompting significantly improves terminology recognition. F1 scores are improved by 7.2 percentage points on average (EA: 43.9%→52.0%, MA: 38.5%→43.5%, NA: 32.5%→46.2%, SA: 41.3%→43.3%), with recall improved by 5.8 points while high precision is maintained (all above 97%).

5.4 Parameter Efficiency Analysis

Table 5 shows the number of parameters by component. Inference-time domain prompting is achieved with only 24.6M trainable parameters (0.44%), providing memory efficiency, knowledge preservation, and prompt-only adaptation.

6 Conclusion

This work presents an adapter-only approach for domain-adaptive speech recognition that freezes both the speech encoder (Whisper-large-v2) and LLM (Gemma-3-4B-Instruct), training only 24.6M adapter parameters (0.44% of total). The causal convolutional downsampler and projection MLP bridge the audio-text

Table 4. Word-level recognition results for domain-specific terminology (top-20 most frequent unseen words per domain, excluding abbreviations and hyphenated terms). Higher is better. Best results in each metric/domain cell are shown in **bold**.

Method	**EA**	**MA**	**NA**	**SA**
F1 (%)				
Ours (Generic Prompt)	43.90	38.50	32.50	41.30
Ours (Domain Prompt)	**52.00**	**43.50**	**46.20**	**43.30**
Precision (%)				
Ours (Generic Prompt)	97.20	96.30	100.00	100.00
Ours (Domain Prompt)	**97.70**	**98.50**	**100.00**	**100.00**
Recall (%)				
Ours (Generic Prompt)	28.40	24.00	19.40	26.00
Ours (Domain Prompt)	**35.40**	**27.90**	**30.10**	**27.70**

Table 5. Number of parameters by component

Component	**Parameters**	**Training Mode**
Whisper Encoder	636.8M	Frozen
Gemma-3 LLM	5.0B	Frozen
Conv Downsampler	14.8M	Trainable
MLP Projection	9.8M	Trainable
Total Trainable	24.6M (0.44%)	–
Total Parameters	5.6B	–

modality gap while preserving the frozen LLM's linguistic knowledge. Experiments on academic domains demonstrate 26.8% average relative WER reduction over Whisper-large-v2 without domain-specific training. Inference-time domain prompting further improves performance in three of four domains (average 1.38% relative improvement) and enhances domain-specific terminology recognition (7.2 percentage point F1 improvement).

The adapter-only architecture addresses key limitations of existing speech-LLM methods: catastrophic domain transfer failures in linear projection approaches and catastrophic forgetting in joint fine-tuning methods. The proposed design enables practical deployment where users maintain a single frozen model and adapt through lightweight modules and text prompts.

Limitations and Future Directions. Current constraints include domain-dependent prompt effectiveness, reliance on LLM pre-training coverage, English-only evaluation, limited model-scale analysis, missing runtime efficiency reporting (e.g., latency/throughput), and the absence of a full adapter ablation under identical training budgets (e.g., causal vs. non-causal downsampling). Future work includes automated prompt optimization, multilingual extension, comprehensive adapter ablations, model-scale sensitivity analysis, and hierarchical prompting strategies for improved terminology recognition.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference (LREC). pp. 4218–4222 (2020)
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS). pp. 1877–1901 (2020)
3. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4960–4964 (2016)
4. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., Wei, F.: BEATs: Audio pre-training with acoustic tokenizers. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 5178–5193 (2023)
5. Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K.C., Li, J., Ghosh, S., Balam, J., Ginsburg, B.: SALM: Speech-augmented language model with in-context learning for speech recognition and translation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 13521–13525 (2024)
6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. `https://lmsys.org/blog/2023-03-30-vicuna/` (2023)
7. Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., Zhou, J.: Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919 (2023)
8. Cui, W., Yu, D., Jiao, X., Meng, Z., Zhang, G., Wang, Q., Guo, S.Y., King, I.: Recent advances in speech language models: A survey. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 13943–13970 (2025)
9. Deng, K., Woodland, P.C.: Decoupled structure for improved adaptability of end-to-end models. Speech Communication **163**, 103109 (2024)
10. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: CLAP: Learning audio concepts from natural language supervision. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)

11. Fang, Q., Feng, Y.: Understanding and bridging the modality gap for speech translation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). pp. 15864–15881 (2023)
12. Fathullah, Y., Wu, C., Lakomkin, E., Li, K., Jia, J., Shangguan, Y., Mahadeokar, J., Kalinli, O., Fuegen, C., Seltzer, M.: AudioChatLlama: Towards general-purpose speech abilities for LLMs. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 5522–5532 (2024)
13. Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025). https://doi.org/10.48550/arXiv.2503.19786, `https://arxiv.org/abs/2503.19786`
14. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning (ICML). pp. 1764–1772 (2014)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)
16. Hono, Y., Mitsuda, K., Zhao, T., Mitsui, K., Wakatsuki, T., Sawada, K.: Integrating pre-trained speech and language models for end-to-end speech recognition. In: Findings of the Association for Computational Linguistics (ACL). pp. 13289–13305 (2024)
17. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
19. Hu, S., Zhou, L., Liu, S., Chen, S., Meng, L., Hao, H., Pan, J., Liu, X., Li, J., Sivasankaran, S., Liu, L., Wei, F.: WavLLM: Towards robust and adaptive speech large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 4552–4572 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.263
20. Kim, D., Lee, G., Shim, K., Shim, B.: Preserving pre-trained representation space: On effectiveness of prefix-tuning for large multimodal models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 797–819 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.44
21. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 19730–19742 (2023)
22. Li, S., Tang, H.: Multimodal alignment and fusion: A survey. International Journal of Computer Vision **134** (2024), `https://api.semanticscholar.org/CorpusID:274280969`
23. Ling, S., Hu, Y., Qian, S., Ye, G., Qian, Y., Gong, Y., Lin, E., Zeng, M.: Adapting large language model with speech for fully formatted end-to-end speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11046–11050 (2024). https://doi.org/10.1109/ICASSP48485.2024.10448204
24. Ma, Z., Yang, G., Yang, Y., Gao, Z., Wang, J., Du, Z., Yu, F., Chen, Q., Zheng, S., Zhang, S., Chen, X.: An embarrassingly simple approach for LLM with strong ASR capacity. arXiv preprint arXiv:2402.08846 (2024)

25. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: An ASR corpus based on public domain audio books. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210 (2015)
26. Peyser, C., Meng, Z., Hu, K., Prabhavalkar, R., Rosenberg, A., Sainath, T.N., Picheny, M., Cho, K.: Improving joint speech-text representations without alignment. In: Proceedings of Interspeech. pp. 1354–1358 (2023)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
28. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 28492–28518 (2023)
29. Rousseau, A., Deléglise, P., Estève, Y.: TED-LIUM: an automatic speech recognition dedicated corpus. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC). pp. 125–129 (2012)
30. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Zhang, C.: SALMONN: Towards generic hearing abilities for large language models. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR) (2024)
31. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
32. Yu, W., Tang, C., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Zhang, C.: Connecting speech encoder and large language model for ASR. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12637–12641 (2024)
33. Zhang, Y., Liu, Z., Bu, F., Zhang, R., Wang, B., Li, H.: Soundwave: Less is more for speech-text alignment in LLMs. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 18718–18738 (2025)