Chain-of-Thought Distillation for ASR Error Correction with Multimodal Large Language Models

Shaomeng Yang[1,2], Jiaming Luo[4], Jinran Wang[1,5], Rongfeng Su[1,3], Yongjie Zhou[6], Lan Wang[1,3*], Nan Yan[1,3*]
[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences, China
[4] Shanghai Jiao Tong University, China
[5] Wuhan Research Institute of Posts and Telecommunications, China
[6] Shenzhen Kangning Hospital, Shenzhen, Guangdong, China
E-mail: {sm.yang, jr.wang2, rf.su, lan.wang, nan.yan}@siat.ac.cn, leojm2017@sjtu.edu.cn, zhouyj2023@mail.sustech.edu.cn

***Abstract*— Large language models (LLMs) have demonstrated strong performance in automatic speech recognition (ASR) error correction. However, current approaches rely heavily on annotated datasets for fine-tuning LLMs, which are often costly and challenging to obtain in domain-specific scenarios. To address this limitation, we propose a novel framework that employs Chain-of-Thought Distillation (CoTD) to automatically generate annotated data using a single-modal teacher model at scale. The distilled knowledge is subsequently transferred to a multimodal student model capable of processing both audio and text modalities. Experimental results indicate that fine-tuning multimodal large language models with a combination of distillation data and a limited account of human-annotated data achieves a relative CER drop compared to the baseline, and a comparable performance to the fine-tuned model utilizing five times the amount of human-annotated data.**

I. INTRODUCTION

Recent advancements in automatic speech recognition (ASR) have significantly improved transcription accuracy through deep learning architectures, such as end-to-end neural networks and self-supervised pretraining frameworks [1-5]. These innovations have enabled ASR systems to handle diverse acoustic conditions and speaker variations with increasing robustness. However, enduring technical barriers persist in contemporary speech recognition systems, particularly within specialized operational domains. These challenges stem from interdependent factors: the prevalence of discipline-specific terminologies and complex syntactic structures, acoustic interference from environmental noise coupled with speaker variability in accents and nonstandard articulation patterns, and insufficient linguistic resources for underrepresented languages characterized by scarce annotated corpora and morphological complexity. The resultant transcription inaccuracies manifest as systematic errors rather than random noise, predominantly through phonemic confusions (particularly in tonal languages), morphological-syntactic dislocations, and semantic inconsistencies. Such deficiencies critically undermine the functional integrity of language processing applications, including but not limited to conversational interfaces, multilingual machine translation pipelines, and sentiment analysis engines. This operational reality necessitates post-processing correction mechanisms. Consequently, researchers have drawn attention to the task of ASR error correction.

Traditional ASR error correction methods rely on rule-based systems or statistical language models [6-7], which often struggle to generalize across domains. Recent efforts leverage large language models (LLMs) to address this limitation [8-10], capitalizing on their contextual understanding and generative capabilities. While LLMs exhibit notable capabilities through fine-tuning approaches, their practical deployment in real-world scenarios is constrained by an inherent reliance on extensive annotated training corpora. This dependency poses a formidable obstacle, particularly in technical domains such as psychological counseling, biomedical research, legal analysis, and specialized engineering fields, where both the scarcity of domain-specific textual data and the requisite expertise for accurate annotation create substantial implementation barriers. The annotation process in these specialized contexts frequently necessitates involvement from credentialed professionals, thereby escalating operational costs and temporal investments to prohibitive levels. Furthermore, the quality and consistency of human-generated labels in low-resource domains remain persistent challenges that directly impact model generalizability. These fundamental limitations continue to motivate research into alternative methodologies that could potentially reduce annotation dependence while maintaining model efficacy.

*Corresponding authors.

979-8-3315-7206-8/25/$31.00

This work proposes a novel framework that mitigates reliance on annotated data by distilling error-correction knowledge from a single-modal teacher LLM to a multi-modal student model using Chain-of-Thought Distillation (CoTD). The teacher model generates synthetic correction trajectories through chain-of-thought reasoning, simulating human-like error analysis and refinement processes. These trajectories are then used to train the student model, which integrates both audio and text modalities to enhance contextual disambiguation. By combining distilled data with limited human-annotated samples, our approach achieves a 57.1% relative character error rate (CER) reduction over the baseline, while maintaining competitive performance compared to models trained on five times the amount of human-labeled data. This paper introduces a scalable, annotation-efficient paradigm for ASR error correction, decoupling knowledge generation from multi-modal inference. Extensive experiments demonstrate that hybrid training with distilled and human-annotated data achieves near-supervised performance with minimal annotation costs, offering practical value for low-resource domains. We bridge the gap between data-intensive LLM fine-tuning and real-world ASR correction demands, advancing toward adaptable, resource-efficient speech processing systems.

II. Related work

A. ASR error correction

ASR error correction aims to rectify grammatical inaccuracies and enhance text fluency in ASR outputs, serving as a crucial post-processing step in speech recognition systems. Early approaches primarily relied on rule-based systems and statistical language models [6-7]. While effective in domain-specific scenarios, these methods required extensive expert-crafted linguistic rules and demonstrated limited generalizability.

The advent of deep learning has facilitated the application of pre-trained language models for text refinement. Salazar et al. [11] employed a pre-trained BERT model to score multiple ASR hypotheses for optimal transcription selection. Subsequent studies introduced transformer-based architectures for semantic correction: Zhao et al. [12] proposed a BART-based semantic error correction system, while Hrinchuk et al. [13] developed an encoder-decoder transformer framework to "translate" ASR outputs into grammatically and semantically correct text. Although these end-to-end models automatically learn error pattern mappings through data-driven training, their performance remains constrained in low-resource scenarios due to dependency on large-scale annotated datasets.

The emergence of LLMs has instigated a paradigm shift in ASR error correction methodologies. Chen et al. [10] proposed leveraging external LLMs for error correction by utilizing N-best decoding hypotheses to extract informative elements for accurate transcription prediction. Li et al. [14] enhanced correction capabilities through fine-tuning a multilingual LLM spanning over 100 languages, enabling cross-lingual transfer of error correction knowledge when processing 1-best hypotheses from various speech foundation models. However, text-driven LLMs may disregard critical acoustic cues embedded in speech signals.

Recent research explores multimodal fusion to transcend text modality limitations. Shu et al. [15] pioneered this direction with ACR-Net, which jointly leverages acoustic features and confidence scores through cross-attention mechanisms for multimodal alignment. Chen et al. [9] integrated acoustic data into the generative correction process to improve mapping from N-best ASR hypotheses to accurate transcriptions. Wei et al. [16] further demonstrated that multimodal augmentation achieves superior performance compared to unimodal LLM approaches through comprehensive error correction experiments.

B. Knowledge Distillation

The core concept of knowledge distillation lies in transferring knowledge from a teacher model to a more lightweight student model, thereby maintaining performance in resource-constrained scenarios. Hinton et al. [17] pioneered the concept of using "softened" output probabilities (i.e., soft labels) from the teacher model as supervisory signals to guide the student model in learning implicit inter-class relationships. Subsequent research has expanded this framework through approaches such as model distillation using intermediate representations [18]. However, these methods were primarily designed for vision tasks and face significant challenges when directly applied to language models, particularly regarding sequence generation uncertainty and semantic coherence.

With the exponential growth of pretrained language model scales, distillation methods tailored for LLMs have emerged as a critical research focus. Early work by Sanh et al. [19] successfully compressed BERT to 40% of its original parameter count through dynamic masked language modeling combined with teacher soft-label distillation. Wang et al. [20] advanced this direction by focusing on transferring the self-attention module of teacher models to enhance the student model's semantic comprehension capabilities.

Wei et al. [21] first demonstrated that explicitly generating intermediate reasoning steps (e.g., "Step 1: ... Step 2: ...") significantly improves LLMs' performance on multi-step reasoning tasks. Building upon this foundation, Hsieh et al. [22] proposed "Distilling Step-by-Step," which leverages teacher-generated Chain-of-Thought (CoT) rationales as additional supervisory signals. Their approach employs a multi-task learning framework that jointly optimizes answer prediction and reasoning step generation, effectively enhancing student models' complex task-handling capabilities.

III. Methodology

This section presents the proposed method for fine-tuning multi-modal LLM through CoT distillation, as illustrated in Fig.1. Our dataset consists of Chinese psychological counseling recordings, which are processed utilizing FunASR's

end-to-end speech recognition system [1], we employ the paraformer-zh acoustic model for Mandarin recognition, fsmn-vad for voice activity detection (VAD), ct-punc for punctuation restoration, and cam++ for speaker diarization [23-27]. The ASR outputs were cut into segmentations with voice activity detection results, generating audio-text segment pairs. For the knowledge distillation, we implement Qwen2.5-72B language model as the teacher model[2]. Qwen2.5-72B [28] was pre-trained on 18 trillion tokens, it is a powerful open-sourced LLM that can handle complicated text tasks. We employ CoT prompting to design the prompt for error correction task. The processing flow is structured as follows: (1) Syntactic decomposition of input sentences, (2) Systematic detection of orthographic, grammatical, and semantic anomalies across constituent components, (3) Iterative refinement with contextual coherence verification, and (4) Final validation of semantic consistency against original transcripts. To enhance computational efficiency while maintaining reasoning capability, we adopt a one-shot prompting strategy that provides contextual exemplars of error correction with explicit reasoning chains. This approach generates annotated data containing both corrected text outputs and their corresponding reasoning trajectories, which can be distilled for training the student model's ability of reasoning.

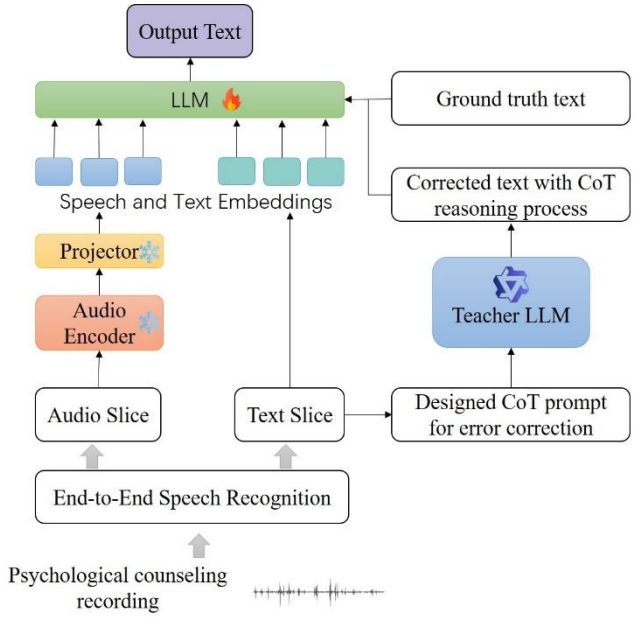

Fig. 1 Overview of proposed method

We subsequently fine-tuned Qwen2-audio[3] model using annotated data introduced above. Qwen2-audio [29], a multi-modal LLM integrating audio and text modalities, contains a Whisper-large-v3-based [30] audio encoder pre-trained on 520,000 hours of audio data, which demonstrate robust capabilities in audio analysis. Through fine-tuning, the model can adapt to the task-specific paradigm of ASR error correction by synergizing audio and textual information to generate more accurate text correction outputs. Therefore, we fine-tuned the model with raw audio signals and ASR-generated transcripts using Low-Rank Adaptation (LoRA) [31] technique. The output labels consist two components: (1) Text correction results with reasoning processes derived from CoT prompting, and (2) A limited set of manually annotated results. The proposed CoTD method involves fine-tuning the model in two sequential stages. Stage 1 focuses on fine-tuning the model using CoT-generated labels. This enabled the model to acquire multi-step reasoning capabilities by learning from structured logical sequences in the labels. Stage 2 involves secondary fine-tuning with human-annotated data, refining the model's ability to produce precise text corrections.

IV. Experiments

A. Experimental Settings

The original dataset comprises Chinese psychological counseling recordings. These recordings were processed through an ASR pipeline. Through VAD results, these recordings were segmented into 26,682 audio-text pairs, each under 30 seconds in duration. From this collection, 1,117 segments were selected as the test set, while 5,133 segments were designated for comparing the difference between auto-annotation and manual annotation. This combined subset of 6,250 segments was meticulously annotated by human to establish ground truth transcripts. The remaining 20,432 segments underwent automated refinement using the Qwen2.5-72B model to modify ASR-generated text outputs.

For error correction processing, we implemented a chain-of-thought prompting strategy to guide large language model interventions. The model outputs consisted of two components: reasoning chains and final corrected text. This enabled the creation of two distinct training datasets: 1) Distillation data incorporating complete reasoning processes alongside final corrections, and 2) Distillation data containing only final correction outputs. Both datasets were subsequently utilized for separate model fine-tuning procedures. To evaluate the model's performance with a limited quantity of manually annotated data, we randomly sampled 1,000 audio segments from the 5,133 available ones (accounting for 20.3% of the total duration). The model was then trained on both the entire dataset and the smaller subset, aiming to discern the impact of manually annotated data on the model training process. To determine whether the distilled data could substitute a portion of the manual data, we extracted 4,133 segments from the entire distilled dataset. These segments were used to supplement the small quantity of existing data, ensuring that the combined amount matched the total volume of manually

[1] https://github.com/modelscope/FunASR

[2] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

[3] https://huggingface.co/Qwen/Qwen2-Audio-7B

annotated data. Therefore, we will apply this limited data for the two-stage training method.

During the fine-tuning phase, we froze parameters in both the speech encoder and the projector, exclusively updating the LLM components. We set rank=8 and alpha=32 for LoRA and fine-tuned for 1 epoch with the learning rate of 1e-4. All experiments ran on NVIDIA A6000 GPUs with batch size equals to 1.

For the baseline model, we adopt the pre-trained model proposed in [10], which is fine-tuned from the LLaMA-7B model on the HP dataset using the H2T-LoRA training method introduced in the paper. This method effectively enhances the model's ASR error correction capability. Since the model is unimodal, we only use the ASR-transcribed text as input to obtain the error-corrected result.

TABLE I

ERROR CORRECTION PERFORMANCE OF FINE-TUNED QWEN2-AUDIO ON DIVERSE DATASETS

Model	**CER (%) ↓**
Baseline	21.9
Qwen2-audio	17.2
Qwen2-audio +CoT distillation data	16.3
Qwen2-audio +Non-CoT distillation data	17.3
Qwen2-audio +Human-annotated data	**8.6**
Qwen2-audio +Human-annotated data(limited)	14.4

B. Experimental Results

This section presents the experimental results of our study. The student model underwent multiple fine-tuning sessions on partitioned datasets, with performance evaluated on the test set using the Character Error Rate (CER) as the primary metric. Table I summarizes the results of fine-tuning the original Qwen2-audio model with different data sources, we apply four different datasets, Non-CoT indicates we only utilize the text result without CoT process as training data, limited data represents we only employ 20% of the total amount of human-annotated data. As presented in Table I, Qwen2-audio achieved better performance than the baseline, however, both direct utilization of distilled outputs (Non-CoT) and incorporation of inference processes (CoT) failed to further improve the results. This phenomenon suggests that knowledge distilled solely from teacher models cannot fully replicate real-world data distributions, thereby impeding task adaptation. In contrast, fine-tuning with human-annotated data yielded a significant CER reduction of 60.7% relative to the baseline, demonstrating the model's ability to acquire task-specific patterns and improve domain adaptability through exposure to authentic data. However, real-world scenarios often face constraints in acquiring large-scale annotated datasets. To address this, we further evaluated models fine-tuned with limited human-annotated samples. The results revealed that constrained training data substantially hampered performance, achieving only a 16.3% CER reduction compared to the original Qwen2-audio model (17.2→14.4). This underscores the necessity of sufficient training data to enable effective task learning.

TABLE II

TWO-STAGE FINE-TUNED QWEN2-AUDIO RESULTS

Dataset	**CER (%) ↓**
Non-CoT distillation data +Human-annotated data(limited)	13.2
CoT distillation data +Human-annotated data(limited)	**9.4**

Building upon the initial distillation-based fine-tuning, we further refined the student model using limited human-annotated data. Table II presents error correction results using a two-stage fine-tuning method. CoT and Non-CoT represent the dataset we used for training in the first stage. Both models employ the limited human-annotated data for the second stage's training. The results demonstrate significant performance improvements in the secondary fine-tuning phase. Notably, distilled data incorporating chain-of-thought reasoning achieved a substantial CER reduction of 57.1% compared to the baseline, with its performance gap narrowing to 9.3% relative to the fully fine-tuned model using exhaustive human-annotated data. Furthermore, distilled data containing only text correction outputs exhibited inferior error rate reduction when combined with limited human-annotated data, suggesting that the reasoning patterns learned during the first-stage fine-tuning synergistically enhanced task-specific capabilities when integrated with authentic supervision in subsequent training. This observation implies that the model effectively assimilated logical inference pathways from the distilled data during preliminary optimization, thereby amplifying its adaptability to domain-specific tasks when exposed to limited real-world data.

In summary, our experimental findings demonstrate that fine-tuning the Qwen2-audio model with human-annotated data from the same domain yields optimal performance in ASR error correction tasks. Notably, under scenarios where manual annotation resources are unavailable and only limited authentic data exist, we have validated an effective alternative approach: the implementation of knowledge distillation through chain-of-thought reasoning outputs generated by a larger-parameter single-modality teacher model. This methodology enabled the construction of augmented training datasets that achieved performance parity comparable to full-scale human-annotated

TABLE III
EXAMPLES OF ASR ERROR CORRECTION

Model	Utterance
Ground Truth	系统地按时地规律去服药(take medicine regularly and on time)
ASR output	系统的按神的规律去服药(take medicine according to God)
CoT distillation	按时按神的规律去服药(take medicine according to God on time)
Human-annotated data(limited)	系统地按照规律去服药(take medicine regularly)
CoT distillation+Human-annotated data(limited)	系统地按时规律地去服药(take medicine regularly and on time)

fine-tuning, demonstrating comparable efficacy in error correction capability enhancement.

C. *Case Analysis*

This section presents a case analysis of output performance across different models, as demonstrated in Table III. For the different output, we assign distinct colors to indicate the error. The model fine-tuned on both CoT distillation data and human-annotated data achieved optimal results, faithfully reconstructing the original sentence. The model trained only on CoT distillation data failed to correct ASR errors, while the model trained only on human-annotated data corrected errors but produced incomplete sentence reconstructions with missing information. This outcome underscores the practical viability of integrating CoT distillation data with authentic datasets for model optimization. Notably, the hybrid training approach enables models to effectively identify erroneous semantic expressions in ASR outputs and implement appropriate corrections. In contrast, models exclusively trained on CoT distillation data failed to detect these errors, revealing inherent limitations in relying solely on reasoning paradigms for error correction tasks.

V. Conclusion

This study aimed to address the dependency on annotated datasets for ASR error correction by proposing a novel framework employing Chain-of-Thought Distillation. The method leverages a single-modal teacher model to automatically generate large-scale annotated data, which is then transferred to a multi-modal student model. The primary contribution lies in demonstrating that combining distilled data with limited human-annotated data significantly reduces the CER, and maintaining competitive performance compared to models trained on five times the amount of human-annotated data. This approach reduces reliance on costly domain-specific annotations and highlights the potential of multi-modal fusion for error correction. However, the study has limitations. First, the experiments were conducted under constrained domain-specific scenarios, and the generalizability of Chain-of-Thought Distillation to cross-domain or multilingual settings remains unverified. Additionally, the distillation process assumes high-quality outputs from the teacher model, which may not hold in low-resource or noisy environments. In the future, we will continue our research by exploring CoTD to cross-modal knowledge transfer, such as incorporating visual or contextual cues for error correction. We are also interested in optimizing the distillation pipeline to mitigate error propagation from the teacher model, obtaining the auto-annotated data with enhanced quality.

VI. Acknowledgment

This research is supported by National Natural Science Foundation of China (U23B2018), National Natural Science Foundation of China (NSFC 62271477), Shenzhen Science and Technology Program (JCYJ20220818101411025), Shenzhen Science and Technology Program (JCYJ20220818102800001), Shenzhen Science and Technology Program (JCYJ20220818101217037), Shenzhen Peacock Team Project (KQTD20200820113106007).

References

[1] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results," CoRR, vol. abs/1412.1602, 2014.

[2] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in International Conference on Machine Learning, 2014.

[3] D. Amodei et al., "Deep speech 2: end-to-end speech recognition in English and mandarin," in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, in ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 173–182.

[4] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015.

[5] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in Proceedings of the 35th International Conference on Neural Information Processing Systems, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[6] H. Cucu, A. Buzo, L. Besacier, and C. Burileanu, "Statistical Error Correction Methods for Domain-Specific ASR Systems," in International Conference on Statistical Language and Speech Processing, 2013.

[7] S. Jung, M. Jeong, and G. G. Lee, “Speech recognition error correction using maximum entropy language model,” in Interspeech, 2004.

[8] Y. Hu et al., “Large Language Models are Efficient Learners of Noise-Robust Speech Recognition,” ArXiv, vol. abs/2401.10446, 2024.

[9] C. Chen et al., “It’s Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition,” in The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024.

[10] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E. S. Chng, “HyPoradise: an open baseline for generative speech recognition with large language models,” in Proceedings of the 37th International Conference on Neural Information Processing Systems, in NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[11] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, “Masked Language Model Scoring,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 2699–2712.

[12] Y. Zhao, X. Yang, J. Wang, Y. Gao, C. Yan, and Y. Zhou, “BART based semantic correction for Mandarin automatic speech recognition system,” in Interspeech, 2021.

[13] O. Hrinchuk, M. Popova, and B. Ginsburg, “Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7074–7078.

[14] S. Li, C. Chen, C. Y. Kwok, C. Chu, E. S. Chng, and H. Kawai, “Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses,” in Interspeech 2024, 2024, pp. 1315–1319.

[15] Y. Shu, B. Hu, Y. He, H. Shi, L. Wang, and J. Dang, “Error Correction by Paying Attention to Both Acoustic and Confidence References for Automatic Speech Recognition,” in Interspeech 2024, 2024, pp. 3500–3504.

[16] V. J. Wei, W. Wang, D. Jiang, Y. Song, and L. Wang, “ASR-EC Benchmark: Evaluating Large Language Models on Chinese ASR Error Correction,” ArXiv Prepr. ArXiv241203075, 2024.

[17] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” in NIPS Deep Learning and Representation Learning Workshop, 2015.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for Thin Deep Nets,” CoRR, vol. abs/1412.6550, 2014.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” CoRR, vol. abs/1910.01108, 2019.

[20] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in Proceedings of the 34th International Conference on Neural Information Processing Systems, in NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[21] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Proceedings of the 36th International Conference on Neural Information Processing Systems, in NIPS ’22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[22] C.-Y. Hsieh et al., “Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes,” in Findings of the Association for Computational Linguistics: ACL 2023, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8003–8017.

[23] Z. Gao et al., “FunASR: A Fundamental End-to-End Speech Recognition Toolkit,” in INTERSPEECH, 2023.

[24] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, “Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition,” in Proc. Interspeech 2022, 2022, pp. 2063–2067.

[25] S. Zhang, M. Lei, Z. Yan, and L. Dai, “Deep-FSMN for large vocabulary continuous speech recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5869–5873.

[26] Q. Chen, M. Chen, B. Li, and W. Wang, “Controllable Time-Delay Transformer for Real-Time Punctuation Prediction and Disfluency Detection,” ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, pp. 8069–8073, 2020.

[27] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, “CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking,” in Interspeech 2023, 2023, pp. 5301–5305.

[28] Q. A. Yang et al., “Qwen2.5 Technical Report,” ArXiv, vol. abs/2412.15115, 2024.

[29] Y. Chu et al., “Qwen2-Audio Technical Report,” ArXiv, vol. abs/2407.10759, 2024.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision.” 2022.

[31] J. E. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” ArXiv, vol. abs/2106.09685, 2021.