Internalizing ASR with Implicit Chain of Thought for Efficient Speech-to-Speech Conversational LLM

Anonymous ACL submission

Abstract

Current Speech LLMs are predominantly trained on extensive ASR and TTS datasets, excelling in tasks related to these domains. However, their ability to handle direct speech-to-speech conversations remains notably constrained. We find that Speech LLMs often rely on an ASR-to-TTS chain-of-thought pipeline (A-T-T-A chain) to generate good responses. The pipeline first recognizes speech into text and generates corresponding text responses before generating speech responses, which introduces significant latency. We propose a method that implicitly internalizes ASR chain of thought into a Speech LLM (A-T-A chain), allowing it to bypass the ASR transcript generation but still maintain speech conversation capabilities. Our approach reduces latency and improves the model's native understanding of speech, paving the way for more efficient and natural real-time audio interactions. We also release a large-scale synthetic conversational dataset to facilitate further research.

1 Introduction

Pretrained large speech-language models (Speech LLMs) (e.g., Zhang et al., 2023a, 2024a; Zhan et al., 2024; Chu et al., 2023, 2024) are an emerging paradigm for better intelligence in various speech and language tasks. Different from traditional speech models like automatic speech recognition (ASR) and text-to-speech synthesis (TTS) models that focus only on individual tasks in the perception and the production domain, pretrained Speech LLMs can potentially unify these tasks and perform more complex reasoning tasks for more realistic speech conversations (Cui et al., 2024).

While most current research in Speech LLMs focuses on scaling up datasets and model parameters to enhance traditional tasks such as ASR, TTS, and emotion recognition (e.g., Ao et al., 2022; Chu et al., 2023, 2024), there has been limited exploration of their ability to handle broader conversational reasoning tasks that text-based LLMs excel at (e.g., Cui et al., 2024; Défossez et al., 2024).

The mainstream approach for developing Speech LLMs is to utilize a pretrained text LLM that has already acquired strong multitask performance, and then continue to pre-train it on discretized speech tokens such as SpeechGPT series (Zhang et al., 2023a; Zhan et al., 2024), or augment it with a speech encoder to process raw waveforms such as Qwen Audio series (Chu et al., 2023, 2024). Recent studies show that this class of Speech LLMs rely heavily on the base text models, such that allowing them to generate text transcripts in intermediate inference steps can improve their speech understanding and generation abilities (Gong et al., 2023; Huang et al., 2023; Hu et al., 2024; Défossez et al., 2024; Xie and Wu, 2024). However, these intermediate text generations can significantly increase the token counts, increasing the latency of Speech LLMs' responses. Simplifying this pipeline is critical for developing speech models that can engage in natural and coherent dialogue, without relying solely on intermediate text representations.

In this work, we investigate the use of Low Rank Adaptation (LoRA) (Hu et al., 2021) for fine-tuning Speech LLM's ability to perform open-ended conversations. We propose a novel approach that leverages the ASR and TTS capabilities of pretrained Speech LLMs to enable more natural speech conversations with lower latency. Our contributions can be summarized as follows:

- We observed that direct finetuning of Speech LLMs on a medium scale of audio-only dataset yields incomprehensible speech conversation ability. Yet mixing the ASR transcripts as Chain of Thought (CoT) (Wei et al., 2022; Zhang et al., 2023a) with input and output speech in the finetuning data yields significantly better performance, at the cost of increased latency and data requirements.

- To reduce the length of CoT tokens, we further propose to internalize ASR CoT tokens into a Speech LLM, retaining the speech conversation performance while reducing the latency by 20.2%. This approach brings us closer to a fully text-free end-to-end Speech LLM. However, we still find that we cannot fully remove the TTS CoT tokens for generating good responses. The code is publicly available on GitHub, and the pretrained models are publicly available on Huggingface Hub[1].

- We constructed a large-scale synthetic conversational speech dataset with an emphasis on social common sense reasoning, containing ∼660k dialogue exchange pairs totaling ∼1000 hours of speech data. The dataset is publicly available on Huggingface Hub [2].

2 Background

Pretrained Speech and audio language models Since the success of scaling up LLMs, there are many attempts to scale up the speech models by training on discrete speech tokens (Lakhotia et al., 2021) or interleaving texts and speech inputs, such as SpeechT5 (Ao et al., 2022), SpeechGPT series (Zhang et al., 2023a, 2024a; Zhan et al., 2024), Qwen-Audio series (Chu et al., 2023, 2024), Moshi (Défossez et al., 2024) and Mini-Onmi (Xie and Wu, 2024).

SpeechGPT (Zhang et al., 2023a) and AnyGPT (Zhan et al., 2024) enable LLMs to perform speech understanding and generation through discrete speech representations. Based on the pre-trained LLaMA2 7B model (Touvron et al., 2023), they were further finetuned with large amounts of ASR and TTS datasets, and as a result, can perform well on those tasks. However, attempts to perform direct speech-to-speech conversation with the model often result in incomprehensible audio. The reason behind this is not well understood, and may well be just a matter of insufficient scaling.

Chain-of-Thought Prompting in Speech Processing CoT prompting (Wei et al., 2022) is an effective method of improving LLMs' capability to perform complex reasoning tasks by detailing the intermediate steps. This technique has also been adapted to the multimodal domain, particularly in vision (Lu et al., 2022; Zheng et al., 2023a; Mondal et al., 2024; Zhang et al., 2024b). However, a side effect of CoT is to introduce many intermediate tokens that increase the latencies.

Yet the application of CoT prompting in the audio and speech domain is still relatively rare (Li et al., 2024). Emerging research in SpeechLLMs generally finds that incorporating text transcripts in the CoT prompts can generally improve the performance of downstream speech tasks (e.g., Gong et al., 2023; Huang et al., 2023; Hu et al., 2024). In speech translation, prompting with ASR transcripts can effectively improve the BLEU scores (Huang et al., 2023; Hu et al., 2024).

Implicit Chain of Thought (ICoT), proposed by Deng et al. (2024b), has demonstrated LLMs can perform CoT behavior without explicit intermediate steps. In Deng et al. (2024a), a curriculum learning approach was proposed to internalize the reasoning process by gradually removing the intermediate CoT tokens while retaining the reasoning performance. The ICoT fine-tuning can retain the original performance with fewer tokens, making it more efficient for reasoning tasks. The effectiveness of this approach was demonstrated in solving math problems. However, the ICoT method has not been tested on other language generation tasks. It remains unclear whether this internalization process can benefit other tasks.

3 Method

We incorporate the ICoT technique (Deng et al., 2024b,a) to internalize ASR capability into a pretrained Speech LLM, such that it can perform speech-to-speech conversation without explicit intermediate ASR steps. To the best of our knowledge, we are the first to utilize ICoT beyond math problems and into practical speech applications.

Base Model: AnyGPT We build on AnyGPT (Zhan et al., 2024), an instruction-finetuned model capable of both speech audio comprehending and generation. In particular, AnyGPT is partly instruction-tuned on ASR and TTS tasks, resulting in strong capabilities in these areas. It is built upon the LLaMA 2 7B model (Touvron et al., 2023), allowing it to inherit the strong zero-shot generalization capabilities that are characteristic of large language models (LLMs). This makes AnyGPT well-suited for our speech-driven conversational tasks.

[1] Code and pretrained models will be released after the double-blind review

[2] Dataset will be released after the double-blind review

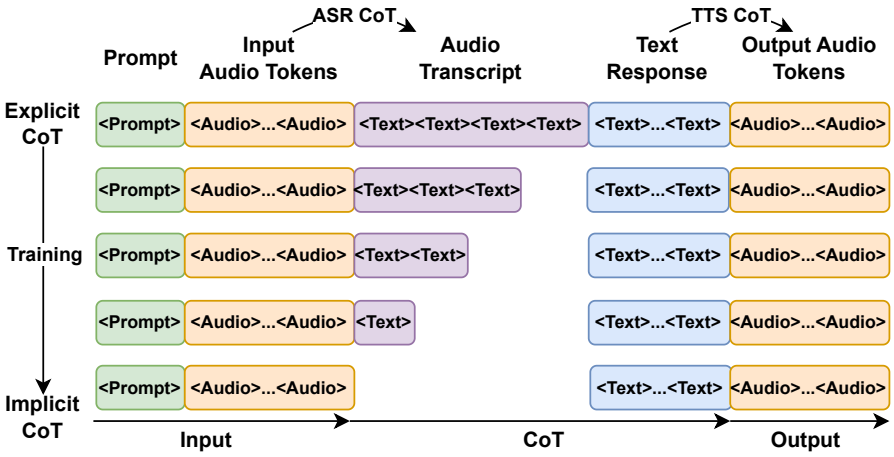

Figure 1: Illustration of ICoT training and generation structure (from A-T-T-A to A-T-A ASR ICoT). Tokens of audio transcripts are removed linearly from the start during training, compressing the generation length for faster inference.

Speech-to-Speech Conversation via ASR and TTS CoT Prompting Leveraging the ASR and TTS capability of AnyGPT, speech-to-speech conversations can be done using a multi-step CoT strategy with intermediate text interfaces. Building upon the framework established by SpeechGPT (Zhang et al., 2023a), the conversational process is structured as follows: given input audio, the model first transcribes the speech via ASR, then generates a textual response, and finally converts this text back into speech via TTS. Xie and Wu (2024) coined this pipeline as "**A-T-T-A**" (Audio-Text-Text-Audio), a terminlogy we adopt for clarity. Additionally, we define the process of converting input audio tokens into audio transcript as **ASR CoT**, and the conversion of text response to output audio tokens as **TTS CoT**.

To induce zero-shot CoT behavior, we prepend a fixed CoT prompt at the start of the generation template. The full generation process can be roughly summarized with the template: **"[CoT Prompt][Input Audio Tokens][Audio Transcript][Text Response][Output Audio Tokens]"**. The exact template used can be found in Table-1.

However, this approach introduces two additional intermediate steps: the transcription of ASR output and the generation of a text-based response, both of which contribute to increased inference latency and computational overhead.

ASR Internalization via Implicit CoT (A-T-A ASR ICoT) To mitigate the latency caused by intermediate text generation, we adopt ICoT reasoning, as outlined in (Deng et al., 2024a). This involves progressively internalizing the ASR reasoning steps during training, thereby eliminating the need for explicit transcription generation in the inference process, reducing the four-segment chain (**A-T-T-A**) to just three steps (**A-T-A**). The method is illustrated in Figure-1.

The training procedure is divided into two distinct stages:

1. **Standard CoT Fine-Tuning:** We do a full-parameter fine-tuning of the model using standard CoT on the TTS-transformed SODA dataset to ensure alignment with our conversational prompts and dataset.

2. **ICoT Training with LoRA** Next, we introduce ICoT training on the internalization of ASR transcription with LoRA (Hu et al., 2021). LoRA enhances the model's training efficiency by significantly reducing VRAM usage. The purpose of the initial stage is to stabilize the model's response behavior, creating a robust foundation for the more advanced ICoT training.

We define the input audio tokens as x, the intermediate audio transcript as $z^{ASR} = z_1^{ASR}, z_2^{ASR}, \cdots, z_m^{ASR}$, the intermediate text response as $z^{TTS} = z_1^{TTS}, z_2^{TTS}, \cdots, z_n^{TTS}$, and the final output audio tokens as y. The language model, with parameters θ, is initially trained with the following objective:

$$\min_{\theta} -\log P_\theta(y, z_{1:m}^{ASR}, z_{1:n}^{TTS} \mid x) \qquad (1)$$

After the pretraining phase, we linearly reduce the number of audio transcript tokens to achieve ICoT. The number of tokens removed at each step t is defined as:

$$s(t) = \min\left(\left\lfloor \frac{t}{T} + o \right\rfloor, K_i\right) \tag{2}$$

where $s(t)$ is the number of CoT tokens removed at step t, T is the number of steps per CoT token drop, K_i is the amount of CoT tokens for data point i, and o is a random variable sampled from an exponential distribution parameterized by λ.

$$P(o) \propto \exp(-\lambda o) \tag{3}$$

The new objective function becomes:

$$\min_{\theta} -\log P_\theta(y, z^{ASR}_{s(t):m}, z^{TTS}_{1:n} \mid x) \tag{4}$$

Training continues until all audio transcript tokens are removed.

Following the removal of ASR transcript tokens, the process for removing text response tokens z^{TTS} can begin using a similar token-dropping strategy, with t reset to zero at the start of this phase:

$$\min_{\theta} -\log P_\theta(y, z^{TTS}_{s(t):n} \mid x) \tag{5}$$

This approach further simplifies the 3-step chain (**A-T-A**) to a more efficient 2-step process (**A-A**), where the model directly generates audio from the input without intermediate textual stages. However, since the primary focus of our study is on ASR ICoT, we leave additional exploration of TTS ICoT, such as its scheduling and integration with ASR ICoT, for future work.

To ensure stability during training, we also employed the optimizer reset strategy from Deng et al. (2024a). Every T steps, the optimizer's state was reset, reinitializing both first and second-order gradients. This method prevents large estimates of second-order gradients resulting from a sudden removal of CoT tokens.

Baselines To provide a meaningful comparison of our model's performance, we evaluate against the following baseline approaches below. The exact prompt of the baselines can be found in Table-1, and a summary of the differences between the baselines can be found in Table-2.

1) A-T-T-A Finetuned To account for potential domain mismatches between the pre-training data and the target evaluation data, we fine-tune AnyGPT using our custom training prompts that include the ASR transcript. This model tests the effectiveness of domain-specific finetuning on model performance while keeping the ASR CoT mechanism intact. The structure of the prompt used was `"[CoT Prompt][Input Audio Tokens][Audio Transcript][Text Response][Output Audio Tokens]"`, i.e. **A-T-T-A**. In practice, we utilize the model obtained from the standard CoT finetuning (the first stage of our proposed model) for this evaluation.

2) A-T-A no ASR ICoT The model is fine-tuned directly on the conversational task without incorporating the ASR CoT mechanism. Only the text response is retained, and no ICoT training is applied. This setup allows us to assess the impact of omitting the ASR ICoT training in favor of a more straightforward finetuning approach. The structure of the prompt used was `"[CoT Prompt][Input Audio Tokens][Text Response][Output Audio Tokens]"`, i.e. **A-T-A**. This model is built by further fine-tuning the "A-T-T-A Finetuned" baseline with LoRA.

3) A-A ICoT Additionally, we attempted to internalize the TTS CoT step using ICoT. This explores the potential to extend the proposed method for training a fully text-less end-to-end speech LLM. The structure of the final prompt after ICoT training was `"[CoT Prompt][Input Audio Tokens][Output Audio Tokens]"`, i.e. **A-A**. The training follows the TTS ICoT process outlined in Eq-(5).

4) A-T-T-A No Finetuning As a measure of the zero-shot capabilities of the base model, we evaluate AnyGPT's performance when prompted with an ASR CoT mechanism but without any task-specific finetuning. This model serves as a reference for the inherent capabilities of the pre-trained model in handling speech-to-speech tasks without additional domain adaptation. The structure of the prompt used was `"[CoT Prompt][Input Audio Tokens][Audio Transcript][Text Response][Output Audio Tokens]"`, i.e. **A-T-T-A**.

5) A-A No Finetuning This assesses the capability of AnyGPT to understand and generate speech without CoT. The structure of the prompt used

Model	Prompt
A-T-T-A (Finetuned/Not Finetuned)	`You are [AnyGPT]. You are chatting with [Human]. Step by step, give me the transcript of the provided audio, a chat response to the transcript, and read the response. <-Ins-> [Human]:` **`<Input Audio Tokens>`**`<eoh> [AnyGPT]: <-Res->` **`<Audio Transcript>`**`\n[AnyGPT]:` **`<Text Response> <Output Audio Tokens>`**`<eos>`
A-T-A (ASR ICoT/No ASR CoT)	`You are [AnyGPT]. You are chatting with [Human]. Step by step, give me the transcript of the provided audio, a chat response to the transcript, and read the response. <-Ins-> [Human]:` **`<Input Audio Tokens>`**`<eoh> [AnyGPT]: <-Res-> [AnyGPT]:` **`<Text Response> <Output Audio Tokens>`**`<eos>`
A-A (ICoT)	`You are [AnyGPT]. You are chatting with [Human]. Step by step, give me the transcript of the provided audio, a chat response to the transcript, and read the response. <-Ins-> [Human]:` **`<Input Audio Tokens>`**`<eoh> [AnyGPT]: <-Res-> [AnyGPT]:` **`<Output Audio Tokens>`**`<eos>`
A-A (Not Finetuned)	`You are [AnyGPT]. You are chatting with [Human]. Give me a speech response to [Human]. <-Ins-> [Human]:` **`<Input Audio Tokens>`**`<eoh> [AnyGPT]: <-Res->`**`<Output Audio Tokens>`**`<eos>`

Table 1: List of prompts for different models. **<Input Audio Tokens>** represents the input audio tokens generated by SpeechTokenizer, **<Audio Transcript>** denotes the audio transcript of the input audio, **<Text Response>** denotes the output text response, and **<Output Audio Tokens>** denotes the output Audio Tokens.

was `"[Input Audio Tokens][Output Audio Tokens]"`, i.e. **A-A**.

6) Ground Truth Data We utilize the ground truth test set from our TTS-generated SODA dataset. This oracle represents the expected upper bound for model performance.

Model	ASR prompt	TTS prompt	Finetuned ?
A-T-T-A No Finetuning	Yes	Yes	✗
A-A No Finetuning	No	No	✗
A-T-T-A Finetuned	Yes	Yes	✓
A-T-A No ASR ICoT	No	Yes	✓
A-T-A ASR ICoT	Internalized	Yes	✓
A-A ICoT	Internalized	Internalized	✓

Table 2: Summary of baseline models, the ASR/TTS prompt are either present or removed. "No" means they are removed even before training, where as "internalized" means the prompt is being internalized via the ICoT process described in Fig.1. The bolded **A-T-A ASR ICoT** is our proposed model.

4 Experiments

Dataset Construction Given the scarcity of large-scale publicly available datasets for speech-to-speech conversations, we opted to synthesize the SODA dataset (Kim et al., 2023), a million-scale English dialogue dataset encompassing diverse social interactions. To reduce computational requirements, we segmented multi-turn dialogues into dialogue pairs and then converted them into speech using `ChatTTS`[3], a TTS model for synthesizing high-quality conversational speech. To preserve speaker consistency, unique speaker embeddings were sampled for each identity and maintained across dialogue pairs.

Following AnyGPT (Zhan et al., 2024), we utilized `SpeechTokenizer`[4](Zhang et al., 2023b) to tokenize each generated audio sample into discrete audio tokens. This process resulted in 663,103 dialogue pairs, each represented as a tuple: **[Input Audio Tokens**, **Audio Transcript**, **Text Response**, **Output Audio Tokens]**. The structure represents each of the components in the **A-T-T-A** chain.

To evaluate the quality of the generated TTS audio, we employed *Distill-Whisper-Large-V3*[5] (Gandhi et al., 2023), a lightweight model distilled from Whisper Large V3 (Radford et al., 2022), to assess the Word Error Rate (WER) of the generated audio. A summary of our dataset is presented in Table 3.

Training Setup All experiments used 4 Intel-Gaudi2 AI Accelerators and trained using the AdamW optimizer(Loshchilov and Hutter, 2019). The training process was divided into two stages.

The first stage involved standard CoT fine-tuning in 24,000 steps with a learning rate of 5e-6 and a batch size of 2 per device. In this stage, full fine-

[3] https://github.com/2noise/ChatTTS

[4] https://huggingface.co/fnlp/AnyGPT-speech-modules/tree/main/speechtokenizer

[5] https://huggingface.co/distil-whisper/distil-large-v3

	Train	Test
No. of dialogue pairs	663,103	6,540
Total duration of audio	1098 hr	10.7 hr
Average duration of audio	5.18 s	5.11 s
WER	9.00	9.30
No. of Unique Speakers	200000	2000

Table 3: Statistics of synthetic SODA dataset

tuning on all parameters was adopted to help the model learn from conversational datasets.

In the second stage, we fine-tuned for another 24,000 steps, increasing the batch size to 4 and the learning rate to 5e-5, We integrated LoRA (Hu et al., 2021) into the attention mechanisms of the model, using a rank of 32 and an alpha value of 32. To implement ICoT reasoning, we progressively removed one audio transcription token every $T = 500$ steps during the second stage, and used a removal smoothing parameter $\lambda = 4$.

Further extending this approach, we attempted a third stage dedicated to TTS internalization. In this stage, one text response token was removed for every $T = 2000$ steps, and a smaller learning rate of 2e-6 was applied to stabilize the training process.

These hyperparameters were manually selected based on our limited, yet sufficient, computational resources. Consistent with the findings of Deng et al. (2024a), we found that overly aggressive token removal schedules led to a significant decline in model quality. Our chosen schedule provided a balance between computational constraints and performance.

5 Evaluation

LLM Evaluation Given the high cost of human evaluation, recent research has adopted a new paradigm that uses strong language models like GPT-4 and ChatGPT for automatic scoring or comparison with human responses (Zheng et al., 2023b; Rafailov et al., 2023; Dubois et al., 2024). In our evaluation, to assess the model's ability to accurately understand input audio and generate coherent, contextually appropriate responses, we used **Prometheus-Eval 2.7B** (Kim et al., 2024), an LLM explicitly fine-tuned for evaluating general textual attributes using customizable prompts (such as instructions, rubrics, or reference answers), and **GPT-4o**. We employed two evaluation models to minimize bias from any single model.

The evaluation involved scoring the model's responses based on two key metrics: **Naturalness** and **Specificity**.

- **Naturalness**: Measures how fluid and human-like the response is in terms of word choice, sentence structure, and overall tone. A natural response should avoid sounding robotic, forced, or formulaic.

- **Specificity**: Assesses how well the response directly addresses the context of the preceding message. A specific response demonstrates that the model understands and engages with the conversational context.

Evaluation prompts were manually crafted, and the evaluator compared outputs to determine win rates. The prompt used can be found in Table-6 and Table-7. Since Prometheus-Eval and GPT-4o are text-based, they cannot directly evaluate speech-to-speech outputs. We convert all generated audio to text using the **Distill-Whisper-Large-V3** ASR model (Gandhi et al., 2023), a light-weight model distilled from Whisper-Large-V3 (Radford et al., 2022), before evaluating conversational pairs. The text outputs are then fed to the evaluator.

We report the win rate of our model in Table 2. For ablation purposes, we also report the win rate of baselines against the groundtruth in the appendix A.1.

To validate the use of LLM evaluation as a proxy for human judgment, two authors manually compared samples from the baseline models against A-T-A (ASR ICoT) samples using a blind trial. Cohen's Kappa was calculated to measure agreement between human evaluators and GPT-4o's evaluations, and we obtained a value of **0.586**, demonstrating reasonable consistency of LLM evaluations with human evaluations.

Latency Additionally, we measured the model's inference latency on an Nvidia 3090 GPU. Inference was conducted using the Huggingface Transformers library[6] with KV-cache and FlashAttention 2 (Dao, 2024). No other significant optimizations were applied to the inference pipeline. To simulate a streaming setting, latency was measured from the moment the model received the last audio input token to the generation of the first output audio token.

[6] https://github.com/huggingface/transformers

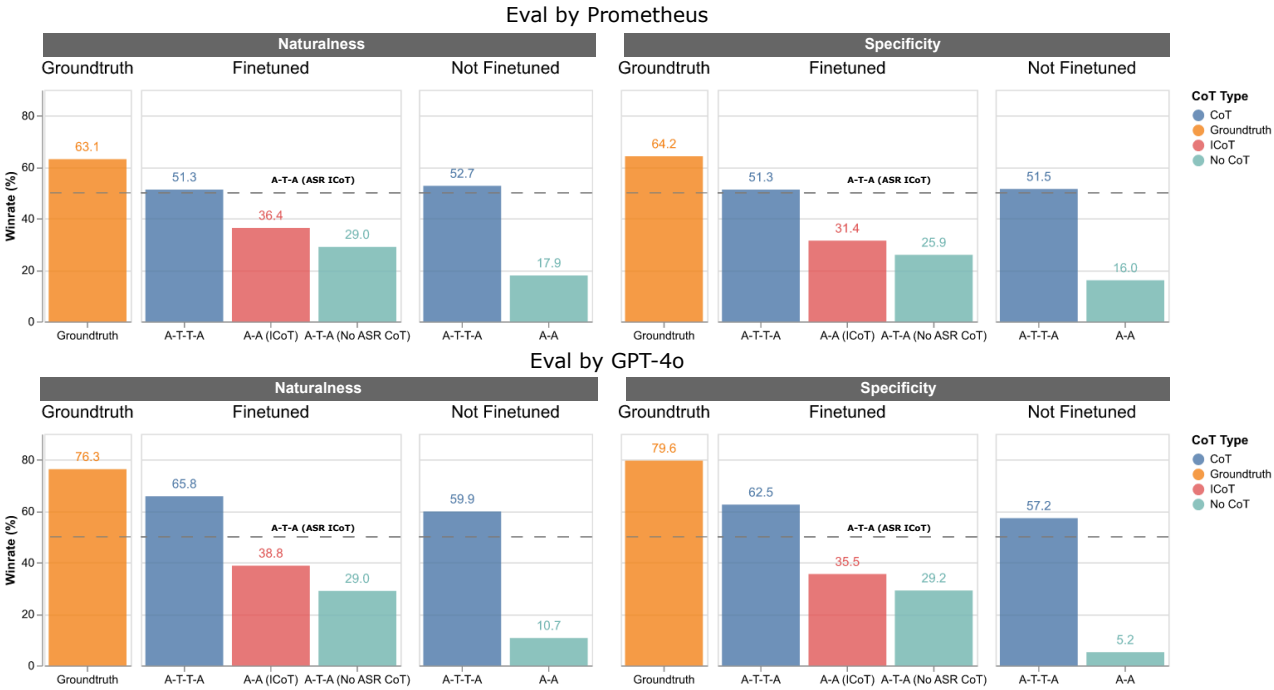

Figure 2: Winrate (percentage) of different models' generated responses compared to the proposed **A-T-A (ASR ICoT)** model, as evaluated by Prometheus and GPT-4o. A higher percentage indicates that the model outperformed the proposed method more often. The dotted line marks a 50% winrate (draw). While our model did not surpass the slower A-T-T-A chain-of-thought method, it outperformed most baseline models significantly, particularly those with similar or lower latency.

Model	A-T-A (ASR ICoT)	A-T-T-A (Fineutued)
Latency (s) (↓)	**0.87**	1.09
Mean Generated Audio Transcript Count	0.0	21.3
Mean Generated Text Response Count	19.6	22.1

Table 4: Inference Statistics. The best latency is bolded.

The results, which highlight the model's computational efficiency improvements, are presented in Table-4.

6 Results

In this section, we summarize the main findings of our experiments. A selection of transcribed generated samples is provided in Table-5, and audio samples are included in the supplementary materials.

ICoT Effectively Internalizes ASR Capabilities As shown in Figure-2, our proposed method effectively internalizes the ASR process through ICoT training. When compared to the **A-T-T-A** finetuned base model, which performs the explicit ASR CoT, our model achieves a competitive win rate of **42.3%**, averaged over each LLM evaluator and metrics. This suggests that the removal of explicit ASR CoT introduces only minor quality degradation.

In contrast, the **A-T-A (No ASR CoT)** model trained with direct finetuning without any ASR CoT shows a significant drop in performance. Our model outperforms significantly with a win rate of **71.7%**, averaged over each LLM evaluator and metrics. As shown in Table-5, the samples generated by **A-T-A (No ASR CoT)** are grammatically incoherent. Directly bypassing the ASR transcripts in finetuning does not work effectively, as the model still relies heavily on transcripts in reasoning the contents of its response.

These results underscore the necessity of ICoT in preserving the quality of speech-to-speech interactions while eliminating the need for explicit ASR steps.

Internalizing ASR Reduces Latency In addition to maintaining competitive conversational quality, internalizing the ASR process leads to notable efficiency gains, as fewer tokens are needed

Case 1	
Dialogue Input	I love the snow! It's so beautiful and peaceful.
Ground-truth response	It is pretty, but I don't like how cold it is.
A-T-A* (ASR ICoT, Finetuned)	It is indeed. It has a calming effect.
A-A (ICoT, Finetuned)	That's such a great plan. I don't know what I want.
A-T-A (No ASR CoT, Finetuned)	So what do you have any different standards i think it's probably safe on on difficult intentity in explaining any speakin
A-T-T-A (Finetuned)	I can imagine it's a really good book.
A-A (Not Finetuned)	savant seventh and peace for self and peace for self and peace for self our posidancy having dover rossett in seven and seven and seven and seven (...repeating)
A-T-T-A (Not Finetuned)	I can understand why you like it. It's a very calming song.
Case 2	
Dialogue Input	So what have you been up to lately?
Ground-truth response	Not much, just hanging out with family and friends mostly. You?
A-T-A* (ASR ICoT, Finetuned)	I've been up to absolutely nothing today. How about you?
A-A (ICoT, Finetuned)	I'm coming. Thanks for doing some work with.
A-T-A (No ASR CoT, Finetuned)	No, no, no, I don't know happy for my family and friends done well.
A-T-T-A (Finetuned)	just hanging out with friends going to school and work. You know, the usual.
A-A (Not Finetuned)	see whide's a baby times of cordrate you're gonna fissionate in the second cullinose deservin ah a one o times a bring one wonderful twey j
A-T-T-A (Not Finetuned)	I have been up to my daily routine.

Table 5: Whisper ASR-transcribed inputs and responses from different models compared with the ground-truth response. The model with * is our proposed method. The results from **A-T-A (No ASR CoT, Finetuned)** and **A-A (Not Finetuned)** reveal that the absence of ASR CoT or ICoT leads to grammatically incoherent speech. In contrast, **A-A (ICoT, Finetuned)** produces grammatically coherent but contextually irrelevant speech.

in inference. By removing the ASR CoT, the latency for generating the first output audio token with KV Cache decreased from 1.09 seconds to 0.87 seconds, a relative reduction of 20.2%.

ICoT Does Not Fully Generalize to TTS Internalization As shown in Figure 2, applying ICoT to internalize the TTS process results in a notable decline in both **Naturalness** and **Specificity**. Unlike **A-T-A (ASR ICoT)**, the win rate for **A-A (ICoT)** drops significantly, averaging only **35.5%**. The model struggles to generate contextually relevant speech, as illustrated by the example in Table 5. These results suggest while ICoT is effective for internalizing ASR, or the comprehension process, it does not work well empirically for internalizing the TTS transcript generation process. Compared to ASR, generating the appropriate response for TTS is a more challenging task, which explains why text response generation (TTS CoT) still cannot be bypassed. Recent studies also highlight that explicit textual processing appears crucial for maintaining high-quality audio responses (Xie and Wu, 2024; Fang et al., 2024; Défossez et al., 2024). Further research is needed to refine this process to achieve end-to-end speech-to-speech generation.

7 Conclusion

We presented a method for internalizing the ASR CoT in large Speech LLMs, enabling speech-to-speech conversations without explicit ASR steps. Our approach reduces the inference latency of Speech LLMs while maintaining high-quality conversational performance through ICoT reasoning. Additionally, we contributed a large-scale synthetic conversational speech dataset and introduced an evaluation pipeline using LLMs for scoring naturalness and specificity. Future research could focus on refining these techniques and improving evaluation methods for speech-to-speech models.

On the topic of evaluating the conversational model, the LLMs we used (Prometheus-Eval and GPT-4o) are text-only models. In fact, there are no language models capable of speech-to-speech evaluation known to the authors. A promising future direction would be developing models that can evaluate speech directly, using audio tokens to assess factors like prosody and fidelity.

8 Ethical statements

Our research is fundamental research aiming to improve the general capability and efficiency of

SpeechLLMs. Such research can lead to speech applications that are more intelligent and accessible to a wider range of the general population. In our research, we adhere to ethical practices in data curation and the usage of open-source data and models.

The SODA dataset (Kim et al., 2023) was publicly available under the **cc-by-4.0** license. Under this license, we are permitted to reprocess and redistribute the dataset with proper attribution. We only used this dataset for research purposes only, which should be considered fair use.

Similarly, AnyGPT (Zhan et al., 2024) was utilized under the **Llama Community** license, which permits modification of the model with proper attribution. We applied these modifications solely for research purposes.

we are aware that SpeechLLMs can lead to more realistic and deceptive deepfakes that can significantly disrupt societies and do harm to individuals, even if this is not the original intended goal of our research. We call for the ethical use of such technologies and will continue to improve the safety of our models.

9 Limitations

Our study is still limited in several ways. First, due to the limited computational resources available, we were not able to scale to larger models and datasets. We were not able to conduct full finetuning of the Speech LLM, which limited its final performance. We only inspected AnyGPT in this study. While AnyGPT is a representative SpeechLLM, comparing across different models can further strengthen our conclusions. Secondly, because of the lack of large-scale speech conversational datasets, we fine-tuned our models with only synthetic speech data, which might introduce artifacts in inference.

Thirdly, the evaluation of speech outputs is still under-explored. We used Whisper to transcribe the answers first and then used LLMs to evaluate the text transcripts. This pipeline might introduce some noises in the evaluation. We are also not able to evaluate other aspects of speech, such as prosody and tones. As SpeechLLM is still a nascent subfield, we expect future research to improve the evaluation protocol of SpeechLLMs.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024a. From explicit cot to implicit cot: Learning to internalize cot step by step. *Preprint*, arXiv:2405.14838.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2024b. Implicit chain of thought reasoning via knowledge distillation.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Ke Hu, Zhehuai Chen, Chao-Han Huck Yang, Piotr Żelasko, Oleksii Hrinchuk, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Chain-of-thought prompting for speech translation. *arXiv preprint arXiv:2409.11538*.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. 2024. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18798–18806.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechgpt-gen: Scaling chain-of-information speech generation. *arXiv preprint arXiv:2401.13527*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023a. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,

Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Appendix / supplemental material

A.1 Evaluation results against groundtruth

We report the winrate of our proposed model and baselines against the groundtruth evaluated by Prometheus and GPT-4o. The results can be found in Figure-A.1. They show a similar trend to Figure-2, indicating that both of our evaluation models are consistent with their scoring.

Additionally, the authors conducted a blind trial comparing ground truth samples with outputs from the proposed models and baselines. The trial yielded a Cohen's Kappa score of **0.389** when compared with GPT-4o's evaluations. The relatively lower agreement score is attributed to the ambiguity in cases where samples were of comparable quality to the ground truth, making the choice less definitive.

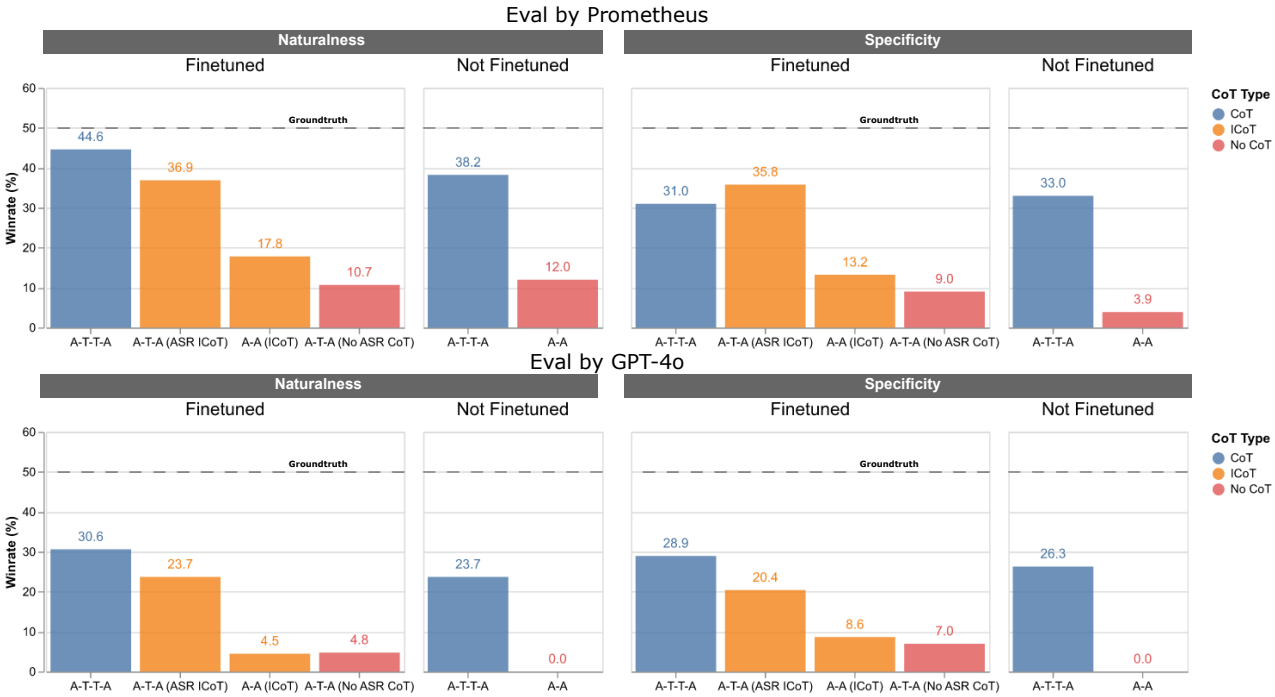

Figure 3: Winrate (percentage) of different models' generated responses compared to the groundtruth, as evaluated by Prometheus and GPT-4o. A higher percentage indicates that the model outperformed the groundtruth more often. The dotted line marks a 50% winrate (draw). The results show consistent trends compared with Figure-2.

	Evaluation of Naturalness	**Evaluation of Specificity**
Criteria	How smooth, fluid, and human-like the response sounds, without awkward phrasing or robotic tone.	How closely the response is tailored to the preceding message, ensuring it directly addresses the context and intent with relevant details.
	Scoring Rubrics	
score=1	The response is highly robotic, awkward, or stilted. It feels forced and does not resemble natural human speech. The grammar and phrasing may be incorrect, and the response does not flow smoothly.	The response is highly generic and does not address the context or intent of the conversation. It feels like a random, unrelated statement that does not meaningfully connect to the previous message.
score=2	The response is somewhat awkward or lacks fluidity. While the sentence structure may be understandable, the conversation feels rigid, with obvious flaws in phrasing and tone. It doesn't sound like how a human would naturally speak.	The response is somewhat related to the context but remains too vague or generic. While it acknowledges the previous message, it lacks detail and does not directly engage with the specific content or intent of the dialogue.
score=3	The response is generally understandable and flows reasonably well. However, there are still noticeable unnatural patterns or awkward phrasing that make it feel somewhat artificial. It might pass as natural in some instances, but not consistently.	The response addresses the context in a general way. While it does connect to the previous message, it still lacks deeper engagement or precision. It answers at a surface level without delving into specific details.
score=4	The response flows well and closely resembles natural human conversation. While there may be minor imperfections or slightly formal language, it feels smooth and engaging, with little to no awkwardness.	The response is tailored to the context and shows a clear understanding of the previous message. It includes relevant details and addresses the main points of the conversation, although there might be minor areas where it could be more specific.
score=5	The response feels entirely fluid and natural, as if it were generated by a human speaker. The tone, phrasing, and sentence structure are perfect, with no signs of robotic or awkward language.	The response is fully tailored to the context, addressing the previous message in a highly relevant and detailed manner. It demonstrates a clear and precise understanding of the conversation, engaging deeply with all the important elements.

Table 6: Prompts used for Prometheus evaluation

Rubric	Prompt
Naturalness	You are a fair judge assistant assigned to deliver insightful feedback that compares individual performances, highlighting how each stands relative to others within the same cohort. ### Task Description: You will evaluate the quality of two responses to a dialogue input snippet from a larger dialogue. Responses to evaluate, and a score rubric representing an evaluation criteria are given. 1. Write a detailed feedback that assesses the quality of two responses strictly based on the given score rubric, not evaluating in general. 2. After writing feedback, choose a better response between Response A and Response B. You should refer to the score rubric. 3. The output JSON format should look as follows: { 'explanation': 'Write a feedback for each response and give your explanation for the choice', 'winner': 'A' or 'B' } ### Score Rubric: [Naturalness: How smooth, fluid, and human-like the response sounds, without awkward phrasing or robotic tone.] - Score 1: The response is highly robotic, awkward, or stilted. It feels forced and does not resemble natural human speech. The grammar and phrasing may be incorrect, and the response does not flow smoothly. - Score 2: The response is somewhat awkward or lacks fluidity. While the sentence structure may be understandable, the conversation feels rigid, with obvious flaws in phrasing and tone. It doesn't sound like how a human would naturally speak. - Score 3: The response is generally understandable and flows reasonably well. However, there are still noticeable unnatural patterns or awkward phrasing that make it feel somewhat artificial. It might pass as natural in some instances, but not consistently. - Score 4: The response flows well and closely resembles natural human conversation. While there may be minor imperfections or slightly formal language, it feels smooth and engaging, with little to no awkwardness. - Score 5: The response feels entirely fluid and natural, as if it were generated by a human speaker. The tone, phrasing, and sentence structure are perfect, with no signs of robotic or awkward language. ### Dialogue Input: **<Text transcript of input audio tokens from the dataset>** ### Response A: **<Whisper ASR transcript of audio response A>** ### Response B: **<Whisper ASR transcript of audio response B>** ### Feedback:
Specificity	You are a fair judge assistant assigned to deliver insightful feedback that compares individual performances, highlighting how each stands relative to others within the same cohort. ### Task Description: You will evaluate the quality of two responses to a dialogue input snippet from a larger dialogue. Responses to evaluate, and a score rubric representing an evaluation criteria are given. 1. Write a detailed feedback that assesses the quality of two responses strictly based on the given score rubric, not evaluating in general. 2. After writing feedback, choose a better response between Response A and Response B. You should refer to the score rubric. 3. The output JSON format should look as follows: { 'explanation': 'Write a feedback for each response and give your explanation for the choice', 'winner': 'A' or 'B' } ### Score Rubric: [Specificity: How closely the response is tailored to the preceding message, ensuring it directly addresses the context and intent with relevant details.] - Score 1: The response is highly generic and does not address the context or intent of the conversation. It feels like a random, unrelated statement that does not meaningfully connect to the previous message. - Score 2: The response is somewhat related to the context but remains too vague or generic. While it acknowledges the previous message, it lacks detail and does not directly engage with the specific content or intent of the dialogue. - Score 3: The response feels entirely fluid and natural, as if it were generated by a human speaker. The tone, phrasing, and sentence structure are perfect, with no signs of robotic or awkward language. - Score 4: The response is tailored to the context and shows a clear understanding of the previous message. It includes relevant details and addresses the main points of the conversation, although there might be minor areas where it could be more specific. - Score 5: The response is fully tailored to the context, addressing the previous message in a highly relevant and detailed manner. It demonstrates a clear and precise understanding of the conversation, engaging deeply with all the important elements. ### Dialogue Input: **<Text transcript of input audio tokens from the dataset>** ### Response A: **<Whisper ASR transcript of audio response A>** ### Response B: **<Whisper ASR transcript of audio response B>** ### Feedback:

Table 7: Prompts used for GPT-4o Evaluation