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

Robin Shing-Hei Yuen
University of British Columbia
robinysh@student.ubc.ca

Timothy Tin-Long Tse
University of British Columbia
ttse05@student.ubc.ca

Jian Zhu
University of British Columbia
jian.zhu@ubc.ca

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 models 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. 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, there has been limited exploration of their ability to handle broader conversational reasoning tasks that text-based LLMs excel at. Bridging this gap is critical for developing speech models that can engage in intelligent dialogue, without relying on intermediate text representations.

In this work, we propose a novel approach that leverages the ASR and TTS capabilities of Speech LLMs to enable natural speech conversations. 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 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 14.5%, and moving closer to a fully text-free end-to-end speech LLM.
- We constructed a large-scale synthetic conversational speech dataset with an emphasis on social common sense reasoning, containing $\sim$660k dialogue exchange pairs totaling $\sim$1000 hours of speech data. The dataset is publicly available on Huggingface Hub [1].

[1] https://huggingface.co/datasets/robinysh/soda_tts

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

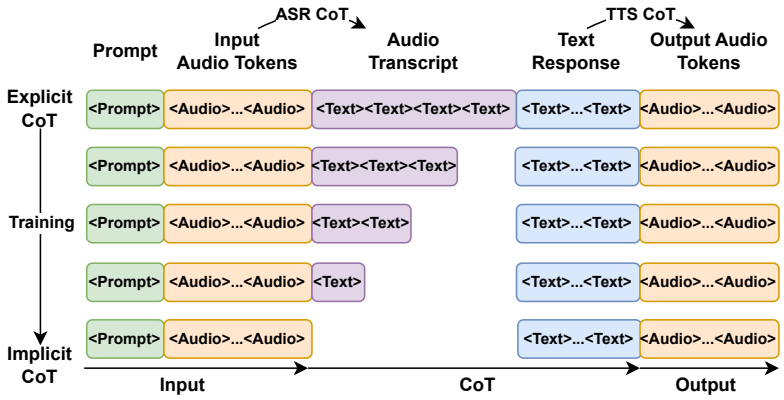

Figure 1: Illustration of ICoT training and generation structure (from A-T-T-A to A-T-A ASR ICoT). Tokens of audio transcripts are removed linearly from the start during training, compressing the generation length for faster inference.

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] and Qwen-Audio series [Chu et al., 2023, 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.

Implicit Chain of Thought 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., 2023, Mondal et al., 2024, Zhang et al., 2024b]. Yet the application of CoT prompting in the audio and speech domain is still relatively rare [Li 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 effectiveness of this approach was demonstrated in solving math problems.

3 Method

We incorporate ICoT [Deng et al., 2024b,a] to internalize ASR capability into a pre-trained Speech LLM, such that it performs speech-to-speech conversation without explicit ASR steps. To our knowledge, this is the first use of 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. 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], inheriting the strong zero-shot generalization capabilities that are characteristic of LLMs. This makes AnyGPT well-suited for our speech conversational tasks.

Speech-to-Speech Conversation via ASR and TTS CoT Prompting Leveraging the ASR and TTS capability of AnyGPT, speech-to-speech conversations can be implemented using a multi-step CoT strategy with intermediate text interfaces [Zhang et al., 2023a]. The conversational process is structured as follows: the model first transcribes the input audio via ASR, then generates a textual response, and finally converts the text 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-6. However, this approach introduces two additional intermediate steps: the transcription of ASR output and the generation of a text-based response, both introduce increased inference latency and computation overhead.

ASR Internalization via Implicit CoT (A-T-A ASR ICoT) To mitigate the latency caused by intermediate text generation, we adopt ICoT reasoning 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**). Figure-1 illustrates this approach. Training consists of two stages. **1)** Full-parameter fine-tuning of the model using standard CoT to ensure alignment with our prompts and dataset. **2)** ICoT training to internalize ASR CoT with LoRA [Hu et al., 2021].

We define input audio tokens as x, intermediate audio transcript as $z^{ASR} = z_1^{ASR}, z_2^{ASR}, \cdots, z_m^{ASR}$, intermediate text response as $z^{TTS} = z_1^{TTS}, z_2^{TTS}, \cdots, z_n^{TTS}$, and final output audio tokens as y. The language model with parameters θ is initially trained using a standard next-token prediction objective with the A-T-T-A format. After this 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:

$$s(t) = \min\left(\left\lfloor \frac{t}{T} + o \right\rfloor, K_i\right) \tag{1}$$

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. The new objective function becomes:

$$\min_{\theta} -\log P_\theta(y, z_{s(t):m}^{ASR}, z_{1:n}^{TTS} \mid x) \tag{2}$$

Training continues until all audio transcript tokens are removed. To ensure stability during training, we also employed the optimizer reset strategy from Deng et al. [2024a].

Baselines To provide a meaningful comparison of our model's performance, we evaluate against the following baseline approaches. The exact prompt of the baselines can be found in Table-6, and a summary of the differences between the baselines can be found in Table-1. **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. **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. **3) A-A ICoT** We attempted to internalize the TTS CoT step using ICoT. **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. **5) A-A No Finetuning** This assesses the capability of AnyGPT to understand and generate speech without CoT. **6) Ground Truth Data** We utilize the ground truth test set from our TTS-generated SODA dataset.

4 Experiments

Dataset Construction Due to limited large-scale, publicly available speech-to-speech conversation datasets, we synthesized TTS audio with SODA [Kim et al., 2023], a million-scale English dialogue dataset encompassing diverse social interactions. Multi-turn dialogues were segmented into dialogue pairs and then converted into speech using `ChatTTS`[2], 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.

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

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 1: Summary of baseline models. "No" means the ASR/TTS prompt is removed even before training, whereas "internalized" means the prompt is being internalized via the ICoT process described in Fig.1. Our proposed model is **bolded**.

	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 2: Statistics of synthetic SODA dataset

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 3: Inference Statistics. The best latency is bolded.

Following AnyGPT [Zhan et al., 2024], we utilized `SpeechTokenizer`[3][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]**. To verify audio quality, we employed *Distill-Whisper-Large-V3*[4] [Gandhi et al., 2023] to calculate the Word Error Rate (WER). A summary of our dataset is presented in Table 2.

Training Setup All experiments used 4 Intel-Gaudi2 AI Accelerators and trained using AdamW [Loshchilov and Hutter, 2019]. Training was conducted in two stages. The first stage involved standard CoT fine-tuning over 24,000 steps with a learning rate of 5e-6 and a batch size of 2 per device. For the second stage, we used LoRA [Hu et al., 2021] for another 24,000 steps, increasing the batch size to 4 and the learning rate to 5e-5. The LoRA was integrated into the attention mechanisms, 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, with removal smoothing parameter $\lambda = 4$. Additionally, a third stage was attempted for TTS ICoT, in which we removed one text response token every $T = 2000$ steps, and applied a smaller learning rate of 2e-6 to ensure stability.

Evaluation We evaluate the model's ability to accurately understand input audio and generate coherent, contextually appropriate responses using **Prometheus-Eval 2.7B** [Kim et al., 2024], an LLM explicitly fine-tuned for evaluating text attributes with customizable metrics, alongside with **GPT-4o**. We employed two evaluation models to minimize bias from any single model. Responses were scored on two metrics: **Naturalness** (fluency and human-likeness), and **Specificity** (relevance and contextual alignment). Evaluation prompts were manually crafted, and the evaluator compared outputs to determine win rates. The prompts can be found in Table-4 and Table-5. Since Prometheus-Eval and GPT-4o are text-based, we transcribed all generated audio via **Distill-Whisper-Large-V3**, before feeding into the evaluators. Win rates against our model are shown in Fig 2, while comparisons against the ground truth provided in the appendix A.1.

To validate the use of LLM evaluation as a proxy for human judgment, two authors blind-tested baseline models against A-T-A (ASR ICoT) samples. The Cohens Kappa between human evaluators and GPT-4o's evaluations was **0.586**, demonstrating reasonable consistency between LLM evaluations and human evaluations.

Latency We measured the inference latency on an Nvidia 3090 GPU using Huggingface Transformers with KV-cache and FlashAttention 2 Dao [2024]. To simulate a streaming setting, latency was measured from when the model received the last audio input token to the generation of the first output audio token. The results are presented in Table-3.

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

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

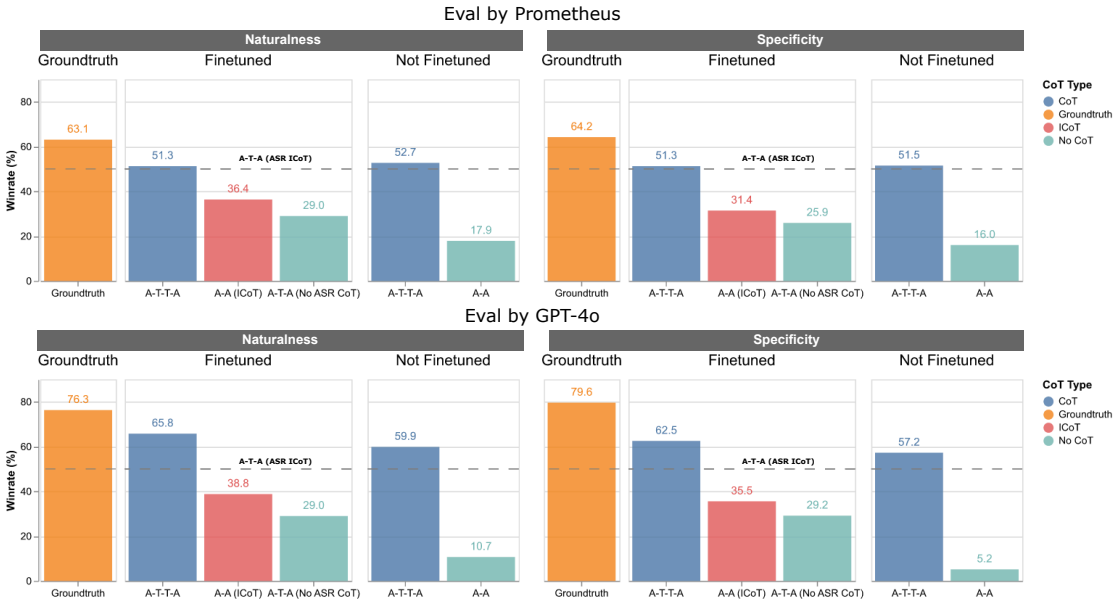

Figure 2: Winrate (percentage) of different models generated responses compared to the proposed **A-T-A (ASR ICoT)** model, as evaluated by Prometheus and GPT-4o. A higher percentage indicates that the model outperformed the proposed method more often. The dotted line marks a 50% winrate (draw). While 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.

5 Results

ICoT Effectively Internalizes ASR Capabilities As shown in Figure-2, our ICoT-trained model internalizes ASR effectively. When compared to the **A-T-T-A** finetuned model with explicit ASR CoT, our model achieves a competitive win rate of **42.3%**, averaged over each LLM evaluator and metrics. This suggests that internalizing 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-7, the samples generated by **A-T-A (No ASR CoT)** are grammatically incoherent. 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 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 7. These results suggest that while ICoT is effective for ASR, further research is needed to refine its application for TTS, where explicit textual processing appears crucial for maintaining high-quality audio responses Xie and Wu [2024], Fang et al. [2024], Défossez et al. [2024].

6 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.

Acknowledgments and Disclosure of Funding

This research was enabled in part by support provided by Advanced Research Computing at the University of British Columbia, the Digital Research Alliance of Canada, LAION and Intel through LAION's BUD-E project. We acknowledge the support of the Discovery Program from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 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, 2022.

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

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

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

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

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step, 2024a. URL https://arxiv.org/abs/2405.14838.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation, 2024b. URL https://openreview.net/forum?id=9cumTvvlHG.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2024. URL https://arxiv.org/abs/2409.06666.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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. SODA: Million-scale dialogue distillation with social commonsense contextualization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.799. URL https://aclanthology.org/2023.emnlp-main.799.

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

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

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

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

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

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

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

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL `https://arxiv.org/abs/2408.16725`.

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

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

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

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

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

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

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.

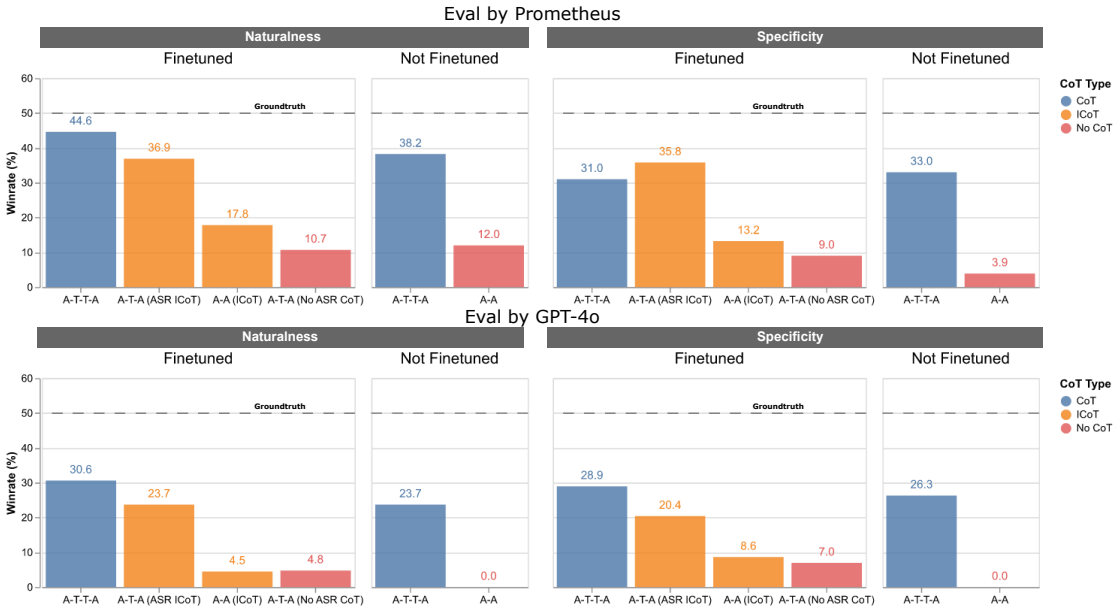

Figure 3: Winrate (percentage) of different models generated responses compared to the groundtruth, as evaluated by Prometheus and GPT-4o. A higher percentage indicates that the model outperformed the groundtruth more often. The dotted line marks a 50% winrate (draw). The results show consistent trends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

A.2 Prompts used in evaluation

Table 4 contains the evaluation criteria that we used to prompt Prometheus [Kim et al., 2024]. Table 5 shows the prompts for the GPT-4o evaluation.

	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 4: 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 5: Prompts used for GPT-4o Evaluation

A.3 Prompts for training and inference

Table 6 contains the prompt used for training and inferencing each of the baselines and proposed model.

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 6: 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.

A.4 ASR Transcribed model output samples

Several samples generated by the baselines and our proposed model were transcribed by **Distill-Whisper-Large-V3**[Gandhi et al., 2023] and presented in Table 7.

	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 7: 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.