PodAgent: A Comprehensive Framework for Podcast Generation

Yujia Xiao[1], Lei He[2], Haohan Guo[1], Fenglong Xie[3], Tan Lee[1]

[1]The Chinese University of Hong Kong, [2]Microsoft, [3]Xiaohongshu Inc.

{yujiaxiao@link.cuhk.edu.hk, helei@microsoft.com}

{hguo@se.cuhk.edu.hk, fenglongxie@xiaohongshu.com, tanlee@ee.cuhk.edu.hk}

Abstract

Existing automatic audio generation methods struggle to generate podcast-like audio programs effectively. The key challenges lie in in-depth content generation, appropriate and expressive voice production. This paper proposed PodAgent, a comprehensive framework for creating audio programs. PodAgent 1) generates informative topic-discussion content by designing a Host-Guest-Writer multi-agent collaboration system, 2) builds a voice pool for suitable voice-role matching and 3) utilizes LLM-enhanced speech synthesis method to generate expressive conversational speech. Given the absence of standardized evaluation criteria for podcast-like audio generation, we developed comprehensive assessment guidelines to effectively evaluate the model's performance. Experimental results demonstrate PodAgent's effectiveness, significantly surpassing direct GPT-4 generation in topic-discussion dialogue content, achieving an 87.4% voice-matching accuracy, and producing more expressive speech through LLM-guided synthesis. Demo page: https://podcast-agent.github.io/demo/. Source code: https://github.com/yujxx/PodAgent.

1 Introduction

Audio programs are an important channel for information acquisition. Compared to video or text media, audio can free your eyes and hands, allowing you to access information conveniently in a variety of scenarios. Podcasts are digital audio programs that are available for streaming or downloading over the Internet. As shown in Figure 1 (Above), the content of a podcast typically consists of viewpoints shared by various individuals. To accommodate diverse interests, podcasts often explore a wide range of topics, including economics, culture, psychology, and more. However, many content creators still face the complex process of transforming creative ideas into a final product. Additionally, providing strong, well-founded viewpoints and producing high-quality podcast-like audio on any given topic remains a significant challenge.

Recent advancements in generative models have made it possible to automatically create high-quality content. Large language models (LLMs) (Ouyang et al., 2022; Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023; Anthropic, 2023b) have achieved breakthrough capabilities in generating coherent and contextually appropriate text. In addition, foundation models for other modalities, such as vision (Blattmann et al., 2023; Midjourney, 2023; Brooks et al., 2024) and audio (Borsos et al., 2023; Wang et al., 2023a; Anthropic, 2023a; Copet et al., 2024), have made remarkable strides in the creation of multimodal content.

While existing models can generate podcast-like content, they have not yet achieved the creation of complete, professionally structured podcast episodes. For instance, audio-enhanced multimodal LLMs (Wu et al., 2023b; Huang et al., 2023b; Zhan et al., 2024) primarily focus on enabling multimodal interactions, but these interactions are typically constrained by short context windows and limited reasoning capabilities. Text-to-Audio (TTA) models (Kreuk et al., 2022; Liu et al., 2023a, 2024; Huang et al., 2023a) can generate various audio types, like speech, sound effects, and music. However, since these models prioritize general audio synthesis, they are inherently limited in producing coherent and intelligent spoken content. Although zero-shot Text-to-Speech (TTS) models (Casanova et al., 2022; Wang et al., 2023a; Tan et al., 2024) can generate high-quality speech for any speaker, they rely on the provided text and lack the ability to generate long-form informative content.

A straightforward approach is to combine the strengths of these models—using LLMs to generate rich text, TTS models to convert it into spoken content, and TTA models to add appropriate sound effects and background music to produce complete,

Findings of the Association for Computational Linguistics: ACL 2025, pages 23923–23937
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

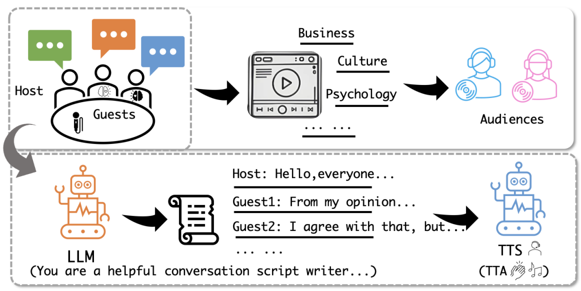

Figure 1: An overview of human-made Podcasts (Above) and the Generative models based PodAgent (Below): LLMs / TTS / TTA are used to generate conversation scripts, speech, sound effect and music.

informative, and professionally structured podcast episodes (Figure 1 Below). This approach naturally aligns with the emerging paradigm of AI agents (Wu et al., 2023a; LangChain, 2023). Empowered by LLMs, various AI agents (Xie et al., 2024; Du et al., 2023; Lu et al., 2024) are created to coordinate multiple AI tools to accomplish complex tasks through perception, decision-making, and action execution. A notable implementation of this approach is WavJourney (Liu et al., 2023b), which leverages LLMs to generate an audio script that connects various models for audio generation. While WavJourney represents a significant step forward with its extensive audio generation workflow, its current implementation still faces challenges in producing complete and intellectually rich content (An example demonstrated in Table 3).

Through observation and analysis of existing automated audio program creation systems, we identify four critical challenges as follows: **Content Depth and Insight Generation**. For a given topic, how to automatically generate rich and insightful viewpoints and provide meaningful analysis? **Natural Dialogue Generation**. How to create engaging conversational content that flows naturally between speakers, maintaining coherence while avoiding repetition? **Appropriate Voice Representation**. How to match suitable voice characteristics with different content and roles, ensuring consistency and authenticity in the audio presentation? **Speech Quality and Expressiveness**. How to generate robust long-form speech with appropriate prosody and emotion that matches the content's intent and maintains listener engagement?

In this work, we present PodAgent, a fully automated and comprehensive framework for creating content-rich and professionally structured audio programs. To tackle the aforementioned challenges, we:

- Design a Host-Guest-Writer system that generates engaging and coherent conversation scripts with diverse, insightful viewpoints from various backgrounds and perspectives for any given topic.
- Build a preset voice pool through comprehensive voice characteristic analysis to enable dynamic role-voice matching that aligns with speaker personalities and content context.
- Involve LLM-predicted speaking style in instruction-following TTS model to obtain high-quality speech output with appropriate prosody and emotion.
- Establish comprehensive evaluation metrics for podcast-like audio generation tasks, including assessments of open-ended topic discussions, voice matching, and voice quality.

2 Related Works

2.1 LLM-powered Agents

LLMs demonstrate remarkable capabilities in emulating human problem-solving through role-specific configurations (Wei et al., 2022; Yao et al., 2024; Shinn et al., 2023). Building upon this foundation, multi-agent systems incorporate LLMs with diverse role specifications to collaboratively address more complex challenges (Liang et al., 2023; Talebirad and Nadiri, 2023; Chan et al., 2023; Park et al., 2024). Within this framework, each agent functions as a domain expert, focusing on specialized areas and contributing unique perspectives. Furthermore, the integration of multi-modal foundation models has greatly enhanced agents' proficiency in handling cross-modal tasks (Huang et al., 2023b; Zhang et al., 2023; Hurst et al., 2024). It is crucial to explore effective problem decomposition strategies and appropriate tool utilization for solving real-world problems.

2.2 Voice Characteristic Analysis

Voice characteristic analysis is essential in our task for effectively assigning suitable voices to speakers in the audio program. This analysis also known as speech captioning, traditionally relies on approaches that classify and recognize predefined categories from speech signals (Issa et al., 2020).

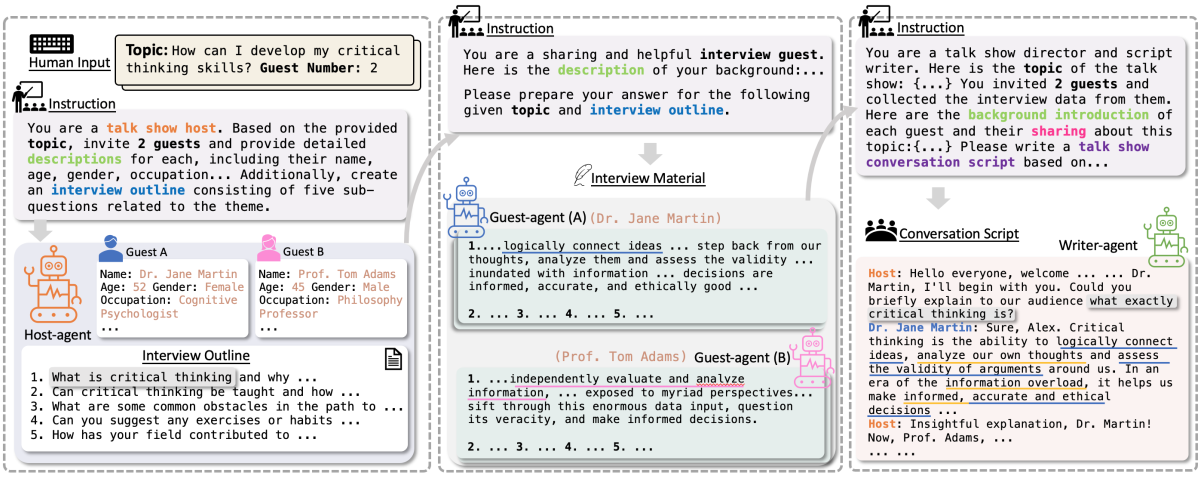

Figure 2: The Workflow of Host-Guest-Writer System. **Left:** The Host-agent generates guest information and an interview outline based on the given topic and number of guests. **Middle:** Guest-agents respond to the interview outlines, offering specialized perspectives aligned with their assigned roles. **Right:** The Writer-agent compiles a complete and coherent conversation script using the gathered interview material.

To address the limitations of insufficient predefined classes, recent studies (Yamauchi et al., 2024; Xu et al., 2023; Zhu et al., 2024) utilize self-supervised learning models for speech feature extraction and description generation. In our analysis, we employ the SpeechCraft (Jin et al., 2024), an open-source speech dataset with fine-grained text descriptions.

2.3 Text-to-Speech synthesis

Recent advances in zero-shot speech synthesis (Casanova et al., 2022; Wang et al., 2023a; Tan et al., 2024) enable voice mimicry from a short utterance of a reference speaker. To enhance style control, instruction-following TTS models (Yang et al., 2024; Guo et al., 2023) bridge textual descriptions with speaking styles. With the growth of the open-source community, an increasing number of outstanding TTS foundation model projects have been released. For instance, Bark (Anthropic, 2023a) is an TTS+ model that extends conventional speech synthesis to include nonverbal cues like laughter, sighs, and crying. CosyVoice (Du et al., 2024a,b) is one of the most recent open-source TTS foundation models, supporting various speech generation tasks like zero-shot voice cloning, multilingual speaking and instruction following.

3 PodAgent

This section introduces PodAgent, a fully automated comprehensive framework for creating informative and professionally structured audio programs. The focus will be on the key contributions of this work: 1) Host-Guest-Writer system for generating conversation scripts 2) Voice-role matching for selecting suitable voices, and 3) speech synthesis enhanced by LLM-predicted instruction.

3.1 Host-Guest-Writer system

We propose a novel Host-Guest-Writer multi-agent system to generate comprehensive and engaging conversational scripts for audio programs, with the workflow presented in Figure 2. In real-world talk-show format programs, hosts typically invite several experts to share insights based on their specialized knowledge of the topic. Inspired by this, the first task of our Host-agent is to formulate appropriate guest profiles. Once established, these profiles are assigned to different Guest-agents, enabling them to provide expertise-based responses.

Rather than implementing computationally intensive turn-by-turn dialogues between agents, which would require managing complex conversation history and turn-taking mechanisms, we adopt a more efficient parallel approach: the Host-agent creates a structured interview outline that serves as a common framework for all Guest-agents to address simultaneously. This allows each Guest-agent to respond to identical questions while maintaining their unique perspectives. Subsequently, all guest responses are processed by a dedicated Writer-agent, which synthesizes the inputs into a cohesive and natural conversational script, effectively eliminating redundancy while preserving the distinct viewpoints of each participant.

In summary, this collaborative framework or-

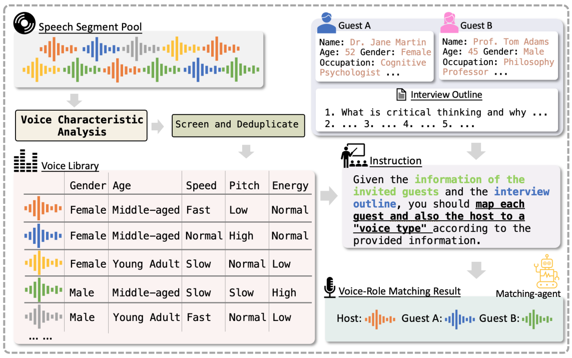

Figure 3: Voice-Role Matching. **Left:** Construction of a voice library through characteristic analysis of diverse speech segments. **Right:** The Matching-agent performs voice-role pairing using the voice library, guest profiles, and interview structure.

chestrates interactions among three specialized agents: the Host-agent, which guides the conversation flow and maintains topic coherence; the Guest-agent, which provides domain expertise and diverse perspectives; and the Writer-agent, which structures and refines the dialogue to ensure natural progression and professional presentation standards.

3.2 Voice-Role matching

After obtaining the conversation script, the next crucial step is to select the appropriate voice for each speaker. This voice-role matching process is critical for creating a natural and immersive listening experience for the audience. Figure 3 demonstrates the voice-role matching process. The first step is to collect speech samples from different speakers as much as possible. Then, the collected speech samples will be analyzed to extract voice characteristics for profiling. After that, we screen the profiled speech samples and de-duplicate segments with similar features, ultimately creating a comprehensive and non-redundant voice library. Details of the data can be found in Appendix A .

The curated voice library will be provided to the Matching agent, along with the guest information and interview outline generated by the Host-agent. The Matching-agent then leverages all the information to make informed and effective voice-role pairings. This process ensures that the selected voices align naturally with each speaker's designated role and expertise, enhancing the authenticity and engagement of the final audio program.

3.3 Instruction-following speech synthesis

To enhance the expressiveness of generated speech, we leverage LLM-predicted speaking styles as instructions to guide the synthesis. As depicted in the top right of Figure 4, the instruction-following TTS system takes three inputs: the text (content to be spoken), a reference voice (speech segment), and an instruction (speaking style). The system then generates speech in the voice of the reference speaker, adhering to the specified speaking style.

4 Experimental Setups

Our experiments will center around the key contributions of this work: topic-based discussion content generation, voice-role matching, and expressive speech synthesis. We follow WavJourney's setup for generating music and sound effects but use a more recent open-source framework, CosyVoice2, for speech generation.

4.1 Datasets

We base our evaluation on a subset of data from (Chiang et al., 2023), which originally contains 80 questions across 8 categories. To align with the topic-discussion scenario, we exclude categories such as "code" and "match" that are less suitable, and ultimately select 4 categories, Generic, Knowledge, Common-sense, and Counterfactual. This resulted in 40 topics, with 10 topics per category, serving as our experimental data. The experimental data is English-based. In addition to it, we also showcase some Chinese-based podcasts on the demo page to provide a broader perspective on PodAgent's capabilities.

4.2 Evaluation on conversation scripts

The dialogue content in podcast programs typically revolves around a given theme, which can vary widely to cater to different audience interests. The discussions typically showcase participants' unique perspectives and personal insights, offering listeners a rich tapestry of viewpoints and thought-provoking ideas. Given this subjective and open-ended nature of podcast content, establishing definitive ground truth or applying standardized quality metrics becomes particularly challenging.

To address this, we design the evaluation methods from two aspects. First, We employ several **quantitative metrics** that measure the lexical diversity, semantic richness and information density of the generated content. These metrics operate

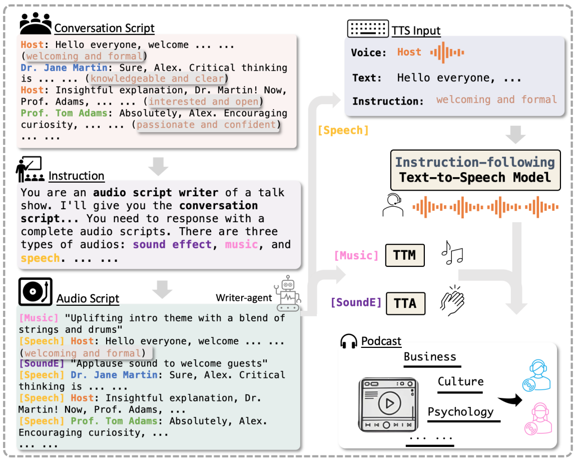

Figure 4: From Conversation Script to Podcast. **Audio Script Generation:** The Writer-agent create the audio script by enriching the conversation script with sound effect and music. **Instruction-following TTS:** Speaking styles are generated along with the conversation script, which can be used as instruction to guide the expressive speech synthesis. **Audio Production:** The generated audio segments are combined to create the final podcast.

independently of any reference or ground truth text, focusing solely on the characteristics of the text itself. Second, we utilize **LLM-as-a-Judge** methodology to perform comparative quality assessments between discussion texts. The specific implementation details of both approaches are outlined below.

- **Distinct-N** (Li et al., 2015), particularly distinct-1 and 2, is used to evaluate the text diversity. It emphasizes the count of distinct n-grams within the text, thereby penalizing those that contain many repeated words. To ensure comparability between texts of varying lengths, we employ a sliding window for normalization of scores. The window size is set as 100 and similar normalization are applied to other quantitative metrics.

$$\text{Distinct-N} = \frac{1}{N_w}\sum_{i=1}^{N_w}\frac{|\text{UniqueNgrams}_i|}{|\text{TotalNgrams}_i|} \quad (1)$$

where N_w is the number of sliding windows.

- **Semantic-Div** This metric is measured by calculating the cosine distance between text segments using BERT (Devlin, 2018) embeddings, providing a robust measure of semantic diversity.

$$\text{Semantic-Div} = \text{mean}\left(1 - \cos\left(\mathbf{e}_i, \mathbf{e}_j\right)\right) \quad (2)$$

$\mathbf{e}_i$ and $\mathbf{e}_j$ are BERT embeddings of different text windows. The cosine similarity between the embeddings is calculated as $\cos\left(\mathbf{e}_i, \mathbf{e}_j\right) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{|\mathbf{e}_i||\mathbf{e}_j|}$.

- **MATTR** (Covington and McFall, 2010) is for measuring lexical diversity by calculating the average Type-Token Ratio (TTR) across a sliding window. This approach reduces sensitivity to text length, providing a robust measure of vocabulary richness. Unlike metrics like Distinct-N, which focus on n-gram diversity, MATTR emphasizes word-level diversity, which useful for analyzing the linguistic richness of natural texts.

$$\text{MATTR} = \frac{1}{N_w}\sum_{i=1}^{N_w}\text{TTR}_i \quad (3)$$

- **Info-Dens** We measure information density using Shannon entropy (Shannon, 1948):

$$\text{Info-Dens} = -\sum_{i=1}^{N} p_i \log_2 p_i \quad (4)$$

where N is the number of unique tokens in the filtered token sequence (excluding stopwords), p_i is the probability of token i in the filtered sequence.

- **LLM-as-a-Judge** The emergence of LLMs as evaluation tools (Zheng et al., 2023) has revolutionized assessment methods for open-ended content generation. This approach proves particularly valuable when dealing with tasks lacking definitive answers and where traditional manual evaluation would be resource-intensive and potentially subjective. We design the evaluation prompt as shown in Figure 5. The key design principles are: 1) The evaluation metrics are primarily based on the template presented in (Zhang et al., 2024) for prompting GPT-4 to annotate the dialogue data. Additionally, we introduce a new metric, speaker diversity, to assess the diversity of viewpoints between different speakers. 2) We opt for a comparative evaluation between two samples, allowing the scores to reflect the relative quality of the dialogues. 3) To address the potential issue of position bias mentioned in (Zheng et al., 2023), we conduct two evaluations for each pair of dialogues - one with "dialogue A vs dialogue B", and another with "dialogue B vs dialogue A". We then average the results to obtain a more fair score. 4) Drawing from the approach described in (Wang et al., 2023b), we first ask the LLM to generate an explanation (evaluation evidence), and then provide the score. This allows the score to be calibrated with the evaluation evidence. 5) We use GPT-4 as the evaluator, ensuring more robust and reliable results.

```
### Dialogues-1:
${dialogue1_to_be_evaluated}

### Dialogues-2:
${dialogue2_to_be_evaluated}

## Instruction:
Compare the coherence, engagingness, diversity,
informativeness, and overall quality of the two-
input dialogue. Additionally, compare the
speaker diversity (except the host). A greater
speaker diversity means the viewpoints provided
by each speaker are more differentiated. Set the
Dialogues-1 as reference with score as 0, rate
each metric from -3 to 3, where a positive score
means Dialogues-2 performs better than
Dialogues-1, while a negative score indicates it
performs worse. Please first provide your
analysis process and then give the score.

### Output Format (json):
{"analysis":..., "coherence": x, "engagingness":
x, "diversity": x, "informativeness": x,
"overall": x, "speaker_difference": x}

### Your Response:
```

Figure 5: Prompt for GPT-4 Evaluator.

4.3 Evaluation on voices

Voice-Role matching To evaluate the effectiveness of PodAgent's voice-role matching mechanism, we conduct a subjective perception study involving 6 participants from diverse backgrounds. The study requires participants to evaluate 40 generated podcast segments, assessing whether the voices of the host and guests are appropriate and coherent with their assigned roles and the discussion topics. The participants provide a binary "pass" or "fail" judgment for each speaker in each segment. The overall pass rate serves as the key metric to quantify the success of the voice-role matching process.

Instruction-following speech synthesis To assess the impact of LLM-predicted speaking styles in PodAgent, we employ two evaluation metrics: preference scores and comparative mean opinion scores (CMOS). Both evaluation metrics are obtained by comparing speech generated with and without instruction guidance. For preference score, 9 evaluators selected between three options for each pair: “A wins”, “No preference”, or “B wins”. The CMOS require judgers to rate the sample pairs on a scale from -3 to 3, with the non-instructed sample as a reference point at 0. The audio samples are speech segments from the generated podcasts.

5 Experimental Analysis

Throughout our experiments, we set the guest number to 2, except for those conducted in the guest number analysis. To ensure experimental consistency and reliable performance, we exclusively employed GPT-4 for all LLM-dependent tasks.

5.1 Analysis on conversation scripts

To evaluate the effectiveness of our Host-Guest-Writer system, we conduct a comparative assessment against a baseline approach. The baseline implementation utilizes GPT-4 to directly generate conversation scripts from given topics, using the following prompt structure:

You are a talk show director and script writer. Here is the topic of the talk show: ... Please Write a corresponding talk show conversation script featuring 1 host and 2 guests.

We chose not to use WavJourney as a baseline due to the dialogue scripts it generates (Table 3) are very short and significantly lower in quality compared to our method. While dialogue scripts generated directly by GPT-4 (Table 4) are obvious content-rich than those created by WavJourney.

Table 1 presents our comparative evaluation results between the proposed system and the baseline. The quantitative metrics are expressed as difference scores ranging from -2 to 2, calculated by subtracting baseline scores from our system's scores. Additionally, we employed the "LLM-as-a-Judge" methodology, utilizing GPT-4 to provide comparative assessments on a scale of -3 to 3. For both evaluation approaches, positive values indicate superior performance by our proposed system, while negative values favor the baseline. Detailed metric descriptions can be found in Section 4.2.

The results demonstrate consistent and substantial improvements across all evaluation dimensions for conversations generated by the Host-Guest-Writer system. With only one minor exception - a marginal decline of -0.005 in the Semantic-Div score for the Generic category - our system outperformed the baseline across all metrics. These comprehensive positive results strongly validate the effectiveness of our Host-Guest-Writer approach in generating high-quality conversational content.

5.2 Analysis on voices

Voice-Role matching Figure 6 illustrates the Voice-Role matching evaluation results. With setup of two guests plus one host (three speakers total), a rating scale of 0-3 is to indicate the number of speakers successfully matching their assigned roles. The findings demonstrate robust performance across all categories, with over 60% of sessions achieved full matches where all voices aligned with their roles.

Metrics \ Categories		Generic	Knowledge	Common-sense	Counterfactual
Quantitative Metrics	Distinct_1	+0.031	+0.034	+0.028	+0.005
	Distinct_2	+0.016	+0.008	+0.011	+0.004
	Info-Dens	+0.707	+0.705	+0.670	+0.558
	Semantic-Div	-0.005	+0.010	+0.019	+0.008
	MATTR	+0.031	+0.034	+0.028	+0.005
LLM-as-a-Judge	Coherence	+0.7000	+0.6500	+0.6500	+0.6500
	Engagingness	+1.4500	+1.4500	+1.4000	+1.2500
	Diversity	+1.6500	+1.2500	+1.5500	+1.0000
	Informativeness	+1.9000	+1.8000	+2.2000	+1.7000
	Speaker-diversity	+1.1500	+1.1000	+1.3000	+1.1000
	Overall	+1.7500	+1.6625	+1.7250	+1.4000

Table 1: Evaluation on the Host-Guest-Scriptwriter System. **Baseline:** Directly ask the GPT-4 to generate a conversation script for a provided topic. **Quantitative metrics:** derived by subtracting the baseline score from the proposed model's score, yielding a range of -2 to 2. **LLM-as-a-Judge** scores range from -3 to 3. Positive values in all metrics indicate that the proposed model outperforms the baseline, whereas negative values suggest the opposite.

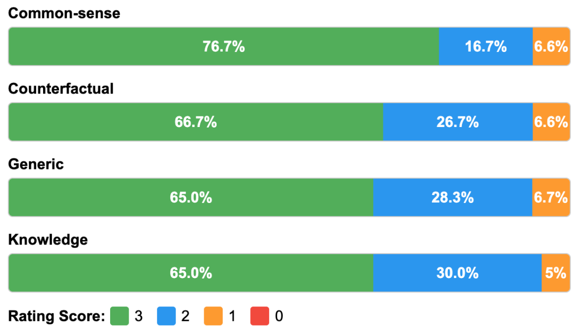

Figure 6: Voice-Role Matching result. This perception test is designed to evaluate whether the tone of each voice actor aligns with their respective roles and scenarios. The rating scale ranges from 0 to 3, where the score indicates the number of speakers that match effectively.

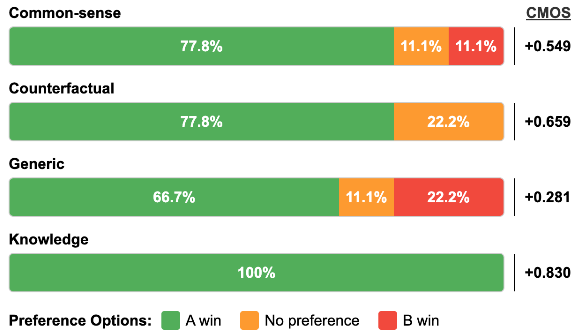

Figure 7: Evaluation on the instruction-following TTS. **Preference Test:** A - speech generation with guidance of LLM-predicted speaking style; B - speech generation without it. **CMOS Test:** Set B as reference, rating A from -3 to 3. Positive means A is better.

Moreover, more than 90% of sessions achieved successful matching for at least two speakers. The pass rates for the categories Common-sense, Counterfactual, Generic, and Knowledge are 90.0%, 86.7%, 86.1%, and 86.7%, respectively.

Instruction-following speech synthesis Figure 7 showcases the preference and the CMOS scores. The analysis demonstrates a clear preference for speech samples generated using LLM-predicted speaking styles across all categories. Furthermore, all CMOS scores are positive, ranging from 0.2 to 0.9, further supporting this conclusion.

5.3 Ablation Study

We perform ablation studies on our Host-Guest-Writer system to verify the effectiveness of three critical factors: guest number, outline, and multi-agent framework. Table 2 presents the results.

Guest number We explore the influence of the guest number by varying it from 1 to 5. Our findings indicate that increasing the number of participants does not always enhance content quality. We found that 2-guest setups consistently produce the most informative scripts based on quantitative metrics. While larger groups offer diverse perspectives, they often introduce redundancy and coordination challenges. Smaller groups foster more focused

Metrics / Methods	**Distinct_1**	**Distinct_2**	**Info-Dens**	**Semantic-Div**	**MATTR**
#Guest = 1	0.7392	0.9768	7.1247	0.1521	0.7392
#Guest = 2	**0.7662**	**0.9789**	7.0971	**0.2310**	**0.7662**
#Guest = 3	0.7367	0.9767	**7.1323**	0.1230	0.7367
#Guest = 4	0.7278	0.9768	6.9253	0.0977	0.7278
#Guest = 5	0.7012	0.9641	7.1328	0.1150	0.7037
#Guest = 2 (w/o outline)	0.7037	0.9569	6.9515	0.1275	0.7037
#Guest = 2 (Single Agent)	0.6559	0.9303	5.9001	0.1277	0.6559

Table 2: Ablation Study. 1. Guest Number; 2. with or without preset outline; 3. multi-agent VS single-agent

interactions and deeper dialogue, allowing participants to fully develop their ideas.

Outline The use of a topic-centered outline, created by the "Host", serves as a crucial structural framework for guiding guest interactions. Our experiments compare performance between scenarios with and without this outline, using two guests in both cases. Results show that *#Guest = 2 (w/o outline)* performs worse than *#Guest = 2* with outline, demonstrating the importance of structured guidance in the Host-Guest-Writer system.

Multi-agent system Our Host-Guest-Writer system is a multi-agent framework that collaborate multiple LLMs with distinct role settings to compose insightful conversation scripts from diverse perspectives. To evaluate its effectiveness, we compare it against a single-agent approach where one LLM handles all tasks. The comparison baseline *#Guest = 2 (Single Agent)* receives the instruction:

You are a talk show director and script writer. Here is the topic of the talk show:... Please follow the steps: 1. Based on the provided topic, invite 2 guests and provide detailed descriptions for each, including their ... 2. Create an interview outline consisting of five sub-questions related to the theme. 3. Write a talk show conversation script based on the unique role, experiences and diverse perspective of each invited guest...

As evidenced in Table 2, the multi-agent collaborative system demonstrates clear performance advantages over the single-agent approach across all evaluated metrics.

5.4 Case Study

In the Appendix B, we provide comparative examples of conversation scripts on the topic: “How can I develop my critical thinking skills?” Table 3 features dialogue content extracted from the audio script generated by WavJourney, which is notably short and lacks depth, with only 4 topic-related exchanges and providing limited information. This is due to that WavJourney generates dialogue as part of an audio script, which restricts multi-turn discussions. Table 4 presents scripts generated by directly asking GPT-4 with baseline instruction presented in Section 5.1. It shows modest improvement but still constrained to 4 turns. Table 5 showcases the Single-Agent version of the Host-Guest-Writer system we discussed in section 5.3. This case achieves richer content through task decomposition but lacks concluding remarks. Table 6 displays the conversation scripts produced by our proposed PodAgent’s Multi-Agent Host-Guest-Writer system, delivering the most comprehensive and well-structured discussion of the topic.

6 Conclusion

In this study, we proposed PodAgent, a comprehensive framework for creating audio programs that addresses the shortcomings of previous automated podcast-like generation methods. The key components of PodAgent include: 1) a Host-Guest-Writer system generating comprehensive, multi-perspective conversation scripts, 2) a preset diverse voice pool for suitable voice-role assignment, and 3) LLM-guided speech generation for enhanced expressiveness. Given the absence of established benchmarks in podcast generation, we designed a thorough experimental setup encompassing both qualitative and LLM-based evaluation of the conversation scripts, as well as voice-related metrics. Our extensive experimental results demonstrate PodAgent’s capability to produce high-quality, complete, and realistic audio programs.

7 Discussion

Limitations Although PodAgent is the first fully-automatic system capable of generating complete and informative podcast-like audio, several limitations in this study require further investigation: 1) Voice Quality. While we used a state-of-the-art open-source TTS foundation model to generate speech, offering improved robustness compared to earlier models, quality issues may still arise when generating large amounts of long-form content. 2) Voice Pool. In this work, reference speech segments were collected from LibriTTS (Koizumi et al., 2023). To produce more natural conversational audio, it is essential to expand the voice pool by incorporating more conversational-style voices. 3) Sound Effects and Music. This study primarily focuses on improving content and voice generation. However, there is room for enhancement in generating sound effects and music, as well as determining their appropriate placement within the audio.

Future Work To improve the podcast-listening experience, beyond just expanding the voice library diversity, a more advanced approach would be to generate new synthetic voices directly based on the desired characteristics, which can be more intelligent and help avoid some of the ethical concerns around real-voice cloning and consent. Additionally, the conversational expression can be further enhanced by adopting a more casual and natural style and incorporating appropriate vocal articulations like laughter, sighs, exclamations, and other non-semantic vocalizations, as incorporating these expressive sounds can make the conversation feel more lively and engaging for the user.

8 Ethics Statement

Since this work involves generating long-form audio content, including speech, music, and sound effects, we address considerations as follows: 1) Copyright and Intellectual Property. The generation of music, sound effects, and voices must respect existing copyright laws. In this work, we rely on open-source datasets and models to ensure compliance with intellectual property rights. Users are encouraged to verify that their use of PodAgent complies with copyright regulations. 2) Voice Cloning and Consent. In this study, we use anonymized, open-source speech data to avoid ethical violations. Users must ensure they have proper authorization and consent when using the system to generate speech resembling real individuals.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2023a. Bark: Text-to-audio model. https://github.com/suno-ai/bark.

Anthropic. 2023b. Claude: Ai assistant. https://www.anthropic.com/. Accessed: 2025-02-04.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, and et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org. Accessed: 2025-02-09.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023a. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *Preprint*, arXiv:2304.12995.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dias Issa, M Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1255–1264.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

LangChain. 2023. Langchain: A framework for building applications with llms. `https://langchain.com/`. Accessed: YYYY-MM-DD.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. 2023b. Wavjourney: Compositional audio creation with large language models. *arXiv preprint arXiv:2307.14335*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Inc. Midjourney. 2023. Midjourney. `https://www.midjourney.com`. Accessed: 2025-02-04.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus. In *Interspeech 2021*, pages 2756–2760.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2(5):9.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2023. Secap: Speech emotion captioning with large language model. *Preprint*, arXiv:2312.10381.

Kazuki Yamauchi, Yusuke Ijima, and Yuki Saito. 2024. Stylecap: Automatic speaking-style captioning from speech based on speech and language self-supervised learning models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11261–11265. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

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*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *Preprint*, arXiv:2312.15407.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. 2024. Unistyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7513–7522.

A Voice Library

English For the voice pool construction detailed in Section 3.2, we utilize speech segments from the LibriTTS-R dataset (Koizumi et al., 2023), an enhanced sound-quality version of the LibriTTS corpus (Zen et al., 2019). This dataset contains approximately 350,000 speech segments from over 2,000 speakers, with a balanced gender distribution. By screening and eliminating duplicate segments with similar voice characteristics, we develop a diverse voice library containing 222 unique speakers.

Madarin We also construct a Mandarin voice pool using the AISHELL-3 speech corpus (Shi et al., 2021), which contains approximately 85 hours of recordings from 218 native Mandarin speakers, totaling over 8,000 utterances. We do selection and deduplication based on the voice characteristic labels provided by SpeechCraft (Jin et al., 2024). As a result, we curated a diverse voice library comprising 172 speech segments from 85 distinct speakers.

B Case Study

Please refer to the examples shown in the tables below and the illustration provided in Section 5.4.

Speaker	**Speaking Content**
Host	Hello everyone, and welcome to our show. Today, we're discussing a very intriguing topic: How can I develop my critical thinking skills? To shed light on this topic, I'm delighted to introduce our first guest. Please welcome, Dr. Jane Doe, a renowned psychology professor.
Jane Doe	Thank you for having me here. I'm excited to share a few insights on improving critical thinking skills.
Host	Fantastic, we also have another guest joining us for the discussion. Please welcome, Mr. John Smith, a successful entrepreneur.
John Smith	Thank you. I'm thrilled to be here to discuss this important topic.
Host	Let's dive right into it then. Firstly, Dr. Doe, could you please explain what is critical thinking?
Jane Doe	Sure. Critical thinking refers to the ability to analyze information objectively and make a reasoned judgment.
Host	Excellent! Now, Mr. Smith, how has critical thinking played a role in your success?
John Smith	Critical thinking has been essential in my decision-making process. It's helped weigh my options logically before making a decision.
Host	Thank you both for your contributions. I believe this discussion helped our audience understand more about critical thinking.

Table 3: Example of conversation scripts generated by **WavJourney**. Generate a talk show featuring one host and two guests. The topic is: How can I develop my critical thinking skills?

Speaker	Speaking Content
Host	Hello and welcome to 'Mind Matters'! I'm your host, Steve, and I'm thrilled to have you join us for another exciting episode. Tonight, we delve into the human mind and explore the topic of critical thinking skills. Joining us tonight are two distinguished guests, Dr. Linda, a cognitive psychologist, and James, a renowned author of self-improvement literature. Welcome to the show, Dr. Linda and James!
Dr. Linda	Thank you, Steve. It's a pleasure to be here and discuss the importance of developing critical thinking skills.
James	Hello, Steve. Thank you for having me. I'm looking forward to sharing some practical tips and strategies.
Host	Fantastic! So to kick us off, Dr. Linda, would you mind giving us a brief overview of what critical thinking is and why it's important?
Dr. Linda	Certainly, Steve. Critical thinking is the ability to analyze information objectively and make a reasoned judgment. It involves the evaluation of sources, such as data, facts, observable phenomena, and research findings. Good critical thinkers can draw reasonable conclusions from a set of information, and distinguish between useful and less useful details.
Host	That's a clear explanation, thank you, Dr. Linda. How about you, James? As an author, how did developing your critical thinking skills influence your writing process?
James	Great question, Steve. Critical thinking plays a huge role. Not only does it help in researching and understanding different viewpoints before forming my own, but also in constructing clear, concise, and persuasive arguments. It's like having a good quality control mechanism in your brain!
Host	Quality control for the brain- I like that! We'll continue to ponder this and delve deeper into ways to develop these critical thinking skills after a short break. Stay tuned!

Table 4: Example of conversation scripts generated by **directly asking GPT-4** to generate a conversation script for a provided topic: How can I develop my critical thinking skills?

Speaker	Speaking Content
Host	It's a pleasure to have you with us on this enlightening journey to uncover the power of the mind. My first guest for today is Dr. Sarah Smith, a renowned psychologist and author, and my second guest is Mr. Peter Green, an innovative education consultant.
	Let's start with the basics, Dr. Smith can you enlighten us on what critical thinking is, and why it's important?
Dr. Sarah Smith	Certainly. Critical thinking involves objective analysis and evaluation of an issue to form a judgement. It's important as it helps us make informed and rational decisions, and question existing ideas and beliefs.
Host	Very insightful, Dr. Smith. Moving on to you, Mr. Green, could you share some strategies we can use to improve our critical thinking skills?
Mr. Peter Green	Of course. You can start with questioning everything, look for evidence, understand different perspectives, and practice active listening. And of course, continual learning and practicing these skills will only help you further develop them.
Host	Wonderful inputs, Mr. Green. Dr. Smith, do you have any captivating experiences or examples, to bring the concept of critical thinking to life?
Dr. Sarah Smith	Yes, in my experience, I've seen clients who have made drastic changes in their life just by applying critical thinking. One particular client was stuck in a cycle of unhealthy relationships but through the process of critical thinking, she was able to identify the problem areas and make better choices in her relationships.
Host	Impressive transformation indeed! Mr. Green, how does your work help in promoting critical thinking?
Mr. Peter Green	My education technology startup, ThinkTank, encourages higher-level thinking by providing stimulating content and puzzles. It's all about breaking down complex problems and finding creative solutions using critical thinking.
Host	That's quite commendable, Mr. Green. Lastly, what challenges do we face in teaching or learning critical thinking and how can they be addressed?
Dr. Sarah Smith	One major challenge is removing cognitive biases. It's a long process, but being aware of them and deliberately challenging them can help overcome this hurdle.
Mr. Peter Green	In addition to what Dr. Smith said, there's also a need for environments like schools, workplaces, or even homes to encourage critical thinking instead of accepting information at face value.

Table 5: Example of Conversation Scripts generated by **Host-Guest-Writer (Single-Agent)**. Topic: How can I develop my critical thinking skills?

Speaker	**Speaking Content**
Host	Hello everyone, welcome to our in-depth talkshow ’Think with Us’! I’m your host, Alex, and today we’ll be touching the cornerstone of decision-making, the art of Critical Thinking. We have two honored guests joining us tonight—Dr. Jane Martin, a renowned cognitive psychologist, and Prof. Tom Adams, a distinguished philosophy professor.
	Welcome to the show Dr. Martin and Prof. Adams! It’s great to have you both with us. Dr. Martin, I’ll begin with you. Could you briefly explain to our audience what exactly critical thinking is?
Dr. Jane Martin	Sure, Alex. Critical thinking is the ability to logically connect ideas, analyze our own thoughts and assess the validity of arguments around us. In an era of the information overload, it helps us make informed, accurate and ethical decisions.
Host	Insightful explanation, Dr. Martin! Now, Prof. Adams, would you suggest that critical thinking can be taught and developed?
Prof. Tom Adams	Absolutely, Alex. Encouraging curiosity, teaching argument validation, understanding biases, encouraging self-reflection and broadening the horizons of our knowledge can all contribute to the development of critical thinking. All it requires is patience and determination.
Host	Dr. Martin, as Prof. Adams mentioned biases, could you elaborate more on the obstacles that can hinder the development of critical thinking?
Dr. Jane Martin	Indeed. Intellectual humility, confirmation bias, cognitive laziness, and navigating the crowded information landscape all pose challenges. Also dealing with the discomfort of conflict between new and old ideas, also known as cognitive dissonance can be an obstacle.
Host	Prof. Adams, would you add some practical tips on improving these abilities?
Prof. Tom Adams	One should engage with different perspectives, debate with others and always find the counter-argument to their own point of view. Logical puzzles and keeping a journal of decision-making processes can significantly enhance critical thinking.
Host	Those are indeed valuable tips, thank you, Prof. Adams. Dr. Martin, how does your field of cognitive psychology contribute to critical thinking?
Dr. Jane Martin	Cognitive psychology provides insights into how we can understand and modify our thinking behaviors. It guides educators on how to teach critical thinking skills and helps individuals understand their own thought processes. We continually strive to improve thinking and enrich lives.
Host	That’s fascinating, Dr. Martin! And what would you say, Prof. Adams, is the contribution of your field?
Prof. Tom Adams	Philosophy is the bedrock of critical thinking. It encourages questioning, exploring diverse perspectives and seeks universal truths. The teachings of great philosophers like Socrates are valuable tools for nurturing critical thinking skills.
Host	That’s absolutely enlightening, thank you both for your insights. Ladies and gentlemen, that’s all the time we have for today. By embracing the art of critical thinking, we can better equip ourselves to navigate this complex world. Until next time, keep questioning, keep learning.

Table 6: Example of Conversation Scripts generated by **Host-Guest-Writer (Multi-Agent)**. Topic: How can I develop my critical thinking skills?