

SpeechRole: A Large-Scale Dataset and Benchmark for Evaluating Speech Role-Playing Agents

Changhao Jiang^{*1}, Jiajun Sun^{*1}, Yifei Cao^{*1}, Jiabao Zhuang^{*1}, Xinmeng Che¹,
Hui Li¹, Xiaoran Fan¹, Ming Zhang¹, Junjie Ye¹, Shihan Dou¹, Zhiheng Xi¹,
Jingqi Tong¹, Yilong Wu¹, Baoyu Fan², Tao Ji¹, Tao Gui¹, Qi Zhang¹, Xuanjing Huang^{†1}

¹Fudan University ²IEIT Systems Co., Ltd.

chjiang25@m.fudan.edu.cn, xjhuang@fudan.edu.cn

Abstract

Speech is essential for realistic role-playing, yet existing work on role-playing agents largely centers on text, leaving Speech Role-Playing Agents (SRPAs) underexplored and without systematic evaluation. We introduce SpeechRole, a unified framework for developing and assessing SRPAs. SpeechRole-Data contains 98 roles and 111k speech-to-speech conversations with rich timbre and prosodic variation, providing large-scale resources for training SRPAs. SpeechRole-Eval offers a multidimensional, controlled benchmark that directly evaluates generated speech relative to synthesized references, preserving paralinguistic cues and measuring interaction ability, speech expressiveness, and role-playing fidelity. Experiments show that end-to-end SRPAs such as GPT-4o Audio achieve strong fluency and naturalness, but remain limited in prosody consistency and emotion appropriateness. Recent open-source end-to-end models substantially narrow the gap with cascaded systems, although performance remains heterogeneous across languages, metrics, and multi-turn interaction. Furthermore, in the demonstrated Qwen2.5-Omni setting, role-specific training further improves end-to-end role-playing capabilities. We release all data, code, and evaluation tools at <https://github.com/yuhui1038/SpeechRole>.

1 Introduction

Recently, role-playing agents have emerged as a promising paradigm for achieving personalized interactions and emotional resonance (Chen et al., 2024a). Simulated characters with distinct traits make interactions more engaging across applications such as digital assistants (Xu et al., 2024) and educational tools (Yu et al., 2024). However, current research predominantly focuses on the text

modality (Li et al., 2023; Shao et al., 2023; Wang et al., 2024a; Sadeq et al., 2024), overlooking the crucial role of speech in authentic interactions.

Speech Role-Playing Agents (SRPAs) are speech-to-speech systems designed to generate spoken responses in character, reproducing distinctive vocal timbre, prosody, and persona traits while maintaining coherence across interactions. Paralinguistic features such as pitch, rhythm, intonation, and timbre are crucial for conveying character personality, emotion, and intent (Chang et al., 2023). Although preliminary studies have explored speech role-playing (Zhang et al., 2025), large-scale datasets and systematic evaluation benchmarks for SRPAs remain scarce.

To address this gap, we introduce SpeechRole, a unified framework for developing and evaluating SRPAs. SpeechRole-Data is a large-scale speech-to-speech dataset containing 111k dialogues across 98 roles, offering diverse vocal traits and prosodic patterns that support the construction of SRPAs. Building on this foundation, SpeechRole-Eval provides a multidimensional benchmark that directly evaluates generated speech without relying on speech-to-text conversion, thus preserving critical paralinguistic information. The benchmark assesses SRPAs along three complementary dimensions: interaction ability, speech expressiveness, and role-playing fidelity. Since its references are synthesized, SpeechRole-Eval is intended as a controlled and scalable benchmark for system comparison rather than a complete measure of natural actor-like role-playing. Together, SpeechRole-Data and SpeechRole-Eval establish a foundation for systematic research on speech role-playing.

Our empirical findings reveal substantial variation across current SRPA systems. GPT-4o Audio demonstrates notable advantages in speech fluency and naturalness, while prosody consistency and emotion appropriateness remain challenging. Recent open-source end-to-end systems substan-

^{*}Equal contribution.

[†]Corresponding author.

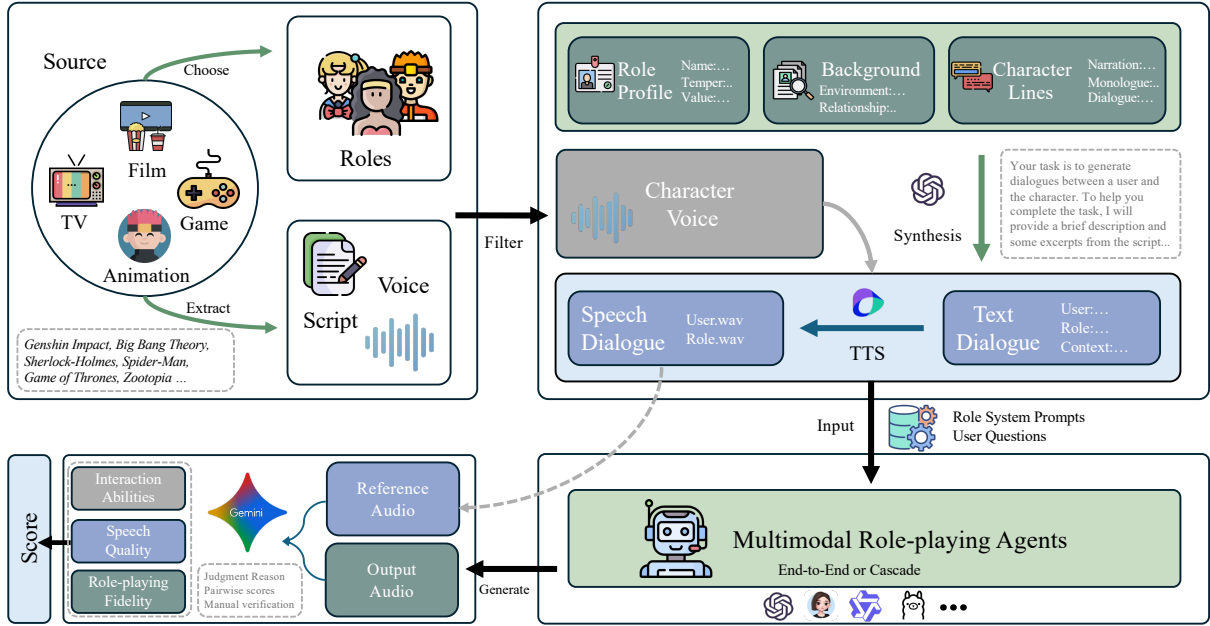


Figure 1: Overview of the SpeechRole framework. It covers role extraction with distinctive timbre and prosody, speech dialogue dataset construction, cascaded and end-to-end speech generation paradigms, and the SpeechRole-Eval benchmark for multidimensional evaluation of interaction abilities, speech quality, and role-playing fidelity.

tially narrow the gap with cascaded systems, especially in single-turn settings, but their performance remains heterogeneous across languages, metrics, and multi-turn interaction. Furthermore, in the demonstrated Qwen2.5-Omni setting, role-specific training significantly improves end-to-end role-playing capabilities, including on unseen roles. Overall, our main contributions are as follows:

1. We construct SpeechRole-Data, a large-scale speech-to-speech corpus with 111k dialogues across 98 roles, offering diverse timbral and prosodic patterns for developing SRPAs.
2. We introduce SpeechRole-Eval, a multidimensional benchmark that directly evaluates generated speech across interaction ability, speech expressiveness, and role-playing fidelity without relying on speech-to-text conversion.
3. We provide a systematic evaluation of representative cascaded and end-to-end SRPAs, together with empirical analysis of their strengths and limitations across multiple dimensions.

2 Related Work

Role-Playing and Persona-Grounded Dialogue. Persona-grounded dialogue has a long research

history prior to recent LLM-based role-playing agents. Early work explored persona consistency and speaker modeling in neural dialogue systems (Li et al., 2016), while datasets such as PersonaChat (Zhang et al., 2018) and EmpatheticDialogues (Rashkin et al., 2019) enabled controllable and emotionally aware conversational modeling. Large-scale open-domain dialogue systems, including DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020), and BlenderBot 3 (Shuster et al., 2022), further advanced conversational coherence, specificity, and long-term interaction quality. More recent work has increasingly explored explicit role-playing capabilities of LLMs, with most efforts focusing on the text modality. Representative examples include ChatHaruhi (Li et al., 2023) and RoleLLM (Wang et al., 2024a), which improve persona simulation and reduce hallucinations, as well as OmniCharacter (Zhang et al., 2025), which demonstrated the importance of incorporating speech for real-time, voice-aware character interactions. Existing benchmarks such as INCHARACTER (Wang et al., 2024b) focus on behavioral consistency and persona fidelity. More broadly, recent LLM benchmarks have examined evaluation protocols and evaluator design (Zhang et al., 2024), sequential learning capability (Dou et al., 2025), context learning (Dou et al., 2026), and retrieval and hierarchical organization in re-

Category1: Internal Reasoning		Category2: Experiential Narration		Category3: Social Communication	
Definition	Type	Definition	Type	Definition	Type
Concerning the character's subjective world, including personal perspectives, emotional responses, and underlying motivations.	Opinion and Emotion Inquiry Reasoning and Motivation Reflection and Change	Emphasizing the character's ability to recall past events, describe actions, and provide relevant background or historical information.	Action Description and Retrospection Skills and Abilities Historical and Background Description	Highlighting interpersonal understanding achieved through evaluative judgments, the provision of advice, and relational reasoning.	Judgment of Others and Events Advice and Life Experience
Example		Example		Example	
 Thor , what was going through your mind when you first woke up chained in Surtur's lair?		 Hu Tao , I've heard Wangsheng Funeral Parlor didn't originally handle funeral rites. Could you tell me about its early history?		 Officer Hopper , what advice would you give to someone who is told they can't achieve their dreams because of who they are?	
 Ah, yes. I awoke in that wretched cage thinking, 'Oh no! Thor's in a cage. How did this happen?' Truth be told, sometimes a bit of capture is needed to get a straight answer from someone. ...		 Hmm? Haven't I told you this before? Wangsheng was founded specifically to combat the grudges of fallen gods. (Sigh) All ancient history now! ...		 I'd tell them: Don't let anyone else decide what you're capable of—not even yourself, on the tough days. When I started out, everyone believed a bunny couldn't be a cop. But I worked hard. ...	
 Were you scared of Surtur and the prophecy about Ragnarok?		 Then how did it evolve into its current role handling matters of life and death?		 What if they're scared of failing or making mistakes?	
 Fear? No, not really. Annoyed, perhaps. Surtur wouldn't stop going on about his destiny and the destruction of Asgard. I was having these terrible dreams of Asgard burning. ...		 Though we no longer battle divine grudges, our techniques for detecting godly remnants still managed to survive. Gradually, funeral services became our primary focus. ...		 It's okay to be scared—I've made plenty of mistakes, big ones. What's important is learning from them and never giving up. ...	

Figure 2: Examples of the three major task categories in SpeechRole.

search synthesis (Zhang et al., 2026). Although these studies highlight the value of fine-grained, capability-oriented evaluation, systematic frameworks for speech role-playing remain scarce, leaving vocal expression, emotional depth, and paralinguistic fidelity underexplored.

Interactive Speech Agents. Speech-interactive LLMs typically follow two paradigms: cascaded and end-to-end. Cascaded systems perform speech dialogue by sequentially combining automatic speech recognition (ASR), text-based LLMs, and text-to-speech synthesis (TTS). This modular design has been widely adopted in systems such as ChatGPT and AudioGPT (Huang et al., 2024). Recent TTS models have improved naturalness, expressiveness, and controllability, although cascaded pipelines can suffer from error accumulation and limited cross-modal context integration. End-to-end systems aim to unify speech perception and generation within a single model. Early work such as SpeechGPT (Zhang et al., 2023) incorporated discrete speech units into LLM vocabularies. More recent architectures, including LLAMA-Omni (Fang et al., 2025) and Qwen2.5-Omni (Xu et al., 2025a), advance end-to-end speech modeling through dual-modality training, streaming capabilities, and low-latency generation, enabling more natural and efficient speech-driven interactions. Despite these advances, systematic evaluation of speech role-playing capabilities in end-to-end settings remains limited.

3 SpeechRole-Data

SpeechRole-Data is constructed to address the scarcity of scalable, persona-consistent speech dialogue resources for SRPAs. Real character dialogues with clean, reusable speech segments are difficult to obtain due to limited availability and copyright constraints. To enable controlled and large-scale data creation while preserving character identity, we adopt a synthetic pipeline combining LLM-based dialogue generation and TTS-based speech synthesis.

We curate 98 characters from television dramas, films, animations, and games, including 18 from ChatHaruhi (Li et al., 2023) and 80 from RoleLLM (Wang et al., 2024a), and collect their narrative profiles and scripts as persona grounding. Conditioned on these materials, an LLM generates user-character dialogues that maintain storyline and personality consistency. In parallel, representative speech segments are extracted to capture each character's timbre and prosody, which are then used for voice cloning. The generated texts and voice samples are combined through TTS to produce full speech-to-speech dialogues. This process yields 111k spoken interactions with controlled role diversity and vocal fidelity, supporting both training and evaluation of SRPAs.

3.1 Text Data Construction

Role Selection. We curate 98 diverse roles, including 18 from ChatHaruhi (Li et al., 2023) and

Splits	Characters	Samples (single-turn/multi-turn)	Speech Hours (user/character)
Train	78	89,461 (43,082/46,379)	875.25 (204.45/670.80)
Out-of-domain	20	21,993 (10,627/11,366)	203.60 (49.84/153.76)
Test	98	392 (196/196)	4.20 (0.94/3.26)

Table 1: Statistics of SpeechRole-Data across train, out-of-domain, and test splits.

80 from RoleLLM (Wang et al., 2024a). The selection balances language, gender, and personality, focusing on characters with distinctive vocal traits to enrich speech modeling.

Role Metadata Extraction. For each character, we extract structured metadata to build a comprehensive profile, including (1) Role Profile (temperament, preferences), (2) Background (social identity, relationships), and (3) Character Lines (dialogues, monologues). This metadata forms the basis for generating personality-consistent dialogues.

Dialogue Generation. Using the extracted metadata as a guide, we employ gpt-4.1-2025-04-14 (OpenAI, 2025) to generate conversations. For each of the 98 roles, we produce approximately 800 single-turn and 800 multi-turn dialogues, creating a substantial initial pool of text data.

Dialogue Deduplication. To improve data diversity and avoid repeated conversational patterns, we perform dialogue-level deduplication using semantic similarity. For each dialogue, we compute its embedding-based similarity with all other dialogues using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) for English and text2vec-bge-large-chinese (Xu, 2023) for Chinese. Dialogues with similarity above 0.9 for English or 0.85 for Chinese are considered near-duplicates, and only one instance is retained.

3.2 Role Voice Collection and Synthesis

Voice Collection and Preprocessing. To acquire authentic reference voices, we collect audio from sources such as the game Genshin Impact, whose assets permit non-commercial academic research use. For other audiovisual works, we only extract short speech fragments for analysis and dataset construction and do not distribute or reproduce any original copyrighted material. All collected audio is anonymized and segmented into brief utterances. Raw streams are extracted using ffmpeg (FFmpeg Developers, 2024) and converted to mono 16 kHz WAV format.

Audio Cleaning and Segmentation. We use the open-source Emilia framework (He et al., 2024) to process raw audio, performing source separation, speaker diarization, and voice activity detection. This pipeline produces clean 3–10 second single-speaker clips. The quality of each clip is assessed using DNSMOS P.835 OVRL (Reddy et al., 2022), and clips with an overall quality score below 3 are filtered out.

Role-Level Speaker Identification. Speaker diarization groups the original audiovisual audio into segments corresponding to different speakers, assigning each speaker a numeric ID but not revealing which character the ID represents. To map each segment to the correct character, we first transcribe the clips using Whisper-large-v3-turbo (Radford et al., 2023). We then use LLMs (gpt-4.1-2025-04-14 and DeepSeek-V3-0324 (DeepSeek-AI et al., 2024)) to infer character identity based on the transcribed text and contextual cues. Finally, all assignments are manually verified for accuracy.

Reference Voice Selection. We select a representative reference voice for each role by ranking all its cleaned clips according to vocal consistency. For each clip, we extract a speaker embedding using the CAM++ model (Wang et al., 2023). We then compute its average cosine similarity to all other clips belonging to the same role and choose the clip with the highest similarity as the reference voice, as it best reflects the role’s characteristic timbre and speaking style.

Role and User Speech Synthesis. Using the collected reference voices, we synthesize role utterances with three state-of-the-art TTS models—CosyVoice2 (Du et al., 2024), F5-TTS (Chen et al., 2024b), and E2 TTS (Eskimez et al., 2024)—each conditioned on the selected reference clip to preserve timbre and prosody. Generating multiple synthesized versions mitigates model-specific artifacts and reduces potential bias introduced by any single TTS system. User ut-

terances are generated using the VolcEngine TTS system (VolcEngine, 2024) with a fixed voice, as user speech does not require persona-specific vocal variation. This process yields multiple speech-to-speech dialogue variants aligned with both the textual content and role-specific vocal characteristics.

3.3 Dataset Statistics

SpeechRole-Data comprises 98 distinct roles and 111k speech-to-speech dialogue samples, covering both single-turn and multi-turn conversations. Each role is paired with a detailed role profile and a reference audio clip that illustrates its characteristic speaking style, including timbre, rhythm, and prosody. Each dialogue sample contains speech data and contextual information relevant to the role, enabling contextually grounded and persona-consistent role-playing.

The dataset is divided into three subsets. The training set contains dialogues from 78 roles. The out-of-domain set includes 20 roles that do not appear in the training data, enabling evaluation of generalization to unseen characters. The test set covers all 98 roles, with four curated dialogue sessions per role, resulting in 392 dialogue-level evaluation instances. In total, the test set comprises 1,448 role utterances and approximately 4.2 hours of speech. Evaluation scores are aggregated across all utterances at the system level, providing substantially denser evidence than the dialogue count alone suggests. This scale is comparable to prior role-playing evaluation benchmarks, such as the 400-sample evaluation in Zhang et al. (2025) and the 294-sample benchmark introduced by Dai et al. (2025). Detailed examples are provided in Figure 2.

To further characterize the acoustic properties of the dataset, we analyze the duration distribution of individual speech segments for both users and roles. As shown in Figure 3, role utterances are generally longer, typically ranging from 10 to 20 seconds, reflecting their narrative or expressive nature. In contrast, user utterances are concentrated between 3 and 6 seconds, highlighting their concise and goal-oriented characteristics.

4 SpeechRole-Eval

Evaluating SRPAs is inherently challenging, as performance depends not only on dialogue coherence but also on vocal expressiveness and persona consistency. Existing evaluations rely heavily on human judgments (Zhang et al., 2025), which, while

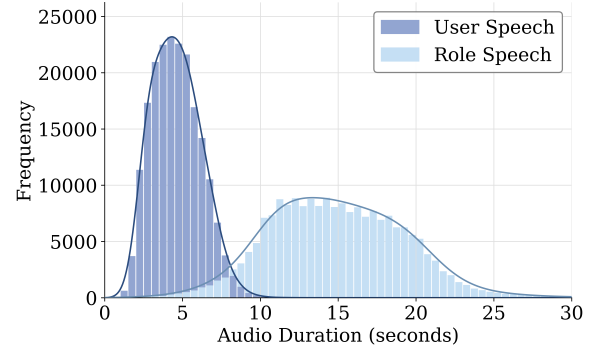


Figure 3: Speech duration distribution of SpeechRole-Data.

informative, are costly and difficult to scale consistently across studies. The absence of a standardized and scalable evaluation protocol further limits systematic comparison between models.

To address these challenges, we introduce SpeechRole-Eval, a controlled benchmark designed to assess SRPAs across three key dimensions: interaction ability, speech expressiveness, and role-playing fidelity. The evaluation protocol leverages a large language model (LLM) with speech understanding capabilities to provide scalable scoring relative to synthesized references. We validate its system-level ranking consistency against human annotations, but do not treat the judge as a fully calibrated substitute for human perceptual assessment. This section describes the task setup, evaluation criteria, and the overall evaluation pipeline.

4.1 Evaluation Metrics

SRPAs must handle not only the linguistic demands of dialogue generation but also the vocal and stylistic requirements of spoken character portrayal. Evaluating such systems therefore requires attention to aspects beyond those considered in traditional text-based settings. To capture the full range of SRPA capabilities, we adopt a three-dimensional evaluation framework.

Interaction Ability: measures the agent’s capacity to produce coherent, contextually appropriate, and instruction-aligned responses. Assessed using Instruction Adherence (IA) and Conversational Coherence (CC).

Speech Quality and Expressiveness: evaluates the naturalness, fluency, and prosodic expressiveness of speech. Assessed using Speech Fluency (SF), Speech Naturalness (SN), Prosodic Consistency (PC), and Emotion Appropriateness (EA).

Because PC and EA depend on subtle and subjective acoustic cues, we interpret them as diagnostic indicators alongside the other metrics rather than as standalone decisive measures.

Role-Playing Fidelity: examines how well the agent maintains character-specific personality and knowledge. Assessed using Personality Consistency (PeC) and Knowledge Consistency (KC).

4.2 LLM-Based Evaluation Procedure

Absolute scoring with LLM-based judges has been shown to be unstable in open-ended generation tasks, as scores are sensitive to prompt phrasing, scale interpretation, and the absence of a single ground-truth response. To mitigate these issues, prior work suggests pairwise comparison as a more reliable alternative, producing relative preference signals that better align with human judgments (Dai et al., 2025).

Following this paradigm, SpeechRole-Eval adopts a pairwise comparison protocol. For each test case, the LLM jointly evaluates a candidate response and a high-quality reference response, assigning comparative scores to both. This design reduces scale ambiguity and improves discriminative consistency. To improve interpretability and judgment stability, the LLM first generates a brief rationale before providing final rating, encouraging more deliberate and structured evaluation.

We employ gemini-2.5-pro (Comanici et al., 2025) to assign scores following the pairwise evaluation protocol, with ratings ranging from 1 to 10. Each test sample is evaluated against three high-quality reference responses synthesized using different TTS backends. This multi-reference design mitigates potential bias introduced by any single synthesis system and improves robustness. For each reference j , the relative score is computed as the ratio between the test and the reference ratings. The final score for an instance is obtained by averaging across references. Formally, for N evaluation instances and $M = 3$ reference variants per instance, the metric is defined as:

$$\text{Score} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{M} \sum_{j=1}^M \frac{s_i^{\text{test}}}{s_{i,j}^{\text{ref}}} \right) \quad (1)$$

Aggregation is performed at the instance level before computing system-level averages. This multi-reference ratio standardizes scores across prompts and reduces variance stemming from reference-

specific artifacts, supporting more stable system-level comparisons within this controlled setting.

5 Experiments

This section presents an experimental evaluation of SRPAs on SpeechRole-Eval. We benchmark representative cascaded and end-to-end SRPA paradigms in both English and Chinese.

In addition to existing baselines, we include SpeechRole-Agent, an end-to-end task-specific model obtained by fine-tuning Qwen2.5-Omni on SpeechRole-Data, to examine the impact of role-specific speech training. To ensure robust evaluation, all systems are assessed using the proposed multi-reference LLM-based protocol described in Section 4. Finally, we conduct a human-LLM agreement study to validate the reliability of the automatic evaluation framework.

5.1 Evaluated Systems

We evaluate a total of eleven SRPAs, including both cascaded and end-to-end systems.

Cascaded Systems. Cascaded SRPAs follow a three-stage pipeline consisting of ASR, text-based reasoning, and TTS. User speech is first transcribed using Whisper-large-v3-turbo (Radford et al., 2023). The transcribed text is then processed by a LLM to generate role-conditioned responses. We consider three open-source LLMs for this stage: Qwen3-8B (Yang et al., 2025), Llama-3.1-8B (Dubey et al., 2024), and Mistral-7B (Jiang et al., 2023). Finally, generated responses are converted into speech using F5-TTS (Chen et al., 2024b), which supports reference-based voice cloning to match the target character’s timbre.

In addition, we include a proprietary cascaded baseline built upon the Alibaba Cloud API (Alibaba Cloud, 2024), which integrates Paraformer Realtime ASR v2 for speech recognition, Qwen-Plus-Character for role-conditioned response generation, and CosyVoice-v2 for speech synthesis. This system also supports reference-based voice cloning.

End-to-End Systems. End-to-end SRPAs directly generate speech responses from speech inputs without explicit intermediate text representations. Such architectures are expected to reduce latency and error accumulation in cascaded pipelines. We evaluate the proprietary GPT-4o Audio (Hurst et al., 2024) and five open-source

Models	IA	CC	SF	SN	PC	EA	PeC	KC	Overall
<i>English Evaluation Results</i>									
Alibaba Cloud API	0.904	0.959	0.895	0.815	0.823	0.797	0.826	0.909	0.866
GPT-4o Audio	0.813	1.023	1.098	1.021	0.900	0.816	0.774	0.922	0.921
Qwen3-8B	0.935	0.893	0.920	0.895	0.924	0.943	0.927	0.906	0.918
Llama-3.1-8B	0.949	0.958	0.975	0.970	0.966	0.964	0.951	0.920	0.957
Mistral-7B	0.827	0.903	0.959	0.865	0.859	0.810	0.786	0.867	0.859
LLaMA-Omni	0.609	0.766	0.690	0.538	0.560	0.502	0.512	0.708	0.611
Qwen2.5-Omni	0.471	0.666	0.743	0.656	0.550	0.452	0.388	0.613	0.568
Qwen3-Omni	0.798	0.782	0.843	0.835	0.757	0.789	0.756	0.844	0.801
MiMo-Audio	0.853	0.915	0.753	0.724	0.716	0.752	0.747	0.827	0.786
Step-Audio-2-mini	0.761	0.914	0.896	0.835	0.645	0.655	0.593	0.716	0.752
SpeechRole-Agent	0.721	0.939	1.001	0.861	0.753	0.660	0.591	0.882	0.801
SpeechRole-Agent (OOD)	0.710	0.921	0.984	0.848	0.739	0.667	0.567	0.827	0.783
<i>Chinese Evaluation Results</i>									
Alibaba Cloud API	0.945	1.029	1.135	1.069	1.032	0.982	0.965	0.881	1.005
GPT-4o Audio	0.832	1.031	1.104	0.924	0.876	0.775	0.750	0.933	0.903
Qwen3-8B	0.717	0.787	0.895	0.788	0.833	0.802	0.771	0.853	0.806
Llama-3.1-8B	0.797	0.803	0.908	0.849	0.847	0.783	0.746	0.753	0.811
Mistral-7B	0.622	0.710	0.848	0.723	0.718	0.627	0.562	0.663	0.684
Qwen2.5-Omni	0.562	0.787	0.987	0.842	0.722	0.587	0.473	0.691	0.706
Qwen3-Omni	0.796	0.757	0.900	0.914	0.832	0.854	0.819	0.853	0.841
MiMo-Audio	0.805	0.892	0.871	0.859	0.807	0.851	0.827	0.837	0.844
Step-Audio-2-mini	0.737	0.890	0.953	0.888	0.695	0.697	0.627	0.725	0.776
SpeechRole-Agent	0.885	1.003	1.125	0.988	0.911	0.812	0.777	0.984	0.936
SpeechRole-Agent (OOD)	0.892	1.002	1.117	0.969	0.898	0.800	0.763	0.986	0.928

Table 2: Main results on the SpeechRole-Eval benchmark in English and Chinese. We report performance across eight evaluation metrics: Instruction Adherence (IA), Conversational Coherence (CC), Speech Fluency (SF), Speech Naturalness (SN), Prosody Consistency (PC), Emotion Appropriateness (EA), Personality Consistency (PeC), and Knowledge Consistency (KC), together with the overall score. SpeechRole-Agent is obtained by fine-tuning Qwen2.5-Omni on SpeechRole-Data and evaluated on all roles, while SpeechRole-Agent (OOD) reports results on roles unseen in the training set only. Models are grouped into three categories: proprietary APIs, open-source cascaded systems, and open-source end-to-end speech models. Bold indicates the best score within each group.

models: Qwen2.5-Omni-7B (Xu et al., 2025a), LLaMA-Omni (Fang et al., 2025), Qwen3-Omni (Xu et al., 2025b), MiMo-Audio (LLM-Core-Team Xiaomi, 2025), and Step-Audio-2-mini (Wu et al., 2025). Together, these systems provide broader coverage of current end-to-end speech modeling approaches. Notably, LLaMA-Omni only supports English and is therefore excluded from Chinese evaluations.

Role-Specific Fine-Tuning. To investigate the impact of role-specific speech supervision, we further fine-tune Qwen2.5-Omni-7B on SpeechRole-Data, producing a specialized SRPA referred to as SpeechRole-Agent. Since the training set contains 78 roles, we report two variants in the bench-

mark results: SpeechRole-Agent, evaluated on all 98 roles, and SpeechRole-Agent (OOD), evaluated on 20 roles unseen in the training set only. This design allows us to examine both fitting ability and generalization to novel characters.

5.2 Overall Results and System Comparison

Table 2 reports the performance of all evaluated systems on SpeechRole-Eval in both English and Chinese. Overall, several trends can be observed across the evaluated systems.

Interaction Ability. Interaction ability metrics (IA and CC) reflect the model’s ability to understand user intent and maintain coherent dialogue flow. Several cascaded systems built upon

Language	Measure	IA	CC	SF	SN	PC	EA	PeC	KC
English	Spearman’s ρ	0.88	0.79	0.83	0.71	0.98	0.77	0.67	0.88
	Kendall’s τ	0.71	0.64	0.71	0.57	0.93	0.62	0.50	0.79
	Human–LLM Agreement	0.86	0.82	0.86	0.79	0.96	0.81	0.75	0.89
Chinese	Spearman’s ρ	0.72	0.70	0.77	0.86	0.64	0.64	0.96	0.89
	Kendall’s τ	0.59	0.59	0.59	0.71	0.43	0.43	0.90	0.81
	Human–LLM Agreement	0.80	0.80	0.80	0.86	0.71	0.71	0.95	0.90

Table 3: System-level ranking correlation and pairwise agreement between LLM-based evaluation and human judgments across evaluation dimensions in English and Chinese.

strong text-based LLMs, such as Llama-3.1-8B and Qwen3-8B, achieve strong performance on these metrics. Recent end-to-end models substantially improve over the earlier Qwen2.5-Omni baseline, although their performance remains heterogeneous. For example, MiMo-Audio obtains strong IA and CC scores, whereas Qwen3-Omni shows a clear degradation from single-turn to multi-turn evaluation, as detailed in Appendix F.3. After fine-tuning on SpeechRole-Data, SpeechRole-Agent improves over Qwen2.5-Omni across both metrics in English and Chinese.

Speech Quality and Expressiveness. Metrics including SF, SN, PC, and EA evaluate the perceptual quality and expressive control of generated speech. GPT-4o Audio achieves strong performance in speech fluency and naturalness, while several cascaded systems also perform competitively due to high-quality TTS components. Recent end-to-end models narrow the earlier gap but exhibit different trade-offs across fluency, naturalness, prosody, and emotion. Because PC and EA show lower agreement than most other metrics, we interpret them cautiously and jointly with the remaining dimensions. Fine-tuning with SpeechRole-Data improves the speech generation ability of SpeechRole-Agent over its Qwen2.5-Omni base.

Role-Playing Fidelity. Role-playing fidelity metrics (PeC and KC) measure whether systems maintain consistent persona traits and role-specific knowledge. Several cascaded systems with strong LLMs achieve stable role-playing fidelity, while recent end-to-end models substantially narrow the gap but do not uniformly dominate across languages and metrics. In the demonstrated Qwen2.5-Omni setting, role-specific training substantially improves SpeechRole-Agent over its base model. Importantly, the improvements

are not limited to roles seen during training. When evaluated only on the 20 unseen roles, SpeechRole-Agent (OOD) maintains similar performance with minor degradation.

Additional Robustness Analyses. The appendix further reports a text/role versus audio/expressiveness decomposition, speaker-similarity checks, open-judge experiments, and a leave-F5-out control. The leave-F5-out rankings remain highly consistent with the original results (Spearman’s $\rho = 0.9762$ in English and 0.9286 in Chinese), reducing concern that reference-backend overlap drives the main comparison.

5.3 Human–LLM Agreement Analysis

To examine whether the proposed LLM-based evaluation framework produces system rankings consistent with human judgment, we conduct a human–LLM agreement study. We randomly sample half of the evaluation instances for each of eight evaluated SRPAs and perform human assessment on both English and Chinese subsets.

Three expert annotators independently rate each response following the eight dimensions defined in Section 4.1. Human ratings are assigned on a 1–5 scale. After annotation, the scores for each model and metric are aggregated by averaging the ratings across annotators and instances.

We then analyze the correlation between the aggregated human scores and the automatic scores produced by SpeechRole-Eval. Since the LLM-based evaluation outputs continuous scores that are not restricted to the same 1–5 scale, we primarily focus on rank-based agreement measures. Specifically, we report Spearman’s ρ to measure rank correlation and Kendall’s τ to measure pairwise ranking consistency between human and LLM evaluations. In addition, we compute Human–LLM Agreement, defined as the proportion of model

pairs for which human and LLM evaluations produce consistent relative ordering.

The results are summarized in Table 3. Overall, strong correlations are observed across most evaluation dimensions in both languages. Spearman’s ρ ranges from 0.67 to 0.98 in English and from 0.64 to 0.96 in Chinese, while Kendall’s τ ranges from 0.50 to 0.93 and from 0.43 to 0.90, respectively. The Human–LLM Agreement scores are consistently high, exceeding 0.75 for most metrics. We note that some fine-grained metrics, such as PC and EA in Chinese, show lower rank agreement (Kendall’s $\tau = 0.43$), reflecting the difficulty of judging subtle pitch, rhythm, pause, and emotional cues. These dimensions are also more subjective among human raters. We therefore interpret PC and EA as diagnostic indicators rather than standalone decisive metrics. Overall, the results support use of the LLM judge as a scalable proxy for system-level comparison, but not as a fully calibrated replacement for human perceptual judgment. Appendix F.5 reports human–human agreement and bootstrap confidence intervals, while Appendix F.7 examines two open audio-capable judges.

6 Conclusion

We present SpeechRole, a unified framework for advancing speech role-playing agents (SRPAs) through both large-scale data and systematic evaluation. The framework includes SpeechRole-Data, a speech-to-speech role-playing dataset covering 98 roles with diverse timbral and expressive characteristics, and SpeechRole-Eval, a multidimensional benchmark that evaluates generated speech across interaction ability, speech expressiveness, and role-playing fidelity.

Experiments on cascaded and end-to-end systems characterize current SRPA capabilities across complementary dimensions. Recent end-to-end models substantially narrow the gap with cascaded systems, while performance remains heterogeneous across languages, metrics, and multi-turn interaction, and prosody control and emotional expression remain particularly challenging. In the demonstrated Qwen2.5-Omni setting, role-specific training improves performance, including on unseen roles. The proposed LLM-based evaluation shows system-level ranking consistency with expert judgments and serves as a scalable proxy for controlled comparison relative to synthesized references. We hope SpeechRole facilitates future

research on expressive and controllable speech role-playing systems.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62576106, 62521004, 62476061, 62376061, 62506079).

7 Limitations

Despite its contributions, this work has several limitations.

First, although SpeechRole-Data covers a diverse set of roles and speaking styles, the number of languages remains limited, with experiments primarily focusing on English and Chinese. Extending the dataset and evaluation framework to additional languages and cultural contexts would improve the generality of the benchmark.

Second, all speech in SpeechRole-Eval is generated using text-to-speech (TTS) systems rather than natural human recordings. While this design allows us to construct controlled and scalable evaluation data with consistent role descriptions and speaking styles, it may not fully capture the variability and acoustic complexity of natural speech. In particular, the short reference clips mainly constrain speaker identity and timbre rather than fully representing a character’s acting style, emotional range, pacing, rhythm, or context-dependent delivery. As a result, the benchmark primarily reflects how SRPA systems perform relative to synthesized reference speech and should not be interpreted as a complete measure of natural actor-like role-playing. Future work could incorporate natural recordings to better evaluate robustness in real-world scenarios.

Third, our automatic evaluation relies on a strong proprietary LLM as the judge. Although the human–LLM study supports system-level ranking consistency, the judge is not a fully calibrated replacement for human perceptual assessment and may miss low-level artifacts such as clipping, vocoder noise, unnatural pauses, muffling, or robotic prosody. Prosodic Consistency and Emotion Appropriateness are especially subjective and show lower agreement than most other metrics, so we treat them as diagnostic indicators. Our open-judge experiments also reveal ceiling effects or compressed scores under the current protocol; developing more discriminative open judges and

complementary objective speech metrics remains important future work.

Fourth, the current evaluation focuses primarily on system-level performance and averaged behavior. Its 392 curated sessions and 1,448 role utterances provide dense utterance-level evidence, but the findings should not be overgeneralized beyond the covered roles, topics, languages, or controlled setting. The benchmark also does not explicitly assess fine-grained controllability, long-term character drift, or robustness under adversarial or ambiguous role instructions. These aspects are critical for real-world deployment of SRPAs and warrant further investigation. Additionally, our evaluation is currently limited to a zero-shot setting. Exploring few-shot or in-context role-conditioning may further enhance the capabilities of SRPAs, and we consider this an important direction for future research.

Finally, the demonstrated training result is specific to fine-tuning Qwen2.5-Omni with SpeechRole-Data, including evaluation on 20 unseen roles. Whether comparable gains transfer to other end-to-end backbones remains an open question. Exploring alternative architectures, training objectives, and tighter integration between language and speech modeling may further improve expressive and faithful role-playing in future systems.

8 Ethics Statement

This work does not involve the release of newly annotated human-labeled datasets. Human annotation is only conducted for evaluation purposes in the Human-LLM agreement study, and the resulting annotations are not included in the released resources. All human evaluations are performed by expert annotators selected from the research team, following a predefined evaluation protocol. No personal data is collected, and all evaluated content is derived from fictional role-playing dialogues.

SpeechRole nevertheless involves recognizable fictional characters, source scripts, audiovisual works, and reference-conditioned voice cloning. These materials raise potential concerns regarding copyright, consent, performer likeness, and downstream misuse. We do not distribute original copyrighted audio extracted from audiovisual works. Released resources are intended for research use under the applicable licenses and terms of use, and users should respect the rights of character cre-

ators, performers, and content owners. Synthesized character speech should be clearly identified as AI-generated and must not be used to impersonate real performers, mislead audiences, or enable unauthorized commercial exploitation. These safeguards reduce, but do not eliminate, the risks associated with character and voice imitation.

References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *CoRR*, abs/2001.09977.
- Alibaba Cloud. 2024. Alibaba cloud platform. <https://www.alibabacloud.com>.
- Che-Jui Chang, Samuel S. Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. [The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents](#). In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, Sydney, NSW, Australia, March 27-31, 2023*, pages 790–801. ACM.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. [From persona to personalization: A survey on role-playing language agents](#). *Trans. Mach. Learn. Res.*, 2024.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. [F5-TTS: A fairytale that fakes fluent and faithful speech with flow matching](#). *CoRR*, abs/2410.06885.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. [Mmrole: A comprehensive framework for developing and evaluating](#)

- multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. [Deepseek-v3 technical report](#). *CoRR*, abs/2412.19437.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification](#). In *Interspeech 2020*, pages 3830–3834. ISCA.
- Shihan Dou, Ming Zhang, Chenhao Huang, Jiayi Chen, Feng Chen, Shichun Liu, Yan Liu, Chenxiao Liu, Cheng Zhong, Zongzhang Zhang, Tao Gui, Chao Xin, Chengzhi Wei, Lin Yan, Qi Zhang, and Xuanjing Huang. 2025. [EvaLearn: Quantifying the learning capability and efficiency of LLMs via sequential problem solving](#). In *Advances in Neural Information Processing Systems*.
- Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, and 1 others. 2026. [CL-Bench: A benchmark for context learning](#). *CoRR*, abs/2602.03587.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. [Cosyvoice 2: Scalable streaming speech synthesis with large language models](#). *CoRR*, abs/2412.10117.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. [E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS](#). In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 682–689. IEEE.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. [Llama-omni: Seamless speech interaction with large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- FFmpeg Developers. 2024. Ffmpeg: A complete, cross-platform solution to record, convert and stream audio and video. <https://ffmpeg.org/>.
- Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. [Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation](#). In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 885–890. IEEE.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jiaotong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. [Audiogpt: Understanding and generating speech, music, sound, and talking head](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 23802–23804. AAAI Press.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#). *CoRR*, abs/2308.09597.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. [A persona-based neural conversation model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.
- LLM-Core-Team Xiaomi. 2025. [Mimo-audio: Audio language models are few-shot learners](#).
- OpenAI. 2025. Gpt-4.1. <https://openai.com/index/gpt-4-1/>.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023.

- Robust speech recognition via large-scale weak supervision.** In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. **Towards empathetic open-domain conversation models: A new benchmark and dataset.** In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.
- Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2022. **Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors.** In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 886–890. IEEE.
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-bert: Sentence embeddings using siamese bert-networks.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian J. McAuley. 2024. **Mitigating hallucination in fictional character role-play.** In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14467–14479. Association for Computational Linguistics.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. **Character-llm: A trainable agent for role-playing.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13153–13187. Association for Computational Linguistics.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. **Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.** *CoRR*, abs/2208.03188.
- VolcEngine. 2024. Volcengine platform. <https://www.volcengine.com>.
- Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023. **CAM++: A fast and efficient network for speaker verification using context-aware masking.** In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 5301–5305. ISCA.
- Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. **Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.** In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14743–14777. Association for Computational Linguistics.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. **Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.** In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1840–1873. Association for Computational Linguistics.
- Boyong Wu and 1 others. 2025. **Step-audio 2 technical report.** *CoRR*, abs/2507.16632.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. **Qwen2.5-omni technical report.** *CoRR*, abs/2503.20215.
- Jin Xu and 1 others. 2025b. **Qwen3-omni technical report.** *CoRR*, abs/2509.17765.
- Ming Xu. 2023. **text2vec: A tool for text to vector.**
- Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024. **MINDECHO: role-playing language agents for key opinion leaders.** *CoRR*, abs/2407.05305.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chenyan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. **Qwen3 technical report.** *CoRR*, abs/2505.09388.
- Jifan Yu, Zheyuan Zhang, Daniel Zhang-li, Shangqing Tu, Zhanxin Hao, Rui Miao Li, Haoxuan Li, Yuanchun Wang, Hanming Li, Linlu Gong, Jie Cao, Jiayin Lin, Jinchang Zhou, Fei Qin, Haohua Wang, Jianxiao Jiang, Lijun Deng, Yisi Zhan, Chaojun Xiao, and 14 others. 2024. **From MOOC to MAIC: reshaping online teaching and learning through llm-driven agents.** *CoRR*, abs/2409.03512.
- 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. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, and Yongbin Li. 2025. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. *CoRR*, abs/2505.20277.

Ming Zhang, Jiabao Zhuang, Wenqing Jing, Kexin Tan, Ziyu Kong, Jingyi Deng, Yujiong Shen, and 1 others. 2026. Can deep research agents retrieve and organize? evaluating the synthesis gap with expert taxonomies. *CoRR*, abs/2601.12369.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LLMEval: A preliminary study on how to evaluate large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19615–19622.

A AI Assistants in Research or Writing

In preparing this manuscript, AI assistants were employed solely to assist with refining the clarity, style, and readability of certain text segments. They were not involved in designing the study, developing or implementing the methodology, collecting or analyzing data, or generating the primary scientific contributions. All substantive research decisions, analyses, and conclusions are fully the responsibility of the authors.

B Evaluation Metrics Details

- **Instruction Adherence (IA):** Do the spoken responses strictly follow the task instruction,



Figure 4: Distribution of eight sub-tasks across three major task categories in SpeechRole.

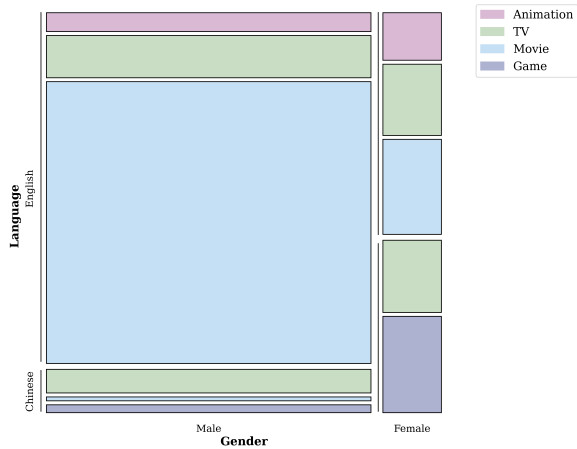


Figure 5: Distribution of the 98 roles in SpeechRole-Data by language, gender, and data source.

remaining fully in character without any out-of-role explanations or assistant-like meta-comments?

- **Conversational Coherence (CC):** Do the responses maintain logical consistency within the dialogue, aligning with previous content without contradictions or abrupt topic shifts?
- **Speech Fluency (SF):** Are the responses delivered fluently, with smooth articulation, appropriate pacing, and minimal disfluencies such as stuttering or unnatural pauses?
- **Speech Naturalness (SN):** Do the responses sound natural, human-like, and free from noticeable artifacts or robotic effects typically associated with synthetic speech?
- **Prosodic Consistency (PC):** Does the prosody, including pitch, stress, and intona-

tion, align with the character’s intended speaking style and remain consistent across the discourse?

- **Emotion Appropriateness (EA):** Are emotional cues in the speech (e.g., anger, joy, sadness) well-aligned with the dialogue context and the character’s emotional state?
- **Personality Consistency (PeC):** Do the responses consistently reflect the character’s personality traits, such as optimism, sarcasm, or authority?
- **Knowledge Consistency (KC):** Are the responses grounded in the character’s established background, knowledge, and relationships, without fabricating out-of-character facts?

C Additional data statistics

Role Type Distribution Figure 5 presents the distribution of characters by language (Chinese and English), source type (game, film, television series, and animation), and gender. English characters, particularly those from films, constitute the majority in the dataset. In contrast, the Chinese subset exhibits a more balanced distribution across various source types. This diversity provides extensive coverage of voices and contexts for role-playing.

Task Type Distribution User prompts can generally be categorized into three major task categories, which can be further divided into eight subtask types, as illustrated in the figure 4. The description and proportion of each subtask type are as follows:

1. Internal Reasoning

- **Opinion and Emotion Inquiry:** Questions about opinions, emotions, inner feelings, or reactions. (14.73%)
- **Reasoning and Motivation:** Questions about why someone did something, their motivations, or decision-making processes. (14.73%)
- **Reflection and Change:** Questions about how practices/opinions have changed. (9.55%)

2. Experiential Narration

- **Action Description and Retrospection:** Questions about what happened, how something was done, or detailed retrospectives. (14.72%)

- **Skills and Abilities:** Questions about how to accomplish something, difficulties with skills, or ability changes. (6.38%)
- **Historical and Background Description:** Questions about history or background of places/organizations/people. (11.55%)

3. Social Communication

- **Judgment of Others and Events:** Questions about relationships, event evaluations, or interpersonal interactions. (14.73%)
- **Advice and Life Experience:** Questions seeking advice, coping methods, or life experiences. (13.61%)

D Prompt Templates

To ensure high-quality data generation and reliable automatic evaluation, the prompt templates are designed with several principles in mind. First, the prompts explicitly specify the target role and conversational context to encourage consistent role-playing behavior. Second, they provide clear structural instructions to guide the model in producing well-formed dialogue outputs. Third, for evaluation prompts, the criteria are explicitly defined to promote consistent scoring across different evaluation dimensions and reduce ambiguity in the judgment process.

Prompts for Batch Data Generation Figure 6 and Figure 7 respectively present the prompts used for generating single-turn dialogue data for English and Chinese characters. Figure 8 and Figure 9 respectively present the prompts used for generating multi-turn dialogue data for English and Chinese characters.

Prompts for Automated Judgement Figure 10 displays the judgment prompt designed for automatically assessing system-generated audio responses by comparing them with the synthesized reference responses.

E SpeechRole-Agent training settings

Figure 11 presents the distributed training script for fine-tuning the Qwen2.5-Omni-7B model on 8×H100 GPUs. The configuration employs bfloat16 mixed-precision training with a learning

rate of $1e-4$ and gradient accumulation. The implementation includes periodic evaluation (every 500 steps) and checkpoint management. Under this setup, the full training run takes approximately 15 hours on 8×H100 GPUs.

F Additional Experiments

In this section, we provide additional empirical analyses to validate the robustness and reliability of the SpeechRole-Eval framework. Specifically, we investigate the impact of different Text-to-Speech (TTS) reference backends on model evaluation and empirically demonstrate the mathematical stability of our scoring formulation.

F.1 Cross-TTS Robustness Analysis

Although natural human recordings serve as the ideal gold standard for evaluating Speech Role-Playing Agents (SRPAs), acquiring high-quality, character-faithful human speech at scale is cost-prohibitive and lacks reproducibility. Consequently, SpeechRole-Eval utilizes TTS-generated speech as reference baselines, serving as a controlled and scalable proxy. However, a potential risk of using synthetic references is that the evaluation might inadvertently reward models that mimic the specific artifacts, styles, or regularities of a single synthesis pipeline, rather than genuinely reflecting human-like speech role-playing capabilities.

Models	CosyVoice2	E2 TTS	F5-TTS
Alibaba Cloud API	0.8866	0.9057	0.8739
GPT-4o Audio	0.9344	0.9236	0.8962
Llama-3.1-8B	0.9630	0.9338	0.9014
LLaMA-Omni	0.5694	0.5541	0.5590
Qwen3-8B	0.9130	0.9200	0.8646
Mistral-7B	0.8382	0.8402	0.8142
Qwen2.5-Omni	0.5938	0.5975	0.5797
SpeechRole-Agent	0.8261	0.8194	0.8238

Table 4: Performance of models evaluated using three independent TTS reference systems. Scores represent the average across three dimensions: Interaction Ability, Speech Quality and Expressiveness, and Role-Playing Fidelity.

To mitigate system-specific biases and ensure the reliability of the benchmark, we conduct a cross-TTS consistency analysis. In our evaluation protocol, we synthesize reference responses using three diverse, state-of-the-art TTS systems: CosyVoice2, E2 TTS, and F5-TTS. We report the overall performance of the models under different TTS references, which is aggregated by averaging

Reference Pair	Spearman ρ
CosyVoice2 vs. E2 TTS	1.0000
CosyVoice2 vs. F5-TTS	0.9524
F5-TTS vs. E2 TTS	0.9524

Table 5: Spearman’s rank correlation (ρ) of system-level rankings across different TTS reference pairs.

the scores across three key dimensions: Interaction Ability, Speech Quality and Expressiveness, and Role-Playing Fidelity.

As shown in Table 4, the absolute scores of the models remain stable across different reference backends. More importantly, SpeechRole-Eval provides highly consistent model rankings. We calculate the Spearman’s rank correlation coefficient (ρ) between the system-level rankings derived from different TTS references. As presented in Table 5, the overall rankings exhibit high consistency. This indicates that the system rankings are not dominated by a single reference backend.

F.2 Consistency Analysis of Scoring Formulations

Regarding the evaluation metric, SpeechRole-Eval adopts a ratio-based scoring function (A/B), where the candidate score A is normalized by the paired reference score B . Since there are inherent variations in the absolute judge scores across different prompts and reference audios, this design effectively mitigates reference-specific variance, rendering the scores more comparable across diverse test cases.

Reference TTS	Spearman ρ
CosyVoice2	1.0000
E2 TTS	0.9762
F5-TTS	0.9524

Table 6: Spearman’s rank correlation (ρ) between rankings derived from the Ratio-based (A/B) and Difference-based ($A - B$) scoring metrics.

To mathematically justify this protocol and prove that our main conclusions are robust, we compute an alternative difference-based metric ($A - B$). This metric evaluates absolute gaps equally, regardless of the absolute quality of the reference baseline. Table 7 details the model performance under both scoring formulations across the three TTS systems.

Models	Ratio-based (A/B)			Difference-based ($A - B$)		
	CosyVoice2	E2 TTS	F5-TTS	CosyVoice2	E2 TTS	F5-TTS
AlibabaCloudAPI	0.8866	0.9057	0.8739	-1.2940	-1.2937	-1.4554
GPT-4o Audio	0.9344	0.9236	0.8962	-1.1633	-1.2806	-1.3811
Llama-3.1-8B	0.9630	0.9338	0.9014	-0.8632	-0.9920	-1.1017
LLaMA-Omni	0.5694	0.5541	0.5590	-4.1516	-4.3293	-4.2398
Qwen3-8B	0.9130	0.9200	0.8646	-1.1924	-1.1331	-1.4032
Mistral-7B	0.8382	0.8402	0.8142	-1.7168	-1.7101	-1.8284
Qwen2.5-Omni	0.5938	0.5975	0.5797	-4.0083	-4.0067	-4.1304
SpeechRole-Agent	0.8261	0.8194	0.8238	-1.9732	-2.0437	-2.0083

Table 7: Comparison of overall model performance using Ratio-based (A/B) and Difference-based ($A - B$) scoring formulations across three TTS reference systems.

To explicitly quantify the consistency between the two metrics, we calculate the Spearman’s rank correlation (ρ) between the system rankings produced by A/B and $A - B$. As demonstrated in Table 6, the rankings derived from both formulations are highly correlated (Spearman’s $\rho \geq 0.9524$ for all TTS systems). This robust consistency confirms that the overarching conclusions of our benchmark are not an artifact of a specific scoring formulation.

F.3 Recent End-to-End Speech Models

To improve coverage of the rapidly evolving end-to-end speech-model landscape, we additionally evaluate Qwen3-Omni, MiMo-Audio, and Step-Audio-2-mini on both the English and Chinese subsets. We use exactly the same eight metrics, three-reference protocol, and score aggregation as in the main evaluation.

All three models substantially outperform the earlier Qwen2.5-Omni baseline, whose overall scores are 0.568 in English and 0.706 in Chinese. Qwen3-Omni matches SpeechRole-Agent’s English overall score (both 0.801), while SpeechRole-Agent remains stronger in Chinese (0.936 versus 0.841). Qwen3-Omni decreases from 0.862 to 0.739 between single- and multi-turn English evaluation and from 0.933 to 0.749 in Chinese, primarily due to lower multi-turn conversational coherence. These results indicate that recent end-to-end systems have substantially narrowed the gap with cascaded systems, especially in single-turn settings, but performance remains heterogeneous across languages, metrics, and multi-turn interaction.

F.4 Text/Role and Audio/Expressiveness Decomposition

We aggregate IA, CC, PeC, and KC as text/role-oriented dimensions and SF, SN, PC, and EA as audio/expressiveness-oriented dimensions. This is a decomposition of the original speech-based evaluation rather than a transcript-only evaluation. We do not introduce ASR transcripts because doing so would change the speech-to-speech setting and could additionally penalize end-to-end models through transcription errors.

The decomposition shows that cascaded systems with strong text LLMs are stable on the text/role dimensions, while GPT-4o Audio is strong in speech fluency and naturalness. The earlier open-source end-to-end models lag especially in interaction and role fidelity, and SpeechRole-Agent improves over Qwen2.5-Omni in both groups.

F.5 Human Agreement and Uncertainty

The human-LLM study is intended to validate system-level ranking consistency rather than establish the audio-capable LLM as a calibrated replacement for human perceptual judgment. We compute human-human agreement by converting each annotator’s scores into within-sheet model rankings, calculating pairwise agreement for each annotator pair, and then averaging across the three pairs. We estimate uncertainty using 2,000 bootstrap resamples over role/mode sheets.

These results show that expressive speech dimensions are more subjective than semantic dimensions, while human-LLM pairwise preference agreement remains around 71% even for the most challenging Chinese PC and EA metrics. Reliabil-

Language	Model	IA	CC	SF	SN	PC	EA	PeC	KC	Overall
English	Qwen3-Omni	0.798	0.782	0.843	0.835	0.757	0.789	0.756	0.844	0.801
English	MiMo-Audio	0.853	0.915	0.753	0.724	0.716	0.752	0.747	0.827	0.786
English	Step-Audio-2-mini	0.761	0.914	0.896	0.835	0.645	0.655	0.593	0.716	0.752
Chinese	Qwen3-Omni	0.796	0.757	0.900	0.914	0.832	0.854	0.819	0.853	0.841
Chinese	MiMo-Audio	0.805	0.892	0.871	0.859	0.807	0.851	0.827	0.837	0.844
Chinese	Step-Audio-2-mini	0.737	0.890	0.953	0.888	0.695	0.697	0.627	0.725	0.776

Table 8: Results for three recent open-source end-to-end speech models under the same SpeechRole-Eval protocol as the main experiments.

Language	Model	Text/Role	Audio/Expr.
English	Alibaba Cloud API	0.899	0.833
English	GPT-4o Audio	0.883	0.959
English	Qwen3-8B	0.915	0.920
English	Llama-3.1-8B	0.945	0.969
English	Mistral-7B	0.846	0.873
English	LLaMA-Omni	0.649	0.573
English	Qwen2.5-Omni	0.534	0.600
English	SpeechRole-Agent	0.783	0.819
English	SpeechRole-Agent (OOD)	0.756	0.809
Chinese	Alibaba Cloud API	0.955	1.054
Chinese	GPT-4o Audio	0.887	0.920
Chinese	Qwen3-8B	0.782	0.830
Chinese	Llama-3.1-8B	0.775	0.847
Chinese	Mistral-7B	0.639	0.729
Chinese	Qwen2.5-Omni	0.628	0.784
Chinese	SpeechRole-Agent	0.912	0.959
Chinese	SpeechRole-Agent (OOD)	0.911	0.946

Table 9: Decomposition of the original speech-based evaluation into text/role-oriented and audio/expressiveness-oriented metric groups.

Metric	Human-Human	Human-LLM / 95% CI
English PC	0.786	0.929 / [0.821, 1.000]
English EA	0.714	0.815 / [0.679, 0.929]
Chinese PC	0.746	0.714 / [0.619, 0.810]
Chinese EA	0.804	0.714 / [0.619, 0.857]

Table 10: Human-human agreement and human-LLM preference agreement with 95% bootstrap confidence intervals for prosody and emotion.

ity boundaries are especially relevant for SF and SN judgments involving low-level artifacts, and for PC and EA judgments involving subtle pitch, rhythm, pauses, and emotional intensity. We therefore use these dimensions diagnostically and interpret them together with the more reliably ranked interaction and role-fidelity metrics.

F.6 Speaker Similarity Analysis

To examine whether the synthesized role speech preserves target timbre, we compute speaker similarity between each character reference clip and the corresponding synthesized role speech. CAM++ (Wang et al., 2023) is the primary speaker encoder, consistent with our reference-voice selec-

TTS backend	CAM++	ECAPA-TDNN	WavLM
CosyVoice2	0.7558	0.7518	0.9531
F5-TTS	0.7879	0.7800	0.9593
E2 TTS	0.8145	0.8098	0.9699
Overall	0.7861	0.7805	0.9607

Table 11: Speaker similarity between character reference clips and synthesized role speech across three encoders.

tion pipeline, while ECAPA-TDNN (Desplanques et al., 2020) and WavLM (Chen et al., 2022) provide robustness checks. The analysis covers all 98 roles and 2,172 valid samples.

The three encoders show consistent backend-level trends, and CAM++ and ECAPA-TDNN have strong per-role correlation (Spearman’s $\rho = 0.7675$). This supports preservation of speaker identity and timbre in the synthesized references, but does not fully validate acting style, emotional range, pacing, rhythm, or context-dependent delivery.

F.7 Open Audio-Capable Judges

We evaluate two open audio-capable judges under the same pairwise protocol. In a pilot study, Step-Audio-2-mini exhibits a complete ceiling effect: all 128 metric judgments with the standard prompt are (10, 10), and all 64 judgments with a stricter prompt remain (10, 10), providing no ranking signal. We then evaluate MiMo-Audio on the full benchmark. Its scores are also highly compressed: 56.6% of judgments are ties, and 91.9% are either (9, 9) or (8, 9). Its system-ranking correlation with human judgments is low in English and moderate in Chinese, indicating inconsistent cross-lingual reliability.

Under the current protocol, neither open model provides sufficiently discriminative and consistently human-aligned judgments. We retain Gemini-2.5-Pro because it shows stronger agree-

Language	Original vs. no-F5 Spearman ρ
English	0.9762
Chinese	0.9286

Table 12: System-ranking consistency between the original three-reference evaluation and the leave-F5-out setting.

ment with experts across both languages. This selection is based on discriminative ability and agreement with human annotations; we do not treat Gemini’s rankings as ground truth or claim robustness across judges. The prompts, raw evaluation outputs, reference audio, and evaluation pipeline are included in the resource release.

F.8 Leave-F5-out Sensitivity Analysis

Because the open-source cascaded systems use F5-TTS and F5-TTS is also one of the three reference backends, we perform a leave-F5-out control. We remove all F5-TTS references and recompute scores using only CosyVoice2 and E2 TTS.

The resulting overall rankings remain highly consistent with the original three-reference results. Only adjacent systems exchange positions, and the main conclusions remain unchanged. This control reduces concern that same-system reference overlap drives the reported cascaded-system results.

G Resource Release and Reproducibility

To support reproducibility and future research on speech role-playing agents, we release the core resources of SpeechRole at <https://github.com/yuhui1038/SpeechRole>. The release includes the train, out-of-domain, and test splits, covering approximately 111k speech-to-speech dialogue samples and over 1,000 hours of speech data. Each sample contains paired user and role speech, dialogue text, contextual information, and role annotations.

In addition to the dialogue data, the released resources include the associated metadata used during data construction, including role profiles, background descriptions, character lines, and role voice data, as permitted by the applicable licenses and terms of use. We also release the data generation pipeline, including dialogue-generation prompts and settings, speech preprocessing, and TTS-based synthesis scripts.

For SpeechRole-Eval, the release provides the exact three-reference audio, evaluation scripts, scoring protocols, automated-judge prompts, raw evalua-

tion outputs, and human evaluation guidelines. We additionally provide the SpeechRole-Agent fine-tuning recipe and the analysis procedures used for human agreement, speaker similarity, and leave-F5-out evaluation. These resources are intended to enable independent reproduction of our benchmark results and facilitate future extensions to additional models, languages, and evaluation settings.

H Human Evaluation Guidelines

This section describes the annotation protocol used in the human-LLM agreement study.

Evaluation Setup. Three expert annotators independently evaluated system responses generated by eight of the evaluated SRPAs on a randomly sampled half of SpeechRole-Eval. Annotators were fluent in the corresponding evaluation language (English or Chinese) and had prior experience with dialogue or speech quality evaluation.

Each evaluation instance consists of: (1) the user speech input, (2) the target role description, and (3) the generated speech response from a system.

Annotators listen to the generated response and assign scores according to the evaluation dimensions described below.

Scoring Scale. All dimensions are rated on a 1–5 Likert scale:

- **5:** Excellent — fully satisfies the evaluation criterion
- **4:** Good — minor imperfections but overall strong performance
- **3:** Acceptable — noticeable issues but still reasonable
- **2:** Poor — significant problems affecting quality
- **1:** Very poor — fails to meet the criterion

Evaluation Dimensions. Instruction Adherence (IA). Measures whether the response follows the user’s request or instruction.

Conversational Coherence (CC). Measures whether the response is logically consistent with the dialogue context.

Speech Fluency (SF). Evaluates whether the speech is fluent without unnatural pauses or disfluencies.

Speech Naturalness (SN). Evaluates how natural and human-like the synthesized speech sounds.

Prosody Consistency (PC). Measures whether the intonation and rhythm are consistent with the intended speaking style.

Emotion Appropriateness (EA). Measures whether the emotional tone matches the dialogue context and role.

Personality Consistency (PeC). Evaluates whether the response maintains the personality traits of the target role.

Knowledge Consistency (KC). Measures whether the response uses knowledge consistent with the role's background or identity.

Annotation Procedure. Annotators evaluate each response independently without access to system identities or automatic evaluation scores. All instances are presented in randomized order to minimize ordering bias.

Final human scores are computed by averaging ratings across annotators and evaluation instances for each system and metric.

Role Name	Category	Language	Gender	Source	Split
hutao	Game	Chinese	Female	Genshin Impact	train/test
raidenShogun	Game	Chinese	Female	Genshin Impact	train/test
wanderer	Game	Chinese	Male	Genshin Impact	train/test
ayaka	Game	Chinese	Female	Genshin Impact	dev/test
zhongli	Game	Chinese	Male	Genshin Impact	train/test
liyunlong	TV	Chinese	Male	Drawing Sword	dev/test
wangduoyu	Movie	Chinese	Male	Hello Mr. Billionaire	train/test
weixiaobao	TV	Chinese	Male	The Deer and the Cauldron	train/test
jiumuzhi	TV	Chinese	Male	Demi-Gods and Semi-Devils	train/test
wangyuyan	TV	Chinese	Female	Demi-Gods and Semi-Devils	train/test
Luna	Movie	English	Female	Harry Potter	dev/test
Penny	TV	English	Female	The Big Bang Theory	dev/test
zhangwuji	TV	Chinese	Male	The Heaven Sword and Dragon Saber	train/test
zhaomin	TV	Chinese	Female	The Heaven Sword and Dragon Saber	train/test
huangrong	TV	Chinese	Female	The Legend of the Condor Heroes	train/test
guojing	TV	Chinese	Male	The Legend of the Condor Heroes	dev/test
wukong	TV	Chinese	Male	Journey to the West	train/test
HAL 9000	Movie	English	Male	2001: A Space Odyssey	train/test
Colonel Nathan R. Jessep	Movie	English	Male	A Few Good Men	train/test
Antonio Salieri	Movie	English	Male	Amadeus	train/test
Stifler	Movie	English	Male	American Pie	train/test
Paul Vitti	Movie	English	Male	Analyze That	train/test
Alvy Singer	Movie	English	Male	Annie Hall	train/test
Violet Weston	Movie	English	Female	August: Osage County	train/test
Willie Soke	Movie	English	Male	Bad Santa	train/test
Gaston	Animation	English	Male	Beauty and the Beast	train/test
The Dude	Movie	English	Male	The Big Lebowski	train/test
Paul Conroy	Movie	English	Male	Buried	train/test
Truman Capote	Movie	English	Male	Capote	train/test
Mater	Animation	English	Male	Cars 2	train/test
Andrew Detmer	Movie	English	Male	Chronicle	train/test
Coriolanus	Movie	English	Male	Coriolanus	train/test
John Keating	Movie	English	Male	Dead Poets Society	dev/test
Wade Wilson	Movie	English	Male	Deadpool	dev/test
Jim Morrison	Movie	English	Male	The Doors	train/test
Queen Elizabeth I	Movie	English	Female	Elizabeth: The Golden Age	dev/test
Jeff Spicoli	Movie	English	Male	Fast Times at Ridgemont High	train/test
Fred Flintstone	Animation	English	Male	The Flintstones	train/test
Freddy Krueger	Movie	English	Male	Freddy Vs.Jason	train/test
Tyrion Lannister	TV	English	Male	Game of Thrones	train/test
James Brown	Movie	English	Male	Get on Up	train/test
Walt Kowalski	Movie	English	Male	Gran Torino	train/test
John Coffey	Movie	English	Male	The Green Mile	train/test
Theodore Twombly	Movie	English	Male	Her	dev/test
Gregory House	TV	English	Male	House M.D.	dev/test
Sonny	Movie	English	Male	I, Robot	train/test
Colonel Hans Landa	Movie	English	Male	Inglourious Basterds	train/test
Judge Dredd	Movie	English	Male	Judge Dredd	dev/test
Juno MacGuff	Movie	English	Female	Juno	train/test
Professor G.H. Dorr	Movie	English	Male	The Ladykillers	train/test

Table 13: Summary of character attributes and dataset partitioning (part 1).

Role Name	Category	Language	Gender	Source	Split
Fletcher Reede	Movie	English	Male	Liar Liar	train/test
Abraham Lincoln	Movie	English	Male	Lincoln	train/test
Frank T.J. Mackey	Movie	English	Male	Magnolia	train/test
Leonard Shelby	Movie	English	Male	Memento	train/test
Harvey Milk	Movie	English	Male	Milk	train/test
Randle McMurphy	Movie	English	Male	One Flew Over the Cuckoo's Nest	train/test
Jack Sparrow	Movie	English	Male	Pirates of the Caribbean: Dead Man's Chest	dev/test
John Dillinger	Movie	English	Male	Public Enemies	train/test
Lestat de Lioncourt	Movie	English	Male	Queen of the Damned	train/test
Tyler Hawkins	Movie	English	Male	Remember Me	dev/test
James Carter	Movie	English	Male	Rush Hour 2	train/test
Jigsaw	Movie	English	Male	Saw	train/test
John Doe	Movie	English	Male	Se7en	train/test
Sherlock Holmes	Movie	English	Male	Sherlock Holmes	dev/test
Shrek	Animation	English	Male	Shrek	train/test
Pat Solitano	Movie	English	Male	Silver Linings Playbook	train/test
Karl Childers	Movie	English	Male	Sling Blade	train/test
Bruno Antony	Movie	English	Male	Strangers on a Train	train/test
Seth	Movie	English	Male	Superbad	train/test
Caden Cotard	Movie	English	Male	Synecdoche, New York	train/test
Travis Bickle	Movie	English	Male	Taxi Driver	train/test
Stanley Ipkiss	Movie	English	Male	The Mask	dev/test
Lyn Cassady	Movie	English	Male	The Men Who Stare at Goats	train/test
Michael Scott	TV	English	Male	The Office	dev/test
Robert Angier	Movie	English	Male	The Prestige	dev/test
Dr. Frank-N-Furter	Movie	English	Male	The Rocky Horror Picture Show	train/test
Jack Torrance	Movie	English	Male	The Shining	train/test
Tom Ripley	Movie	English	Male	The Talented Mr. Ripley	train/test
D_Artagnan	Movie	English	Male	The Three Musketeers	train/test
Thor	Movie	English	Male	Thor: Ragnarok	train/test
James Bond	Movie	English	Male	Tomorrow Never Dies	dev/test
Mark Renton	Movie	English	Male	Trainspotting	train/test
David Aames	Movie	English	Male	Vanilla Sky	train/test
Rorschach	Movie	English	Male	Watchmen	train/test
Jordan Belfort	Movie	English	Male	The Wolf of Wall Street	train/test
Logan	Movie	English	Male	X-Men Origins: Wolverine	dev/test
Judy Hopps	Animation	English	Female	Zootopia	train/test
Doctor Who	TV	English	Male	Doctor Who	train/test
Raylan Givens	TV	English	Male	Justified	train/test
Mary Sibley	TV	English	Female	Salem	train/test
Lucifer Morningstar	TV	English	Male	Lucifer	train/test
Twilight Sparkle	Animation	English	Female	My Little Pony: Friendship is Magic	dev/test
Oliver Queen	TV	English	Male	Arrow	train/test
Klaus Mikaelson	TV	English	Male	The Originals	train/test
Queen Catherine	TV	English	Female	Reign	train/test
Dr. Hannibal Lecter	TV	English	Male	Hannibal	train/test
Coach Eric Taylor	Movie	English	Male	Friday Night Lights	train/test
yaemiko	Game	Chinese	Female	Genshin Impact	train/test

Table 14: Summary of character attributes and dataset partitioning (part 2).

Prompt for Batch Generation of Single-Turn Data, English

System: Your task is to generate {BATCH_SIZE} single-turn dialogues between a user and the character {role_name} from the script {script_name}. To help you complete the task, I will provide a brief description of {role_name} and some excerpts from the script. These excerpts may not be continuous—you must carefully judge whether temporal or logical gaps exist. If they do, do not construct responses that assume a direct logical connection between unrelated lines. Please follow the instructions below strictly:

1. Each dialogue should consist of a single Q&A turn, forming a natural, in-character, and context-aware role-playing interaction.

2. The dialogue should be in the following format:

```
{{
  "user": "User's question",
  "role": "Character's response"
}}
```

3. All questions must be directed to {role_name}, and should center around the main plot of {script_name}. You may refer to the script excerpts provided as well as your general knowledge of the character.

4. All outputs must be formatted as a JSON array. Each dialogue should be an object with two fields:

- dialogue: a list containing exactly one Q&A pair;
- context: the script excerpts that were used or referenced in generating this dialogue.

5. You must generate {BATCH_SIZE} such dialogues.

6. All character responses must remain faithful to their personality and the context of the story.

7. Ensure each question and answer brings unique value, avoiding repetitive content across samples.

Now I will give you one example. This example is not from {script_name} and not about {role_name}, but serves to illustrate the format and completeness of a single-turn dialogue:

[Example]

```
[
  {{
    "dialogue": [
      {{
        "user": "Tony, why did you stop manufacturing weapons even though it made your company so successful?",
        "role": "Because I saw firsthand what my weapons were doing in the wrong hands. That ambush in Afghanistan—it changed everything. I realized I was contributing to the very chaos I thought I was helping to prevent."
      }}
    ],
    "context": "TONY STARK: I saw young American soldiers killed by the very weapons I created to protect them. I can't ignore that.\n(Flashback: Tony being ambushed by Stark Industries missiles in Afghanistan)\nTONY: I'm not a hero. I'm just trying to fix what I broke."
  }}
]
```

Prompt:

[Character Name and Description]

The character is {role_name} from {script_name}, the character description is: {role_description}

[Script Content]

{script}

[JSON format]

```
[
  {{
    "dialogue": [
      {{
        "user": "",
        "role": ""
      }}
    ]
    "context": "",
  }},
  .....
]
```

[Question Design ({BATCH_SIZE} single-turn dialogues, no semantic repetition, all questions must be directed to {role_name}; each dialogue must consist of exactly 1 user-role interaction pair, and each Q&A must be self-contained, logically sound, and in-character)]

Figure 6: Prompt structure for batch single-turn English dialogue generation.

Prompt for Batch Generation of Single-Turn Data, Chinese

System: 你的任务是设计{BATCH_SIZE}段用户与{script_name}中的{role_name}之间的单轮对话。为了帮助你更好地完成任务，我会提供{role_name}的简要人物描述和部分剧本内容。这些剧本片段可能不连续，你需要根据上下文判断是否存在时间或逻辑跳跃。如果不连续，请不要构建上下句之间有强逻辑衔接的对话。

请严格遵循以下设计规则：

1. 每段对话应为一轮完整问答，由用户向{role_name}提问，角色作答，形成自然、符合剧情背景的角色扮演对话。

2. 每轮结构为：

```
{{
  "user": "用户提的问题",
  "role": "角色的回答"
}}
```

3. 所有问题必须是用户直接向{role_name}提问，并围绕{script_name}的核心情节展开。你可以结合剧本片段和你对角色的常识知识设计问题。

4. 所有输出必须是一个符合JSON结构的数组，每段对话为一个对象，对象中包含两个字段：

- **dialogue:** 一个包含一组问答的列表（即仅一个 user-role 对）；

- **context:** 该段对话中所参考或引用的剧本原文内容。

5. 你需要输出共{BATCH_SIZE}段对话。

6. 所有角色回答必须符合人设，语言风格需贴合剧本风格与时代背景。

7. 对话应尽量避免重复提问或答复，确保每段都有独立的信息价值。

接下来我会给你一个样例，这个样例与{role_name}和{script_name}无关，仅用于展示格式和完整性：

[样例]

```
[
  {{
    "dialogue": [
      {{
        "user": "姜子牙，你为什么要讨伐商纣王？这是一个冒天下之大不韪的决定。",
        "role": "我并非一时冲动起兵。商纣王荒淫无道，沉迷酒色，任用奸佞，焚书坑儒，滥杀忠良，百姓民不聊生。作为一名辅佐天下的臣子，我若对他的暴行视而不见，那才是失职。讨伐他，并非为我一人之私，而是为了拯救黎民百姓于水火。"
      }}
    ],
    "context": "姜子牙：「纣王以鹿台为宫，役民无数，掠财千金，尽为享乐。我观民间疲敝，不忍再等。」\n姬发：「父王身死国灭，子民苦难，我夜不能寐。师父，若你不举兵，我愿亲率义军，誓为天下请命。」\n姜子牙沉思片刻：「兵者不祥之器，非不得已不得动之。然今日不动，百姓将无明日。此战，虽难，却势在必行。」"
  }}
]
```

Prompt:

[Character Name and Description]

The character is {role_name} from {script_name}, the character description is: {role_description}

[Script Content]

{script}

[JSON format]

```
[
  {{
    "dialogue": [
      {{
        "user": "",
        "role": ""
      }}
    ],
    "context": "",
  }},
  .....
]
```

[Question Design ({BATCH_SIZE} single-turn dialogues, no semantic repetition, all questions must be directed to {role_name}; each dialogue must consist of exactly 1 user-role interaction pair, and each Q&A must be self-contained, logically sound, and in-character)]

Figure 7: Prompt structure for batch single-turn Chinese dialogue generation.

Prompt for Batch Generation of Multi-Turn Data, English

System: Your task is to generate {BATCH SIZE} multi-turn dialogues between a user and the character {role name} from the script {script name}. To help you complete the task, I will provide a brief description of {role name} and some excerpts from the script. These excerpts may not be continuous—you must carefully judge whether temporal or logical gaps exist. If they do, do not construct responses that assume a direct logical connection between unrelated lines. Please follow the instructions below strictly:

1. Each dialogue should consist of 2 to 4 turns of Q&A, forming a natural, in-character, and context-aware role-playing conversation.

2. Each turn should be in the following format, alternating between user questions and character answers:

```
{{
  "user": "User's question",
  "role": "Character's response"
}}
```

3. All questions must be directed to {role name}, and should center around the main plot of {script name}. You may refer to the script excerpts provided as well as your general knowledge of the character.

4. All outputs must be formatted as a JSON array. Each dialogue should be an object with two fields:

- dialogue: a list of Q&A turns in order;

- context: the script excerpts that were used or referenced in generating this dialogue.

5. You must generate {BATCH SIZE} such dialogues.

6. The number of turns per dialogue can vary (2–4), but the exchange must be coherent and the responses must stay in character.

7. Aim to produce 3 or 4 turns per dialogue when appropriate, as long as the flow remains logical and engaging.

Now I will give you one example. This example is not from {script name} and not about {role name}, but serves to illustrate the format and definition of completeness.

[Example]

```
[
  {{
    "dialogue": [
      {{
        "user": "Tony, why did you stop manufacturing weapons even though it made your company so successful?",
        "role": "Because I saw firsthand what my weapons were doing in the wrong hands. That ambush in Afghanistan—it changed everything. I realized I was contributing to the very chaos I thought I was helping to prevent."
      }},
      .....
    ],
    "context": "TONY STARK: I saw young American soldiers killed by the very weapons I created to protect them. I can't ignore that.\nRHODEY: You're not just a weapons manufacturer anymore.\n ..... "
  }}
]
```

Prompt:

[Character Name and Description]

The character is {role name} from {script name}, the character description is: {role description}

[Script Content]

{script}

[JSON format]

```
[
  {{
    "dialogue": [
      {{
        "user": "",
        "role": ""
      }},
      .....
    ],
    "context": "",
  }},
  .....
]
```

[Question Design ({BATCH_SIZE} multi-turn dialogues, no semantic repetition, all questions must be directed to {role_name}; each dialogue must consist of 2 to 4 user-role interaction rounds, and you are encouraged to generate 3 or 4 rounds whenever possible, as long as the continuation remains logical and in-character)]

Figure 8: Prompt structure for batch multi-turn English dialogue generation.

Prompt for Batch Generation of Multi-Turn Data, Chinese

System: 你的任务是设计 {BATCH_SIZE} 段用户与 {script_name} 中的 {role_name} 之间的多轮对话。为了帮助你更好地完成任务，我会提供 {role_name} 的简要人物描述和部分剧本内容。这些剧本片段可能不连续，你需要根据上下文判断是否存在时间或逻辑跳跃。如果不连续，请不要构建上下句之间有强逻辑衔接的对话。请严格遵循以下设计规则：

1. 每段对话应为一组 2 到 4 轮的连续问答，形成自然的、符合剧情背景的角色扮演对话。
2. 对话应由用户提问、角色回答交替构成，每轮结构为：

```
{{
  "user": "用户提的问题",
  "role": "角色的回答"
}}
```

3. 请确保所有问题都是用户直接向 {role_name} 提问，并围绕 {script_name} 的核心情节展开。你可以结合剧本片段和你对角色的常识知识设计问题。

4. 所有输出必须是一个符合 JSON 结构的数组，每段对话为一个对象，对象中包含两个字段：

- dialogue: 一个按轮次排列的问答列表；
- context: 该段对话中所参考或引用的剧本原文内容。

5. 你需要输出共 {BATCH_SIZE} 段对话。

6. 每段对话的轮次不必相同（可为 2-4 轮），但都应逻辑自然，角色回答应符合人设。

7. 请尽量生成 3 或 4 轮对话，只要情节自然、角色一致即可，不必刻意缩短对话。

接下来我会给你一个样例，这个样例与 {role_name} 和 {script_name} 无关，主要作用是让你明白完整性的定义以及输出的格式。

[样例]

```
[
  {{
    "dialogue": [
      {{
        "user": "姜子牙，你为什么要讨伐商纣王？这是一个冒天下之大不韪的决定。",
        "role": "我并非一时冲动起兵。商纣王荒淫无道，沉迷酒色，任用奸佞，焚书坑儒，滥杀忠良，百姓民不聊生。作为一名辅佐天下的臣子，我若对他的暴行视而不见，那才是失职。讨伐他，并非为我一人之私，而是为了拯救黎民百姓于水火。"
      }},
      .....
    ],
    "context": "姜子牙：「纣王以鹿台为宫，役民无数，掠财千金，尽为享乐。我观民间疲敝，不忍再等。」\n姬发：「父王身死国灭，子民苦难，我夜不能寐。师父，若你不举兵，我愿亲率义军，誓为天下请命。」\n....."
  }}
]
```

Prompt:

[Character Name and Description]

The character is {role name} from {script name}, the character description is: {role description}

[Script Content]

{script}

[JSON format]

```
[
  {{
    "dialogue": [
      {{
        "user": "",
        "role": ""
      }},
      .....
    ]
    "context": "",
  }},
  .....
]
```

[Question Design ({BATCH_SIZE} multi-turn dialogues, no semantic repetition, all questions must be directed to {role_name}; each dialogue must consist of 2 to 4 user-role interaction rounds, and you are encouraged to generate 3 or 4 rounds whenever possible, as long as the continuation remains logical and in-character)]

Figure 9: Prompt structure for batch multi-turn Chinese dialogue generation.

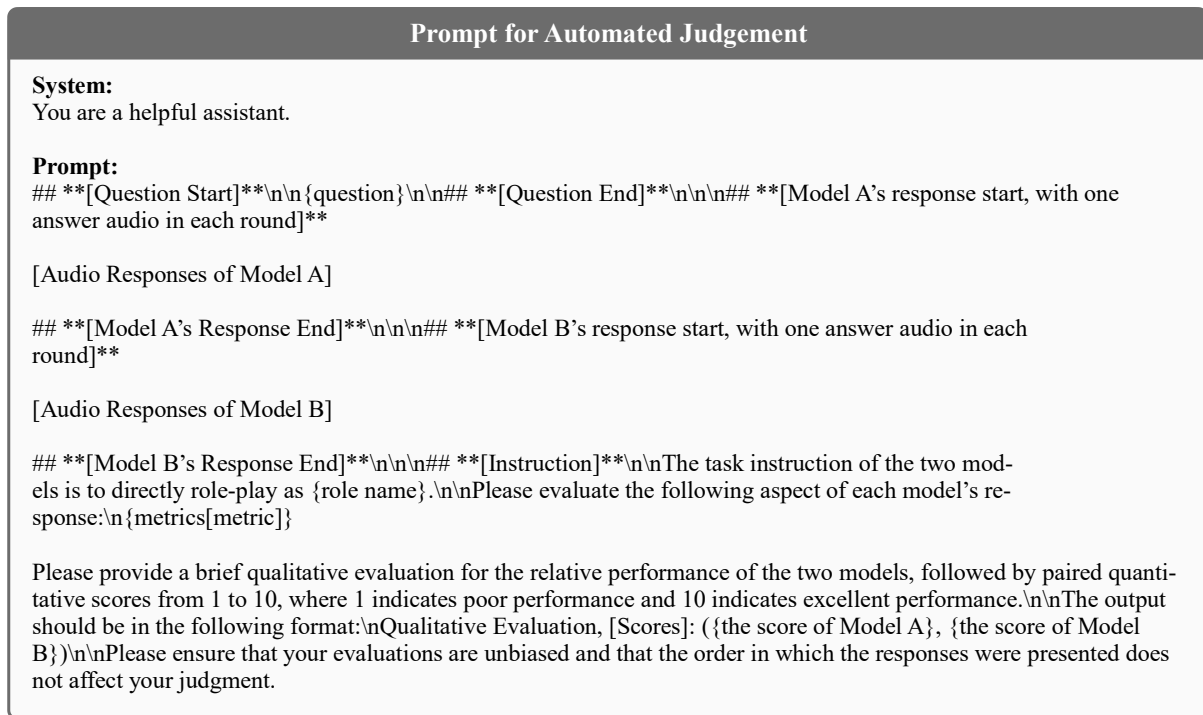


Figure 10: Prompt structure for Automated Judgement.

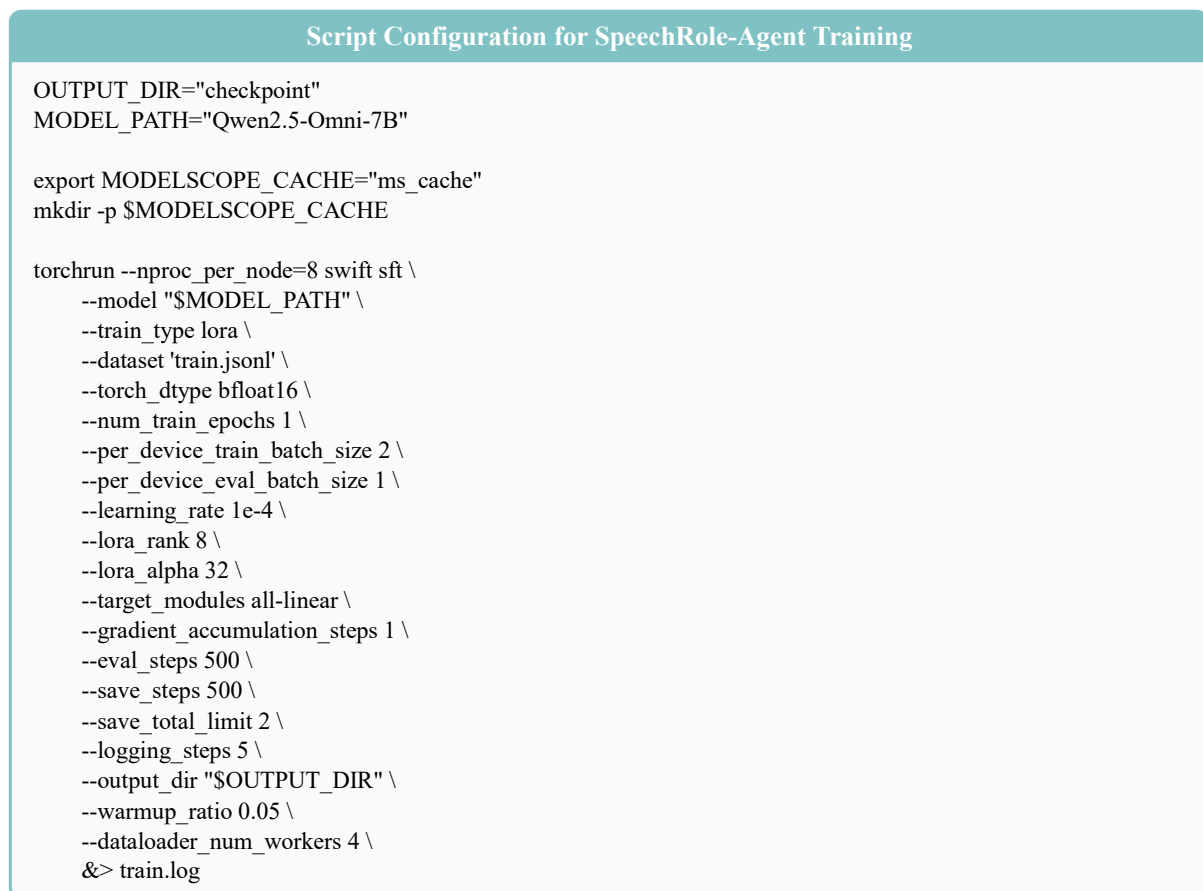


Figure 11: Training script configuration for fine-tuning the Qwen2.5-Omni-7B model.