Speech-DRAME: Human-Aligned Evaluation for Speech Role-Play

Anonymous authors
Paper under double-blind review

Abstract

Role-play has become a key testbed for generative models, expanding from text-only dialogue to multimodal interaction. Extending role-play to speech captures prosody, emotion, and delivery, but also poses new evaluation challenges. Although audio large language models (ALLMs) used as zero-shot judges already exhibit non-trivial capability, current pipelines built on them suffer from three persistent issues: limited sensitivity to paralinguistic cues, collapse of diverse performance aspects into coarse scores, and reliance on synthetic speech references that fail to reflect real-world roles. We present **Speech-DRAME**, a unified framework that addresses these limitations through: (i) *Speech-DRAME-EvalBench*, an evaluation benchmark with bilingual human-annotated data, contrastive negative samples, and protocols for training and testing speech evaluation models (`SEval`), and (ii) `DRAME-Eval`, a fine-tuned `SEval` that builds on top of strong ALLM backbones to further improve agreement with human raters. Speech-DRAME distinguishes between two complementary evaluation strategies: *Archetype Evaluation*, a top-down approach measuring adherence to broad role archetypes, and *Realism Evaluation*, a bottom-up approach grounded in real human speech that emphasizes nuanced role quality. On the *in-distribution* test sets, where the spoken responses largely come from the same synthesis bank that supplies `DRAME-Eval`'s training data, fine-tuning raises Pearson correlation with human ratings from 0.480 to 0.629 on archetypes and from 0.390 to 0.625 on realism over the strongest zero-shot ALLM-judge baseline (few-shot prompting reaches 0.495 and 0.432, respectively). On a separate *real-recording* test set drawn entirely from professional and amateur human acting, however, correlation remains modest: `DRAME-Eval` reaches an average Pearson correlation of 0.247, with the weakest results on identity- and trait-related dimensions. We report this synthetic-to-real gap as a measured open problem rather than a solved one, and treat closing it as the central direction for future work.

1 Introduction

Role-play has been recently recognized as a cornerstone of interactive systems, from initial text-based focuses to the growing ecosystem of multimodal dialogue agents (Chen et al., 2024a; Tseng et al., 2024; Dai et al., 2025b; Ma et al., 2024; Jiang et al., 2025). With the rapid advancement of large language models (LLMs), role-play has evolved beyond scripted interactions into dynamic character-driven conversations, supporting applications in education, entertainment, and human–AI collaboration (Wu et al., 2024; Agatsuma et al., 2024; Fung & Laing, 2024). Extending role-play to speech-based interaction offers an even richer communication channel, capturing prosody, emotion, and character consistency in ways that text alone cannot (Ding et al., 2025; Huang et al., 2025a).

Evaluating speech role-play, however, remains a major challenge. Current pipelines often rely on ALLMs as judges, typically in a zero-shot setting (Huang et al., 2025c; Chiang et al., 2025; Jiang et al., 2025; Zhang et al., 2025a). Recent ALLM judges already exhibit non-trivial capability for spoken role-play assessment, and we adopt them as strong baselines throughout this work. Yet, even with these capable backbones, current zero-shot judging pipelines fall short along three persistent axes:

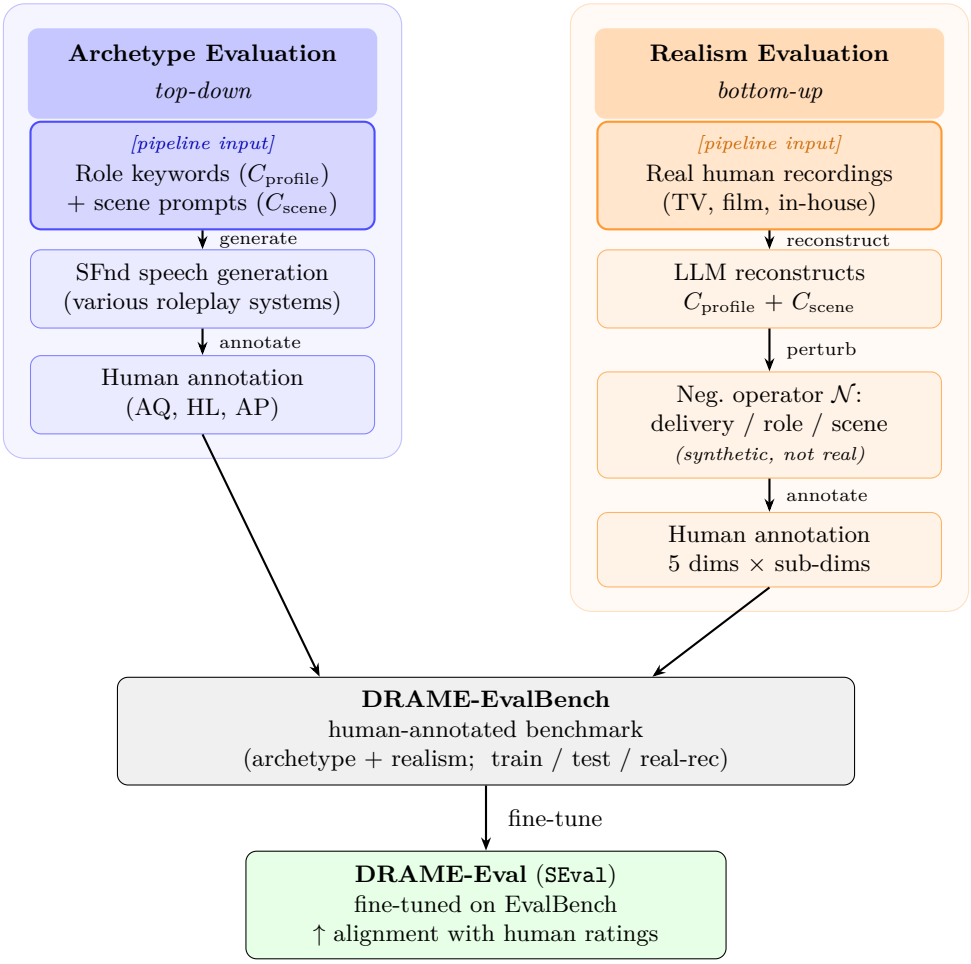


Figure 1: Speech-DRAME formalizes speech role-play with a dual evaluation strategy: **Archetype Evaluation** (top-down, stereotype-driven) and **Realism Evaluation** (bottom-up, human-grounded). The colored header band at the top of each lane names the evaluation paradigm; it is *not* a pipeline stage. Darker-bordered boxes labeled *[pipeline input]* identify where data enters each pipeline: role keywords (C_{profile}) + scene prompts (C_{scene}) for archetype, and real human recordings for realism. Archetype annotation covers three dimensions: Audio Quality (AQ), Human-likeness (HL), and Appropriateness (AP). Both pipelines feed into **DRAME-EvalBench**, which is used to train **DRAME-Eval**, a fine-tuned `SEval` with improved alignment to human ratings.

- they lack sensitivity to paralinguistic cues such as intonation, pacing, and emotional dynamics that define role delivery (Zhang et al., 2025a; Huang et al., 2025c; Chiang et al., 2025; Zhou et al., 2025),
- they collapse diverse aspects of performance into limited coarse dimensions, obscuring critical dimensions of role-play quality (Chiang et al., 2025; Huang et al., 2025c), and
- they depend on synthetic speech references rather than real human speech (Zhang et al., 2025a; Jiang et al., 2025), and rarely test the judge against *contrastive negative* conditions in which the spoken delivery, character identity, or scene context have been deliberately mismatched.

The third axis is particularly underexplored: with positive-only synthetic data, it is hard to tell whether a judge truly attends to paralinguistic cues or simply pattern-matches the transcript. We therefore make contrastive negative samples a first-class component of our benchmark (Sec. 5), in line with the broader trend in NLP and vision-language evaluation (Nie et al., 2020; Shekhar et al., 2017; Thrush et al., 2022).

To overcome these limitations, we present **Speech-DRAME** (Speech Detailed Role-play Assessment with Modeling and Evaluation), a human-aligned evaluation framework that moves beyond synthetic or judge-only pipelines. Speech-DRAME is distinguished by three aspects: [i] a dual evaluation paradigm that balances archetype-driven generality with human-grounded realism (see Sec. 3.2)[1], [ii] human annotations that provide perceptual grounding absent in prior resources (see Sec. 4 and 5), and [iii] alignment between evaluation models (`SEval`) and human perception, which enables systematic analysis of the evaluation models themselves and lays the groundwork for downstream benchmarking of speech foundation models (`SFnd`) in future work (see Sec. 6). Together, these aspects are intended to make Speech-DRAME a reproducible framework for measuring speech role-play.

As illustrated in Fig. 1, the present paper focuses on the *evaluation* side of speech role-play and contributes at two levels: (i) **DRAME-EvalBench**, an evaluation benchmark consisting of human-annotated datasets for archetype and realism role-play, including curated contrastive negative samples that probe whether judges attend to fine-grained delivery, role identity, and scene cues; and (ii) **DRAME-Eval**, a fine-tuned evaluation model (`SEval`) trained on EvalBench that, while building on the broad capability of recent ALLM judges, further improves agreement with human ratings. We deliberately scope this paper to the human-aligned evaluation problem; using `DRAME-Eval` as an automatic judge to rank `SFnd` systems on a held-out role-play benchmark is left to future work. Our contributions are three-fold:

- We introduce a **formal task definition** of speech role-play and its evaluation via speech foundation models (`SFnd`) for generation and speech evaluation models (`SEval`) for assessment, unifying linguistic and paralinguistic dimensions.
- We construct **DRAME-EvalBench**, a bilingual evaluation benchmark with human annotations under both archetype- and realism-based strategies, supporting the training and testing of `SEval` models. To stress-test perceptual grounding, EvalBench also includes contrastive *negative* samples that perturb the scene, the character profile, or the spoken delivery (Sec. 5).
- We propose **DRAME-Eval**, a fine-tuned `SEval` trained on EvalBench. While zero- and few-shot ALLM judges already exhibit non-trivial capability for speech role-play assessment, DRAME-Eval further improves their agreement with human ratings on the in-distribution test set, demonstrating the value of human supervision on top of strong ALLM backbones.[2]

Through this work, we aim to move from acting in text to acting in speech, equipping the community with the models, evaluation strategies, and benchmarks needed to measure and improve speech role-play.

2 Related Works

Text-based role-play evaluation. Early work on role-play focuses on text-only dialogue settings. Benchmarks such as RoleLLM, CharacterEval, RoleEval, RoleMRC, and CharacterBox (Wang et al., 2024a; Tu et al., 2024; Shen et al., 2023; Wang et al., 2025d; Lu et al., 2025) examine persona fidelity and character consistency, typically relying on LLM-as-judge pipelines for scalability. InCharacter (Wang et al., 2024b) probes personality fidelity through psychological-interview prompts, while PingPong (Gusev, 2024) grounds role-play evaluation in user-emulator multi-model interactions. More recent benchmarks emphasize bilingual, multi-turn, and user-centric scenarios (e.g., RMTBench (Xiang et al., 2025)) and large-scale customizable persona training (OpenCharacter (Wang et al., 2025g)), while atomic-level evaluation of persona fidelity (Shin et al., 2025) highlights that single-score judgments hide subtle out-of-character drift in long generations. MMRole (Dai et al., 2025a) extends role-play evaluation to the multimodal (vision+text) regime, anticipating the cross-modal challenges that also arise in speech. However, studies such as PersonaEval (Zhou et al., 2025) reveal fundamental reliability issues: even in pure text, judge models struggle with

[1] *Archetype Evaluation* (top-down) assesses whether synthesized speech fits broad, culturally shared role schemas — e.g., whether a "firefighter" response sounds appropriately authoritative and calm. *Realism Evaluation* (bottom-up) is anchored in real human recordings and assesses fine-grained delivery quality: prosodic dynamics, emotional expressiveness, character consistency, contextual relevance, and semantic match. The two strategies are complementary: the former enables scalable, schema-driven benchmarking; the latter provides perceptual grounding in authentic human acting.

[2] Relevant codes, pre-trained models, and datasets will be open-sourced by the time of acceptance.

speaker tracking and contextual grounding. These limitations become more pronounced in speech, where prosody, emotional nuance, and vocal style quality are essential components of character portrayal.

Speech-native role-play benchmarks. Recent efforts extend role-play evaluation to spoken interaction. OmniCharacter-10K (Zhang et al., 2025a) emphasizes character voice traits, while resources such as Vocal-Bench, SOVA, mTalk, and VoiceBench (Liu et al., 2025; Hou et al., 2025a; Du et al., 2025; Chen et al., 2024b) broaden evaluation to vocal communication, assistant ability, and stylistic realism. In parallel, omni-capable SFMs such as StepAudio and KimiAudio (Huang et al., 2025a; Ding et al., 2025) include role-play scenarios in their tests, alongside the recent wave of end-to-end spoken dialogue models such as Moshi (Défossez et al., 2024). Closest to our aims is a fast-growing cluster of role-play–specific speech benchmarks: Speech-Role (Jiang et al., 2025) provides a 111k-dialogue end-to-end speech-to-speech benchmark for role-playing agents; VStyle (Zhan et al., 2025) offers a bilingual benchmark whose four categories explicitly include a role-play split alongside acoustic-attribute, instruction-following, and implicit-empathy splits; VoxRole (Wu et al., 2025) curates 13k character-grounded film dialogues from 1,228 unique characters with paralinguistic-appropriateness metrics; and AudioRole (Li et al., 2025) contributes a 1k+-hour TV-series dataset with the dual-aspect ARP-Eval framework. Concurrent generation-side efforts such as Voila (Shi et al., 2025c) build voice-language foundation models for real-time voice role-play with persona-aware speaker control, and PersonaPlex (Roy et al., 2026) extends Moshi-style full-duplex models with hybrid voice+text persona prompts. Reinforcement-learning–based shaping of role-playing behavior in audio LLMs has also been explored (Fu et al., 2026). Yet these resources either rely heavily on LLM-generated dialogues and TTS synthesis (e.g., SpeechRole, VStyle), or, even when anchored in real film/TV audio (e.g., VoxRole, AudioRole), score outputs with proprietary judge models and fully automatic metrics, releasing no training-grade human ratings and offering coarse, often opaque rubrics.

Audio-LLM evaluation infrastructure. Beyond role-play, several recent benchmarks evaluate audio LLMs more broadly along dimensions that overlap with realism evaluation. AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025a) provide universal evaluation across speech, sound, and music understanding; Dynamic-SUPERB Phase-2 (Huang et al., 2025b; 2024a) expands the instruction-tuning evaluation surface to 180 tasks. SD-Eval (Ao et al., 2024) explicitly targets paralinguistic and environmental information in spoken dialogue understanding, and CP-Bench (Wang et al., 2025e) benchmarks contextual paralinguistic reasoning in speech-LLMs on in-the-wild data. A parallel strand evaluates voice assistants and conversational agents: VoiceAssistant-Eval (Wang et al., 2025c) jointly tests listening, speaking, and viewing across $\sim$10k instances; MULTIVOX (Selvakumar et al., 2025) probes voice-assistant abilities across diverse styles; SOVA-Bench (Hou et al., 2025b) systematizes the speech conversation ability of voice assistants; and StyleBench (Zhao et al., 2026) measures fine-grained speaking-style intensity control across emotion, speed, volume, and pitch. WildSpeech-Bench (Zhang et al., 2025b) and ParaS2S (Yang et al., 2025b) extend the evaluation target to end-to-end speech-to-speech models with paralinguistic awareness as an explicit axis; ParaS2S in particular pairs a TTS-generated query benchmark with a GRPO-based alignment framework that distills the benchmark's GPT-based scoring pipeline into a reward model, illustrating a complementary path in which the judge is treated as a learned reward rather than as a target for human-aligned supervision. These efforts are complementary to Speech-DRAME: they evaluate generic understanding, voice-assistant behavior, or general dialogue, whereas we evaluate role-play–specific perceptual fit, with human-rated training data and contrastive negatives constructed for that purpose.

Automatic judging with audio-capable LLMs. A line of work explores automatic assessment using ALLMs. Systems such as AudioJudge and Audio-Aware LLMs as Judges (Manakul et al., 2025; Chiang et al., 2025) evaluate emotion, style, and delivery; Auto-ATT (Wang et al., 2025h) scales this to Turing-test-style assessments. In parallel, a complementary thread treats the audio LLM itself as a learned quality predictor: ALLD (Chen et al., 2025) aligns an audio LLM via distillation to produce descriptive, MOS-consistent speech-quality judgments, and SpeechLLM-as-Judges (Wang et al., 2025b) formulates general-purpose, interpretable speech-quality evaluation as an LLM-as-judge problem. While promising for rapid scoring, these approaches show high prompt sensitivity, weak capture of paralinguistic cues, and limited interpretability. Judge-only pipelines further lack stable criteria, as model versions, prompts, and hidden reasoning steps change over time. Speech-DRAME differs from this line of work in two ways: (i) we condition each judgment on a

structured *role* ($\boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}}$) rather than an unconditional quality target, and (ii) we explicitly stress-test the judge with the contrastive-negative operator $\mathcal{N}$ (Sec. 5).

Expressive TTS and instruction-following evaluation. Work on expressive and role-play-oriented TTS evaluation includes InstructTTSEval (Huang et al., 2025c), EmergentTTS-Eval (Manku et al., 2025), MINT-Bench (Chen et al., 2026) for multilingual instruction-following TTS, SpeechParaling-Bench (Liu et al., 2026) for fine-grained paralinguistic-aware speech generation, and discrete-speech LM benchmarks (Wang & Székely, 2024; Chang et al., 2024). Persona-driven TTS frameworks such as P2VA (Lee et al., 2025) also bridge persona descriptions and voice attributes, complementing Speech-DRAME's evaluation focus on the receiving (judge) side. The Interspeech 2024 Discrete Speech Unit Challenge (Chang et al., 2024), in particular, provided community-wide protocols for evaluating discrete-unit-based speech generation across resynthesis, ASR, and TTS tracks, and motivates the broader push toward shared, reproducible evaluation infrastructure that Speech-DRAME extends to role-play. These efforts highlight prosody, emotional control, and style following but remain centered on synthetic outputs and judge-driven scores. They do not provide fine-grained, human-grounded perceptual labels across heterogeneous role-play conditions.

Annotation methodologies: how Speech-DRAME differs. Beyond task focus, Speech-DRAME fundamentally diverges in *how* labels are produced. Table 1 positions Speech-DRAME against a much broader range of contemporary role-play, persona, paralinguistic, and expressive-TTS benchmarks — one row per method — so that the comparison is not narrowed to three pre-selected competitors. The table makes two recurring patterns visible: (i) almost all existing speech-side resources rely on synthetic / TTS-generated audio, with at most a calibration-only slice of real human speech; and (ii) almost none combine *contrastive negatives*, *per-rater confidence*, and *released training-grade human labels* in one resource. Speech-DRAME aims at exactly that intersection.

Our distinction. Speech-DRAME advances role-play evaluation by unifying: (i) a dual evaluation paradigm separating archetype-oriented generality from realism grounded in authentic human speech, (ii) bilingual, rubric-aligned human annotations, and (iii) `SEval` alignment to human ratings, enabling transparent calibration and error analysis. Together, these components move beyond synthetic corpora and judge-only scoring to provide a human-grounded and reproducible framework for evaluating speech role-play performance.

3 Speech Role-play: Formulation and Evaluation Strategies

This section establishes the symbols, models, and data-construction operators that the rest of the paper relies on. Sections 4 and 5 (data curation) and Section 6 (modeling) all back-reference the definitions introduced here.

3.1 Problem Formulation

What is a "role"? We treat a *role* not as a single atomic concept but as a pair ($\boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}}$):

- $\boldsymbol{C}_{\text{profile}}$ — the **character profile**: the persistent identity of the speaker (occupation, social identity, personality traits, named character, age, etc.). In Sec. 4, $\boldsymbol{C}_{\text{profile}}$ is instantiated by *archetype-level keywords* (e.g., "firefighter"); in Sec. 5, it is instantiated by *detailed character descriptions* retrieved from real media.
- $\boldsymbol{C}_{\text{scene}}$ — the **scene specification**: the situational context in which the speaker acts (an event, an interlocutor relationship, an emotional setting). $\boldsymbol{C}_{\text{scene}}$ is again instantiated either by short stereotyped scene prompts (Sec. 4) or by *local scenes* grounded in real plot context (Sec. 5).

All experiments and benchmarks in this paper are single-turn, so we present the formulation in its single-turn form below. The straightforward multi-turn extension — with an additional preceding-dialogue context $\boldsymbol{C}_{\text{speech}}$ — is given in Appendix 8.

Speech role-play (generation). A *speech foundation model* (`SFnd`) maps the role to a spoken response $\boldsymbol{S}_r$:

$$\boldsymbol{S}_r \;=\; \text{SFnd}(\boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}). \tag{1}$$

Table 1: Annotation methodology comparison across recent text-based role-play, speech-native role-play, paralinguistic / conversational, and expressive-TTS benchmarks. Each row is one resource; cells describe the *primary* declared property (auxiliary uses such as calibration-only slices are noted in-cell). **Mod.**: Modality of the rated artifact (T=text, S=speech, M=multimodal). $\boldsymbol{S}_r$ **src.**: dominant source of the spoken response (Syn=synthetic/TTS, Real=real human, Mix=both, "–"=not applicable for text-only). **Labels**: dominant per-sample label source for the *rated artifact* (Judge=judge-model or fully automatic scoring, Human=human raters, Both); resources whose input speech is real and human-curated but whose references or scores are model-generated (e.g., SD-Eval's GPT-4–generated reference responses) are classified by the score/reference source. **Lang.**: primary language(s). **Neg.**: contrastive negatives that perturb scene / role / delivery (✓/✗/n.s. = not stated). **Strategy**: data-collection approach classified following this paper's taxonomy — *Arch.* (top-down archetype: synthesized speech driven by predefined role schemas), *Real.* (bottom-up realism: anchored in real human recordings), *Both*, or "—" for text/multimodal resources where the distinction does not apply. **SEval-ready**: human-labeled split released for training speech-evaluation models. Cells where the cited paper does not state a property are marked "n.s.". Citations point to the entries above. †ParaS2S' main benchmark consists of TTS-synthesized queries; the authors additionally probe generalization on IEMOCAP/MELD real speech.

Method	Cite	Mod.	$\boldsymbol{S}_r$ src.	Labels	Lang.	Neg.	Strategy	SEval-ready	Focus
Text-based role-play									
RoleLLM	(Wang et al., 2024a)	T	–	Judge	EN+ZH	✗	—	✗	Persona fidelity
CharacterEval	(Tu et al., 2024)	T	–	Both	ZH	✗	—	✓	RPCA, reward model
RoleEval	(Shen et al., 2023)	T	–	Judge	EN+ZH	✗	—	✗	Knowledge / persona
RoleMRC	(Lu et al., 2025)	T	–	Judge	EN	✗	—	✗	RP+instruction
CharacterBox	(Wang et al., 2025d)	T	–	Judge	EN+ZH	✗	—	✗	Virtual-world RP
InCharacter	(Wang et al., 2024b)	T	–	Judge	EN+ZH	✗	—	✗	Personality fidelity
PingPong	(Gusev, 2024)	T	–	Judge	EN+RU	✗	—	✗	RP w/ user emul.
RMTBench	(Xiang et al., 2025)	T	–	Judge	EN+ZH	✗	—	✗	Multi-turn user-RP
OpenCharacter	(Wang et al., 2025g)	T	–	Judge	EN	✗	—	✗	Persona training
PersonaEval	(Zhou et al., 2025)	T	–	Both	EN+ZH	✗	—	n.s.	Judge reliability
AtomicPersona	(Shin et al., 2025)	T	–	Both	EN	✗	—	n.s.	Atomic OOC
Multimodal role-play									
MMRole	(Dai et al., 2025a)	M	–	Both	EN+ZH	✗	—	✓	Multimodal RP
Speech-native role-play / character voice									
OmniCharacter-10K	(Zhang et al., 2025a)	S	Syn	Judge	EN+ZH	✗	Arch.	n.s.	Character voice
SpeechRole	(Jiang et al., 2025)	S	Syn	Judge	EN+ZH	✗	Arch.	n.s.	End-to-end RP
VStyle	(Zhan et al., 2025)	S	Syn	Judge	EN+ZH	✗	Arch.	n.s.	Voice style + RP
VoxRole	(Wu et al., 2025)	S	Real	Judge	EN	✗	Real.	n.s.	Film-grounded RP
AudioRole	(Li et al., 2025)	S	Mix	Judge	EN	✗	Real.	✗	TV-grounded RP
Speech-side conversational / paralinguistic-aware benchmarks									
SD-Eval	(Ao et al., 2024)	S	Real	Judge	EN	✗	Real.	✗	Paraling. dialogue
CP-Bench	(Wang et al., 2025e)	S	Real	Human	EN	✗	Real.	✗	Contextual paraling.
ParaS2S	(Yang et al., 2025b)	S	Syn†	Judge	EN	✗	Arch.	✗	Paraling. S2S + RL align
SpeechParaling-Bench	(Liu et al., 2026)	S	Syn	Judge	EN+ZH	✗	Arch.	n.s.	Paraling. generation
StyleBench	(Zhao et al., 2026)	S	Syn	Judge	EN	✗	Arch.	n.s.	Style intensity
Expressive / instruction-following TTS evaluation									
InstructTTSEval	(Huang et al., 2025c)	S	Syn	Judge	EN+ZH	✗	Arch.	n.s.	Instr.-following TTS
EmergentTTS-Eval	(Manku et al., 2025)	S	Syn	Judge	EN	✗	Arch.	n.s.	Prosodic / expressive
MINT-Bench	(Chen et al., 2026)	S	Syn	Both	Multi	✗	Arch.	n.s.	Multi-ling. instr. TTS
Audio-LLM / speech-quality judges									
ALLD	(Chen et al., 2025)	S	Real	Human (MOS)	EN	✗	Real.	✓	Speech-quality LLM
SpeechLLM-as-Judges	(Wang et al., 2025b)	S	Mix	Both	Multi	✗	n.s.	n.s.	Interpretable judge
Speech-DRAME (Ours)	—	S	**Mix (Real+Syn)**	**Human**	**EN+ZH**	✓	**Both**	✓	**Role-play, dual perspective**

Here $\boldsymbol{S}_r$ represents the spoken output (waveform or latent representation), capturing both the linguistic dimension (coherence with the scene) and the paralinguistic dimension (prosody, style, emotion) expected of the role.

Evaluation. A *speech evaluation model* (`SEval`) assesses $\boldsymbol{S}_r$ using the same conditioning signals:

$$\boldsymbol{E}_r \;=\; \mathrm{SEval}(\boldsymbol{S}_r,\, \boldsymbol{C}_{\mathrm{profile}},\, \boldsymbol{C}_{\mathrm{scene}}). \tag{2}$$

$\boldsymbol{E}_r$ is the per-dimension human-aligned score that we use as both the supervision target for training `SEval` models (Sec. 6) and the unit of comparison when ranking `SFnd` systems (Sec. 4–5).

Curation operator $\mathcal{D}$. The data underlying $\boldsymbol{E}_r$ is not given for free; it is produced by a **curation pipeline** $\mathcal{D}$ that converts raw inputs into evaluation samples. We instantiate $\mathcal{D}$ in two complementary ways, mirroring the dual evaluation perspectives introduced in Sec. 3.2. Let $\mathcal{K}_p$ and $\mathcal{K}_s$ denote vocabularies of *role* and *scene* keywords; let T_{prompt} be a template-based expansion that turns a keyword pair into a full profile/scene specification; let $\mathcal{R}$ denote a corpus of real human recordings paired with surrounding media context σ; and let Φ_{retr} denote a retrieval-and-LLM reconstruction operator. Then:

$$\mathcal{D}^{\text{top-down}}: \ (k_p,\, k_s) \in \mathcal{K}_p \times \mathcal{K}_s \ \xrightarrow{T_{\text{prompt}}} \ (\boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}) \ \xrightarrow{\texttt{SFnd}} \ (\boldsymbol{S}_r,\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}), \tag{3}$$

$$\mathcal{D}^{\text{bottom-up}}: \ (\boldsymbol{S}_r,\, \sigma) \in \mathcal{R} \ \xrightarrow{\Phi_{\text{retr}}} \ (\boldsymbol{S}_r,\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}). \tag{4}$$

Both operators produce a tuple $(\boldsymbol{S}_r,\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}})$ as the unit of analysis: $\mathcal{D}^{\text{top-down}}$ by synthesizing $\boldsymbol{S}_r$ from a keyword-expanded role, and $\mathcal{D}^{\text{bottom-up}}$ by reconstructing the role around a real recording. $\mathcal{D}^{\text{top-down}}$ underlies Sec. 4 (archetype data); $\mathcal{D}^{\text{bottom-up}}$ underlies Sec. 5 (realism data).

Negative-sample operator $\mathcal{N}$. For both data flows, we also define a **negative-sample operator** $\mathcal{N}$ that perturbs one component of a positive sample $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ while holding the others fixed, yielding a triple in which the speaker's delivery, identity, or scene context has been deliberately mismatched:

$$\mathcal{N}(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}}) \in \big\{(\tilde{\boldsymbol{S}}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}}),\ (\boldsymbol{S}_r, \tilde{\boldsymbol{C}}_{\text{profile}}, \tilde{\boldsymbol{C}}_{\text{scene}}),\ (\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \tilde{\boldsymbol{C}}_{\text{scene}})\big\}. \tag{5}$$

The three branches correspond to *delivery-mismatch*, *role-mismatch*, and *context-mismatch* negatives respectively, all elaborated in Sec. 5. Two points deserve emphasis. First, $\mathcal{N}$ can produce *multi-property* perturbations in which more than one component is changed simultaneously — for example, $(\tilde{\boldsymbol{S}}_r, \boldsymbol{C}_{\text{profile}}, \tilde{\boldsymbol{C}}_{\text{scene}})$ changes both delivery and scene at once. Second, the role-mismatch branch *always* pairs $\tilde{\boldsymbol{C}}_{\text{profile}}$ with a correspondingly updated $\tilde{\boldsymbol{C}}_{\text{scene}}$: because a character profile and scene must be mutually consistent to form a coherent annotation target, changing the profile alone while leaving the original scene in place would introduce a contradiction rather than a controlled mismatch. We therefore jointly replace both when perturbing the role identity. By construction, the perturbed components produced by $\mathcal{N}$ (e.g., $\tilde{\boldsymbol{C}}_{\text{scene}}$, $\tilde{\boldsymbol{S}}_r$) are *not real*: they are generated specifically to probe whether `SEval` attends to fine-grained perceptual cues. This is made explicit in all downstream tables and figures.

3.2 Role-play Evaluation with Dual Perspectives: Archetype and Realism

Speech role-play evaluation can be conceptualized through multiple lenses depending on the nature of available data and the desired granularity of perceptual judgment. In Speech-DRAME, we formalize *two complementary evaluation perspectives*: *Archetype-based* (top-down, $\mathcal{D}^{\text{top-down}}$ in Eq. 3) and *Realism-based* (bottom-up, $\mathcal{D}^{\text{bottom-up}}$ in Eq. 4). Both perspectives are grounded in distinct theoretical traditions within psychology, sociology, and performance studies, and reflect how humans assess roles in interaction: through both top-down, culturally shared schemas and bottom-up, contextually enacted behaviors. We organize this subsection in three layers: (i) the theoretical motivation for each perspective; (ii) how each perspective instantiates the symbols of Sec. 3.1; and (iii) a side-by-side contrast (Table 2) that the rest of the paper refers back to.

3.2.1 Theoretical motivation

Archetypes as collective schemas. The notion of archetypes originates in Jungian psychology, where they represent culturally recurring prototypes embedded in the collective unconscious (Jung, 2014). Media and performance studies show that archetypes guide audience expectations, providing *top-down interpretive frames* for judging whether a character "fits" familiar categories such as hero, mentor, or caregiver (Campbell, 2008). Because these schemas are widely shared, archetype-based judgments scale efficiently across scenarios and enable broad comparisons.

Realism as situated performance. In contrast, theories of social interaction emphasize the *fine-grained enactment* of roles in situated contexts. Goffman's dramaturgical sociology conceptualizes everyday communication as performance, where believability depends on subtle paralinguistic cues—prosody, hesitation, pacing, and affect (Goffman, 1949). Interactional sociolinguistics further shows how micro-level features such as intonation contours and turn-taking patterns shape perceptions of authenticity (Gumperz, 1982). Realism-oriented evaluation therefore attends to detailed acoustic and pragmatic cues that fall outside schematic expectations.

Complementary perspectives. Performance theory characterizes role behavior as both "restored" (scripted, repeatable structures) and "emergent" (situated enactment) (Schechner, 2017), while sociolinguistic theories distinguish between *ascribed categories* and *performed identities* (Bucholtz & Hall, 2005). These viewpoints motivate the need for dual evaluation: archetype-based assessments capture macro-level expectations rooted in cultural schemas, whereas realism-based assessments capture micro-level delivery and situational fit. Neither perspective subsumes the other; each reveals a different dimension of how listeners interpret role performance. Speech-DRAME therefore adopts both, enabling analyses that are at once scalable and sensitive to the nuances of authentic spoken interaction.

3.2.2 Mapping each perspective onto the formulation

We now connect each perspective to the symbols of Sec. 3.1.

Archetype Evaluation (top-down, $\mathcal{D}^{\text{top-down}}$). $\boldsymbol{C}_{\text{profile}}$ is instantiated by a *stereotypical role definition* such as "firefighter" or "ER doctor"; $\boldsymbol{C}_{\text{scene}}$ is instantiated by a concise scene prompt (e.g., "comfort a child trapped in fire"). The spoken response $\boldsymbol{S}_r$ is then synthesized by an `SFnd`, and $\boldsymbol{E}_r$ is computed against three context-aware dimensions (Sec. 4.2). This strategy provides scalable judgments grounded in shared schemas and supports broad cross-scenario comparisons, particularly under limited annotation resources (Gemmeke et al., 2017). Its reliance on broad prototypes, however, limits sensitivity to fine-grained delivery such as emotional modulation or subtle prosodic control (Holt & Lotto, 2010).

Realism Evaluation (bottom-up, $\mathcal{D}^{\text{bottom-up}}$). Here $\boldsymbol{S}_r$ is given first — drawn from real recordings of professional and non-professional speakers in natural dialogue and media — and $\boldsymbol{C}_{\text{profile}}$ and $\boldsymbol{C}_{\text{scene}}$ are reconstructed around it via retrieval and LLM-assisted summarization. $\boldsymbol{E}_r$ is computed across the five fine-grained dimensions detailed in Sec. 5.2: prosodic dynamics, emotional expressiveness, character consistency, contextual relevance, and semantic match. To stress-test `SEval` models, the bottom-up flow additionally produces contrastive negatives via the operator $\mathcal{N}$ (Eq. 5). Although annotation-intensive, this approach is essential for applications where realism and subtlety are central to perceived performance.

Integrated perspective. Archetype-based and realism-based evaluations illuminate distinct facets of spoken role performance: the former provides macro-level judgments aligned with culturally shared expectations, offering scalability and coarse-grained comparability; the latter captures the fine-grained qualities that make a performance believable in real interactional settings. By incorporating both, Speech-DRAME supports analyses that span broad categorical expectations as well as nuanced, context-sensitive delivery.

3.2.3 Side-by-side contrast

Table 2 summarizes the two perspectives along the dimensions of the formulation. Sections 4 and 5 are organized so that each row of this table is operationalized in turn.

In the following two sections (Sec. 4 and 5), we detail the data collection processes corresponding to the two columns of Table 2, respectively.

4 DRAME-EvalBench Data Curation — Archetype Data Collection

This section instantiates the *top-down* (archetype) column of Table 2: the curation operator $\mathcal{D}^{\text{top-down}}$ (Eq. 3), the corresponding annotation rubric, and the resulting dataset. The parallel *bottom-up* (realism)

Table 2: Top-down (Archetype) vs. bottom-up (Realism) evaluation in Speech-DRAME, expressed in the symbols of Sec. 3.1. "GT" = ground truth.

Aspect	**Archetype (top-down)**	**Realism (bottom-up)**
Curation operator	$\mathcal{D}^{\text{top-down}}$ (Eq. 3)	$\mathcal{D}^{\text{bottom-up}}$ (Eq. 4)
Source of $\boldsymbol{S}_r$	Synthesized by an `SFnd`	Real human recording (TV/film/recordings)
Source of $\boldsymbol{C}_{\text{profile}}$	Stereotypical role keyword + template	Retrieved from media context
Source of $\boldsymbol{C}_{\text{scene}}$	Stereotyped scene prompt	LLM-summarized local plot scene
Granularity of $\boldsymbol{E}_r$	3 context-aware dims (Sec. 4.2)	5 fine-grained dims w/ sub-dims (Sec. 5.2)
Negative samples ($\mathcal{N}$)	Not used	Delivery / role / context mismatches
Scale vs. depth	Broad, scalable, schema-driven	Deep, annotation-heavy, perceptual
Used to benchmark	`SFnd` vs. stereotype expectation	`SEval` vs. human perception, `SFnd` realism

Table 3: High-level statistics of DRAME-EvalBench. *Negatives* counts samples produced by the operator $\mathcal{N}$ in Eq. 5; these are explicitly synthetic / mismatched and not real recordings.

Aspect	**Archetype (Sec. 4)**	**Realism (Sec. 5)**
Curation operator	$\mathcal{D}^{\text{top-down}}$ (Eq. 3)	$\mathcal{D}^{\text{bottom-up}}$ (Eq. 4)
Languages	English + Mandarin	English (5k) + Mandarin (9k)
Source of $\boldsymbol{S}_r$	Synthesized by 7–8 `SFnd` systems	Real human recordings (TV, film, original)
# scenarios	552 (per language)	14k unique (positive)
# annotated samples (total)	8,280	15k (12k tr / 2k test / 1k real-rec test)
Train / Test split	6,780 / 1,500	12,000 / 3,000
Negatives via $\mathcal{N}$ (*not real*)	—	$\tilde{\boldsymbol{C}}_{\text{scene}} + \tilde{\boldsymbol{C}}_{\text{profile}} + \tilde{\boldsymbol{S}}_r$ (Sec. 5.1.2)
Annotation rubric	3 dims + content pass (Sec. 4.2)	5 dims w/ sub-dims (Sec. 5.2)

column is operationalized in Sec. 5, which is deliberately structured to mirror this section so that the two flows can be read side-by-side.

High-level statistics of DRAME-EvalBench. Before detailing the archetype pipeline, Table 3 summarizes the high-level statistics of the full DRAME-EvalBench database, covering both the archetype data introduced in this section and the realism data introduced in Sec. 5. Detailed per-language and per-model breakdowns are deferred to Appendix H and I.

4.1 Archetype Data Curation Pipeline

Archetype Data Collection follows a top–down perspective rooted in sociology and role theory, where social behavior is interpreted through shared stereotypes and generalized expectations (Goffman, 1949; Campbell, 2008). This strategy abstracts role-play into broad and easily recognizable archetypes, enabling scalable judgment of whether a generated response aligns with salient role features. In the formulation of Sec. 3.1, the role $(\boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ here corresponds to a (stereotype keyword, scene keyword) pair expanded through templates, and $\boldsymbol{S}_r$ is synthesized by an existing `SFnd`.

Role dimensions ($\boldsymbol{C}_{\text{profile}}$). We operationalize the character profile $\boldsymbol{C}_{\text{profile}}$ through six principal archetype categories, each representing a stereotypical dimension of role identity or behavior, including

- (1) **Occupation / Functional Role (Modern Daily Life)** (e.g., a firefighter rescuing a trapped civilian).
- (2) **Occupation / Functional Role (Fictional or Fantasy)** (e.g., a martial arts cult leader confronting orthodox rivals).

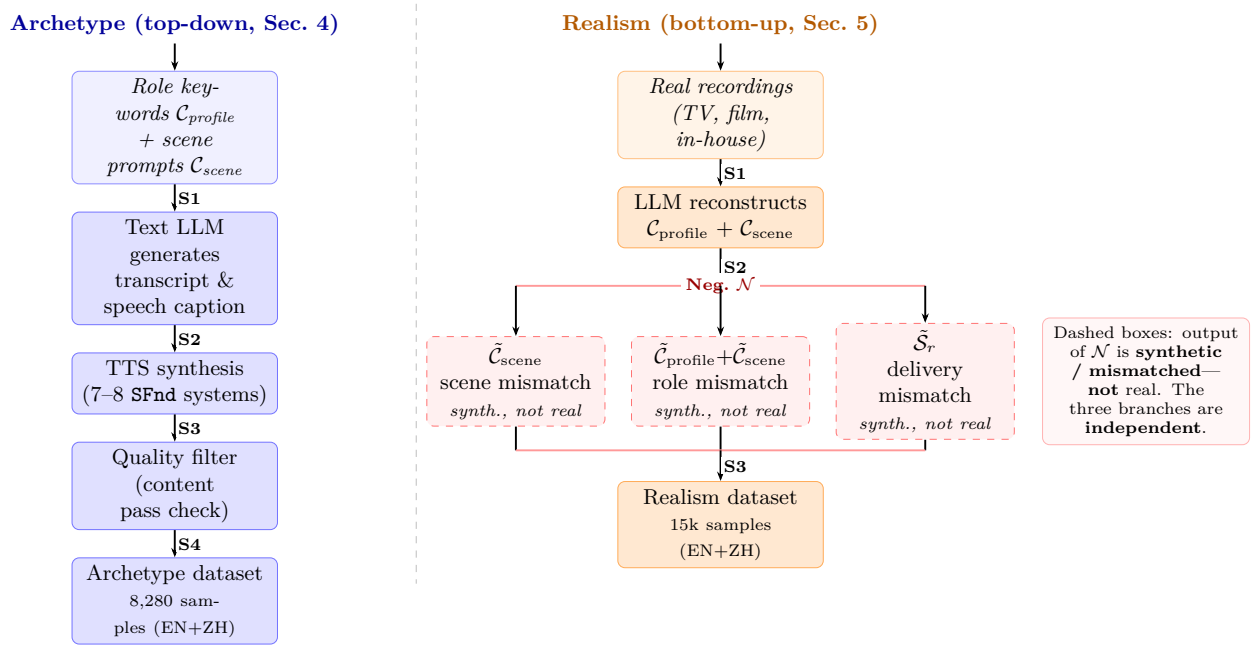


Figure 2: Data curation pipelines for the two evaluation strategies. **Left (blue):** Archetype top-down pipeline (Sec. 4) — role/scene keywords drive LLM generation and TTS synthesis. **Right (orange):** Realism bottom-up pipeline (Sec. 5) — real recordings anchor the pipeline; the negative operator $\mathcal{N}$ applies three *independent* (not sequential) perturbation branches in parallel: scene mismatch, role mismatch, and delivery mismatch. All perturbed outputs are explicitly synthetic/mismatched, not real recordings.

- (3) **Social Identity** (e.g., a doting parent worrying about a child, or a "Buddhist youth" eschewing competition).
- (4) **Named / Specific Characters** (e.g., Sun Wukong mocking celestial authorities in *Journey to the West*).
- (5) **Personality Traits** (e.g., "hot-blooded" in a competitive setting, or "complaining" in response to poor service).
- (6) **Relative Roles** (e.g., parent or stranger).

Scene dimensions ($\boldsymbol{C}_{\text{scene}}$). In addition to the profile $\boldsymbol{C}_{\text{profile}}$, we incorporate scene specifications $\boldsymbol{C}_{\text{scene}}$ that anchor role-play in stereotypical situations. A central focus is on **everyday external events or state changes** (e.g., receiving a promotion, being caught in gossip), which provide concrete scenarios for testing archetypical responses.

Exclusion of fine-grained controls. We deliberately excluded explicit emotional states, prosodic instructions, or linguistic constraints. These dimensions are better suited to realism-oriented or style-controlled evaluation, whereas archetype evaluation benefits from concrete, stereotype-level scenarios that sharpen judgments and improve human agreement.

Pipeline overview. At a high level, archetype evaluation data are generated through a three-stage pipeline: (1) selecting keywords that specify role and scene; (2) expanding them into full prompts via structured templates; and (3) synthesizing prompts into speech with existing `SFnd` systems. This modular pipeline ensures both scalability and reproducibility, while keeping prompts vivid enough for reliable judgments. Implementation details, including template design and alternative prompting strategies are provided in the Appendix C.

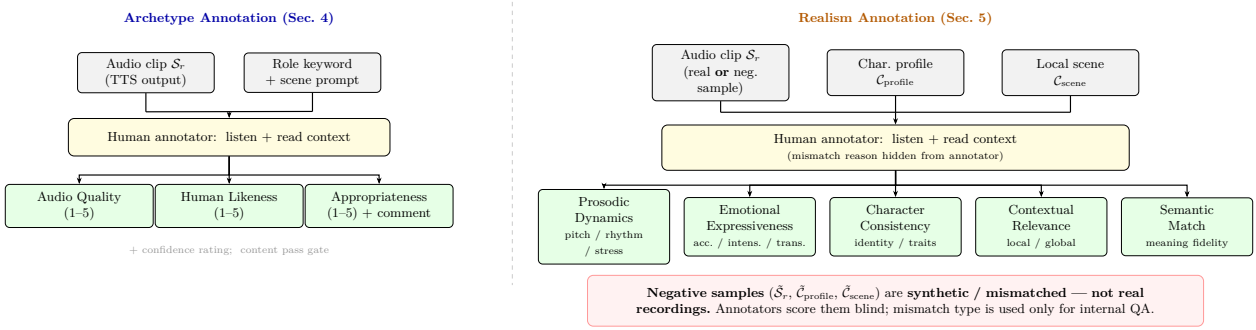


Figure 3: Annotation flows for both evaluation strategies (side by side). **Left (blue):** Archetype annotation (Sec. 4.2) — annotators rate three dimensions (Audio Quality, Human Likeness, Appropriateness) from TTS output + role/scene context. **Right (orange):** Realism annotation (Sec. 5.2) — annotators rate five dimensions (Prosodic Dynamics, Emotional Expressiveness, Character Consistency, Contextual Relevance, Semantic Match) from real *or* negative samples; mismatch type is hidden from annotators and used only for internal QA. Negative samples are explicitly *synthetic / mismatched* — not real recordings.

Based on the generated data (paired scene and synthesized speech), expert human annotators are recruited for detailed annotation on semantic check (i.e., *Content Pass*) and three different spoken role-play dimensions (i.e., *Audio Quality*, *Human-likeness*, *Appropriateness*), which is detailed in following Section 4.2.

Overall, archetype evaluation emphasizes breadth and recognizability over subtlety. By grounding role-play in socially recognizable categories and stereotypical situations, it provides a practical balance between scalability and interpretability, complementing the more fine-grained realism evaluation described in Section 3.2.

4.2 Archetype Annotation Protocol and Scoring Dimensions

The curation pipeline above produces tuples $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ as defined in Eq. 2, where $\boldsymbol{C}_{\text{profile}}$ and $\boldsymbol{C}_{\text{scene}}$ are stereotype-level role/scene prompts and $\boldsymbol{S}_r$ is the synthesized response. To convert these outputs into reliable supervision signals $\boldsymbol{E}_r$ for both training and benchmarking `SEval` models, we apply a dedicated archetype-oriented annotation protocol. As in Sec. 5.2 for the realism flow, we use the term *scoring dimensions* (rather than "evaluation dimensions") to keep the human-rating axes terminologically separate from the `SEval` model itself.

Because archetypes rely on broad, easily recognizable expressive patterns rather than fine-grained behavioral detail, the annotation rubric focuses on dimensions that listeners naturally use when judging whether a speaker fits a stereotypical role. The protocol consists of three dimensions:

- **Audio Quality** (context-free). A preliminary assessment of whether the signal is sufficiently clear for annotators to evaluate role compatibility (Kirkland et al., 2023; Cooper et al., 2024; Perrotin et al., 2025). Noticeable artifacts, distortions, or unstable timbre may obscure vocal cues relevant to archetype identification.
- **Human-likeness** (context-free). This dimension evaluates whether the speech exhibits properties of natural human vocal production, including stable timbre and plausible prosody. Studies in voice perception indicate that listeners attribute personality and social characteristics only when the voice falls within a human vocal range (Cooper et al., 2024). As such, human-likeness serves as a prerequisite for making interpretable role judgments.
- **Appropriateness** (context-dependent). The central dimension of archetype evaluation. It measures whether the expressive style, including prosodic patterns, affective coloration, and trait-related vocal cues, aligns with the archetype implied by the role and scene description. This dimension requires considering contextual framing, consistent with sociological accounts of role enactment and expectation (Welch et al., 2021; Lavan, 2023; Guldner et al., 2024).

Together, these dimensions determine whether a spoken performance is (i) technically interpretable, (ii) human-sounding, and (iii) compatible with the intended archetype. The rubric avoids detailed prosodic or emotional sub-dimensions—addressed separately in realism evaluation—because archetype judgments depend more on global stylistic patterns than on fine-grained contextual behavior.

Annotators also complete a **Content Pass** check to verify that the spoken response conveys the intended meaning of the prompt. Full scoring scales, examples, and instructions are provided in Appendix C.

4.3 Candidate Models for Evaluation

We evaluate a diverse set of speech foundation models spanning both English and Mandarin. In total, we include 8 English and 7 Mandarin models, covering open-source and proprietary systems, and balancing end-to-end and cascaded architectures. Table 4 provides the full list of models.

Table 4: Candidate speech foundation models for evaluation across English and Mandarin. `G` and `Q` in for cascaed models refer to `Gemini2.5Pro` and `Qwen3-30B-Instruct`.

Model	License / Availability	Architecture	Reference
English Models (8)			
`GPT-4o-realtime`	Proprietary	Unknown	(OpenAI, 2025b)
`Doubao`	Proprietary	Unknown	(ByteDance, 2025)
`GPT-4o`(Advanced Voice Mode)	Proprietary	Unknown	(OpenAI, 2025a)
`GLM4-Voice`	Open-source	End-to-end	(Zeng et al., 2024)
`KimiAudio`	Open-source	End-to-end	(Ding et al., 2025)
`Qwen-2.5-Omni`	Open-source	End-to-end	(Xu et al., 2025)
`Q-CosyVoice2`	Open-source	Cascaded	(Yang et al., 2025a; Du et al., 2024b)
`G-Cosyvoice2`	Proprietary + Open-source	Cascaded	(Comanici et al., 2025; Du et al., 2024b)
Mandarin Models (7)			
`GPT-4o-realtime`	Proprietary	Unknown	(OpenAI, 2025b)
`Doubao`	Proprietary	Unknown	(ByteDance, 2025)
`GLM4-Voice`	Open-source	End-to-end	(Zeng et al., 2024)
`KimiAudio`	Open-source	End-to-end	(Ding et al., 2025)
`Qwen-2.5-Omni`	Open-source	End-to-end	(Xu et al., 2025)
`Q-CosyVoice2`	Open-source	Cascaded	(Yang et al., 2025a; Du et al., 2024b)
`G-Cosyvoice2`	Proprietary + Open-source	Cascaded	(Comanici et al., 2025; Du et al., 2024b)

Collection for Proprietary Models. For proprietary systems, we adopt two collection strategies. For GPT-4o-realtime, responses are obtained directly via the `gpt-4o-realtime` API. For Doubao and GPT-4o (Advanced Voice Mode), we collect outputs by recording from their official web interfaces. All recordings are captured through the interaction device itself, ensuring that no additional processing or distortion is introduced into the final audio.

End-to-End Model Generation. For end-to-end systems, the audio prompt is directly used as input. The system prompt is set to default values to ensure consistent speech generation quality.

Cascaded Model Generation. For cascaded systems, we only use Cosyvoice2 as the TTS model in our collection of data. We select `Qwen3-30B-Instruct` and `Gemini2.5Pro` as the text LLM candidates (Yang et al., 2025a; Comanici et al., 2025).

During the generation, we adopt a retrieval-augmented generation (RAG) strategy that combines text LLMs with zero-shot TTS models:

- Step 1: We first build a speech database (LibriSpeech and GigaSpeech for English, WenetSpeech for Mandarin (Panayotov et al., 2015; Chen et al., 2021; Zhang et al., 2022)). Each sample is annotated with Gemini-based speech captions describing salient acoustic attributes.

- Step 2: Given a character and scene description, the text LLM imagines a candidate caption and generates a corresponding transcript (see below prompts for the text LLM).
- Step 3: We retrieve acoustically similar samples from the database using `Qwen3-Embedding-7B` (Zhang et al., 2025d).
- Step 4: The retrieved samples guide the zero-shot TTS system, which synthesizes speech conditioned on both the transcript and matched acoustic profile.

4.4 Archetype Evaluation Data Summary.

The resulting archetype evaluation dataset comprises 552 distinct scenarios for both Mandarin and English, evaluated across eight models,[3] yielding 8,280 samples. In total, the dataset is split into 6,780 samples for training and 1,500 for testing. The split is constructed so that no role-keyword $\times$ scene-keyword combination appears in both partitions, ensuring that test evaluation is not inflated by memorized (keyword, scenario) pairs. For benchmarking, an additional 1,250 scenarios spanning seven role/scene categories are prepared to facilitate archetype role-play benchmarking. More detailed statistics and analyses are provided in Appendix H.

5 DRAME-EvalBench Data Curation — Realism Data Collection

This section instantiates the *bottom-up* (realism) column of Table 2. In contrast to Sec. 4, where the spoken response $\boldsymbol{S}_r$ is synthesized from stereotype keywords, here $\boldsymbol{S}_r$ is given first as a real human recording and the role ($\boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}}$) is reconstructed around it (curation operator $\mathcal{D}^{\text{bottom-up}}$ in Eq. 4). On top of this positive flow, we additionally apply the negative-sample operator $\mathcal{N}$ of Eq. 5 to obtain contrastive negatives that probe whether `SEval` attends to delivery, identity, and scene cues. Throughout this section, items produced by $\mathcal{N}$ are explicitly marked as *not real*: the perturbed components are constructed for stress-testing and are never served to annotators as ground truth.

5.1 Realism Data Curation Pipeline

Realism Data Collection adopts a bottom–up perspective grounded in real human speech data, with the goal of capturing the nuanced delivery, prosody, and contextual fit that synthetic archetype prompts cannot fully represent. Rather than relying on stereotypes, this strategy leverages real-world recordings and carefully constructed contrasts, enabling fine-grained assessment of model performance under realistic conditions.

5.1.1 Positive samples: real speech with reconstructed role

The positive flow proceeds in three sequential stages, all operating on the same underlying spoken segment $\boldsymbol{S}_r$:

(P1) Acquire $\boldsymbol{S}_r$ — ground-truth audio from TV and film. Spoken segments are curated from publicly available media, ensuring coverage across diverse roles and communicative settings. This corresponds to the "$\boldsymbol{S}_r$" input of Eq. 4.

(P2) Reconstruct $\boldsymbol{C}_{\text{profile}}$ — retrieved character profile. Each audio segment is paired with a concise character description derived from the media context. The detailed prompt design for generating these character profiles is provided in Appendix E.1.

(P3) Reconstruct $\boldsymbol{C}_{\text{scene}}$ — generated local scene. We supplement the retrieved character information with an automatically generated local scene description summarizing the surrounding plot or setting. Appendix E.2 includes representative prompts and generation details.

[3]The model list is discussed in Appendix 4.3. Note we only include seven models for English due to the deactivation of the ChatGPT-4o advanced voice model.

The output of this positive flow is a tuple $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ — directly the same signature as Eq. 2. To make the example concrete: in a clip from a courtroom drama, $\boldsymbol{S}_r$ is the actor's spoken line; $\boldsymbol{C}_{\text{profile}}$ is the reconstructed description of the *character* ("a stoic defense attorney in his late 40s, controlled but tense"); and $\boldsymbol{C}_{\text{scene}}$ is the reconstructed description of the surrounding *scene* ("cross-examination of a hostile witness during a high-stakes trial").

5.1.2 Negative samples: contrastive perturbations via $\mathcal{N}$

A central challenge in realism-oriented assessment is determining whether a judge model can detect subtle inconsistencies between how a line is delivered and the narrative or role context in which it is supposed to occur. Real human speech encodes role-dependent clues — intonation, emotional stance, personality expression, pacing, and interactional intent — that cannot be captured by transcript alone. To evaluate whether models attend to these cues, we apply the negative-sample operator $\mathcal{N}$ defined in Eq. 5: starting from a positive sample $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$, we perturb *exactly one* component, holding the others fixed.

Why constructed negatives. The importance of such constructed negatives is well-established across evaluation research. In speech, benchmarks such as NISQA and ConferenceSpeech 2022 rely heavily on simulated degradations and synthetic perturbations to expose model weaknesses that natural corpora fail to capture (Rao et al., 2021; Mittag et al., 2021). The same principle appears in language and multimodal datasets: NLI benchmarks introduce adversarial hypotheses and counterfactual contradictions (Nie et al., 2020), while FOIL-COCO and Winoground mismatch caption semantics to test grounding (Shekhar et al., 2017; Thrush et al., 2022). Across all these domains, human-validated mismatches reveal failure modes that positive-only datasets systematically overlook.

Three parallel branches of $\mathcal{N}$. Eq. 5 admits three parallel branches; *any one positive sample can give rise to a negative under any branch*, but a given generated negative belongs to exactly one branch. The three branches are:

- **Negative Local Scenes (context-mismatch, $\tilde{\boldsymbol{C}}_{\text{scene}}$).** Probes whether judges detect *scene-level narrative inconsistencies*. For each authentic segment we generate $\tilde{\boldsymbol{C}}_{\text{scene}}$ at varying mismatch depths, from subtle semantic perturbations (incorrect interpersonal dynamics or emotional framing) to entirely unrelated scenarios. Generated via prompt-based text generation; detailed pipeline in Appendix E.
- **Negative Character Profiles (role-mismatch, $\tilde{\boldsymbol{C}}_{\text{profile}}$).** Probes whether judges recognize *persona and identity mismatches*. We pair a true speech segment with an incorrect character description or swap key personality traits, social roles, and emotional dispositions. Generated via prompt-based text generation; detailed pipeline in Appendix E.
- **Negative Speech Samples (delivery-mismatch, $\tilde{\boldsymbol{S}}_r$).** Probes whether judges detect mismatches in prosody, emotional intent, vocal identity, or speaking style. We resynthesize the transcript with diverse TTS voices, modify expressive style, or deliberately alter prosodic alignment. Sequential per-system pipelines (Speaker Assignment, Voice Clone, etc.) are detailed in Sec. 5.3.

All three branches produce content that is explicitly synthetic / mismatched and therefore not real. Together, they cover complementary aspects of realism: $\tilde{\boldsymbol{C}}_{\text{scene}}$ tests contextual grounding, $\tilde{\boldsymbol{C}}_{\text{profile}}$ tests persona and role consistency, and $\tilde{\boldsymbol{S}}_r$ tests paralinguistic and delivery appropriateness. This triad mirrors the multi-axis perturbation strategies used in the human-labeled datasets cited above, and ensures that realism assessment captures distinctions that are easily perceived by human listeners but often overlooked by current speech models.

Annotation conditions. During annotation, raters see context information drawn from one of (i) ground-truth stories, (ii) stories with negative local scenes ($\tilde{\boldsymbol{C}}_{\text{scene}}$), or (iii) stories with negative character profiles ($\tilde{\boldsymbol{C}}_{\text{profile}}$); for the target speech, different types of synthesized speech and ground-truth speech are mixed. Annotators assign realism scores *without* being exposed to the underlying mismatch reasons. Those reasons

Table 5: Role-play evaluation dimensions and corresponding sub-dimensions with keyword-style summaries. The table highlights core cues used for scoring across prosody, emotion, character, context, and semantics.

Dimension	Sub-dimension	Discussion Focus
Prosodic Dynamics	Pitch Variation	Contour richness; range; monotony vs. expressiveness
	Rhythmic Naturalness	Pacing; timing; pause placement; flow
	Stress & Emphasis	Accenting; highlight key words; communicative salience
Emotional Expressiveness	Emotion Accuracy	Correct emotion type; scene-appropriate affect
	Emotion Intensity	Strength calibration; under-/over-shooting
	Emotion Transition	Smooth shifts; narrative-motivated changes
Character Consistency	Voice Identity Matching	Timbre stability; general impression match
	Traits Embodiment	Personality cues; role-specific traits
Contextual Relevance	Local Scene Fit	Situational alignment; Environmental interaction
	Global Story Coherence	Character likeness; world-rule adherence; continuity
Semantic Match	Semantic Match	Meaning fidelity; key details preserved; no hallucination

are visible only to the quality-assurance team, who use them as reference for internal validation and consistency checks. While individual mismatch reasons may not be perfect, they support quality control and maintain reliability across the dataset.

A worked example: *Yunjin.* To make the operator $\mathcal{N}$ concrete in the body of the paper, consider an opera segment from *Genshin Impact* included in our released set. The positive tuple $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ is: $\boldsymbol{S}_r$, the actor's operatic narration of a battle scene; $\boldsymbol{C}_{\text{profile}}$, "Yunjin, a passionate opera performer who narrates heroic deeds in traditional singing style"; $\boldsymbol{C}_{\text{scene}}$, "a fierce battlefield narration in which armored riders charge into enemy lines." Applying the three branches of $\mathcal{N}$ yields contrastive triples that share the *same audio* (or transcript) but break realism in different ways: $\tilde{\boldsymbol{C}}_{\text{scene}}$ replaces the battlefield with a quiet teahouse, a calm tea gathering, or a modern café (mild $\rightarrow$ severe context-mismatch); $\tilde{\boldsymbol{C}}_{\text{profile}}$ replaces Yunjin with an introverted software engineer, a logical roboticist, or an anti-traditionalist street performer (mild $\rightarrow$ severe role-mismatch); and $\tilde{\boldsymbol{S}}_r$ resynthesizes the same lyrics with one of the catalog/cloning TTS configurations of Sec. 5.3, replacing the operatic delivery with a flatter or stylistically unrelated voice. Crucially, the transcript is unchanged across all three branches, so a transcript-only judge cannot tell positives from negatives; a model that scores realism correctly therefore must use prosody, identity, or scene cues. The full anchored example, including the original verses and per-level mismatch rationales, is reproduced in App. E.4.

Real-world Recording. As noted in Sec. 2, most prior efforts have relied heavily on synthetic data, mostly from TTS systems based on curated professional role-play voices. While such data can achieve high fidelity, it often lacks the diversity and spontaneity of real-world delivery, limiting the generalizability of evaluation models. To address this gap, we expand realism evaluation beyond synthetic and curated voices by incorporating newly recorded speech from both professional and amateur voice actors. This design introduces a broader spectrum of delivery styles, from polished professional acting to more natural and less controlled amateur performance.

5.2 Annotation Protocol and Scoring Dimensions

This subsection specifies how human annotators produce $\boldsymbol{E}_r$ in Eq. 2 for the realism flow. Concretely, given a tuple $(\boldsymbol{S}_r, \boldsymbol{C}_{\text{profile}}, \boldsymbol{C}_{\text{scene}})$ — possibly perturbed by $\mathcal{N}$ as described in Sec. 5.1.2 — annotators rate the spoken performance $\boldsymbol{S}_r$ against the provided *character profile* $\boldsymbol{C}_{\text{profile}}$ and *local scene* $\boldsymbol{C}_{\text{scene}}$ using a structured multi-dimensional rubric. We use the term *scoring dimensions* (rather than "evaluation dimensions") below to disambiguate the human-rating axes from the `SEval` model itself. The rubric is designed to capture the core perceptual factors that listeners naturally rely on when judging spoken role-play, focusing on five complementary dimensions:

- **Prosodic Dynamics** (Pitch Variation, Rhythmic Naturalness, Stress & Emphasis): Prosody is the primary carrier of expressiveness in speech and strongly shapes a listener's perception of naturalness, intent, and dramatic nuance (Bänziger & Scherer, 2005). These sub-dimensions capture whether the speaker delivers a rhythmically coherent, suitably varied, and communicatively emphasized performance.

- **Emotional Expressiveness** with progressive gating (Emotion Accuracy → Intensity → Transition): Emotional realism underpins believable role-play (Ye et al., 2023; Russell, 1980; Johnstone & Scherer, 2000). The gated structure reflects the perceptual dependency among aspects of emotion: listeners first check *whether* the expressed emotion is appropriate (Accuracy), then *how strongly* it is conveyed (Intensity), and finally *how it evolves* over time (Transition). This avoids collapsing distinct failure modes (e.g., wrong emotion vs. correct emotion but unnatural transitions).

- **Character Consistency** (Voice Identity Matching, Traits Embodiment): Role-play requires maintaining a stable character. This dimension ensures that speakers remain identifiable across turns and that their vocal performance conveys the persona's defining traits (e.g., age, personality, profession). It captures aspects missing from standard speech-quality metrics and cannot be inferred from transcript-level correctness alone (Podesva & Callier, 2015).

- **Contextual Relevance** (Local Scene Fit, Global Story Coherence): Spoken role-play is fundamentally contextual rather than isolated (Grosz & Sidner, 1986). This dimension evaluates whether the performance responds appropriately to the immediate situation (*local fit*) and remains consistent with broader narrative constraints (*global coherence*). Models frequently fail at this long-range consistency, making it critical for realism-oriented evaluation.

- **Semantic Match**: This captures whether the spoken realization faithfully reflects the intended semantic content. While prosody and emotion can vary stylistically, semantic fidelity is a hard requirement for meaningful evaluation and for differentiating expressive errors from content errors.

These dimensions are chosen to jointly cover the full stack of perceptual factors involved in human assessment of situated speech: from low-level acoustic expressiveness (prosody), to affect (emotion), persona stability (character), contextual grounding, and semantic correctness. They exclude redundant or highly system-dependent metrics (e.g., signal-level artifacts, basic intelligibility scores), focusing instead on factors that are both *context-aligned* and *specific to role-play*. Annotators also provide a prompt-quality check and confidence ratings. Full scales, anchors, and gating logic are presented in Appendix F.

5.3 Details for Negative Speech Samples ($\tilde{S}_r$)

This subsection details the third branch of the negative-sample operator $\mathcal{N}$ in Eq. 5, namely the construction of $\tilde{\boldsymbol{S}}_r$ via TTS-based resynthesis. We employed a diverse set of TTS systems — both commercial and non-commercial — spanning a range of speaker-control mechanisms. Table 6 summarizes the systems and settings, while the rest of this subsection organizes the underlying strategies along two axes.

5.3.1 Two parallel families of speaker-control strategies

We group the strategies into two *parallel* families based on what is supplied to the TTS system as the speaker handle: *catalog-based* (the TTS exposes a fixed library of speakers / styles) and *ground-truth-conditioned* (the TTS is conditioned on the original speech itself):

- **Catalog-based.** The TTS system has a published catalog of speakers/styles, and we pick one entry per sample.
  - *Speaker Assignment.* We leveraged the publicly available speaker-style information released by each platform. For each input sample, `deepseek-v3` was used to select the best-matching catalog speaker based on the original speech caption. For `Gemini2.5TTS` only, the *character profile* $\boldsymbol{C}_{\text{profile}}$ (not the local scene) was additionally appended as an instruction to guide synthesis.

  - *Similar Voice Matching.* We used each platform's built-in voice-search function to identify the closest catalog voice to the ground-truth recording, and resynthesized the utterance with that voice.
  - *Voice Design.* The character profile $\boldsymbol{C}_{\text{profile}}$ was first converted into a descriptive voice specification by prompting `deepseek-v3` (Liu et al., 2024a). The description was submitted to the TTS service to generate candidate voices, and the voice that best matched the speaker style of the original recording (judged via Gemini2.5Pro) was selected before generation.

- **Ground-truth-conditioned.** The TTS is conditioned on audio derived from the ground-truth speaker, with three settings forming a sequence of progressively tighter conditioning:
  1. *Voice Clone (intermediate-tag).* Cloning uses the ground-truth voice as the source but routed through an intermediate representation (voice information / tag) rather than direct conditioning.
  2. *Identical Voice Clone.* The ground-truth voice is used directly as the conditioning input, with no intermediate tag.
  3. *Same-Speaker Voice Clone.* Instead of conditioning on the *specific* ground-truth utterance, we use *another* utterance from the same speaker, testing robustness to cross-utterance generalization.

For models using voice prompts in the ground-truth-conditioned family, we additionally apply an in-house enhancement model to first process the prompt audio so as to minimize the effect of audio events and background noise.

Table 6: TTS systems and generation settings used in Realism Evaluation.

System	**Type**	**Prompt Type**	**References**
OpenAI `TTS-1 HD`	Proprietary	Speaker Assignment	API Call (OpenAI, 2025c)
`Gemini2.5pro-preview-TTS`	Proprietary	Speaker Assignment with Instruction	API Call (Google Cloud, 2025)
ElevenLabs `Multilingual V2`	Proprietary	Similar Voice Matching	API Call (ElevenLabs, 2025)
	Proprietary	Voice Design	API Call (ElevenLabs, 2025)
	Proprietary	Voice Clone	API Call (ElevenLabs, 2025)
`HumeTTS`	Proprietary	Voice Design	API Call (Hume AI, 2025)
`OpenAudio-S1`	Proprietary	Identical Voice Clone	API Call (Fish Audio, 2025)
	Proprietary	Same Speaker Voice Clone	API Call (Fish Audio, 2025)
MiniMax `speech-02-hd`	Proprietary	Voice Design	API Call (Fish Audio, 2025)
	Proprietary	Voice Clone	API Call (Fish Audio, 2025)
`InhouseTTS1`	Proprietary	Identical Voice Clone	-
	Proprietary	Same Speaker Voice Clone	-
`InhouseTTS2`	Proprietary	Identical Voice Clone	-
	Proprietary	Same Speaker Voice Clone	-
`InhouseTTS3`	Proprietary	Identical Voice Clone	-
	Proprietary	Same Speaker Voice Clone	-
`InhouseTTS4`	Proprietary	Identical Voice Clone	-
	Proprietary	Same Speaker Voice Clone	-
`InhouseTTS5`	Proprietary	Identical Voice Clone	-
	Proprietary	Same Speaker Voice Clone	-
`CosyVoice2`	Open-source	Identical Voice Clone	(Du et al., 2024a)
	Open-source	Same Speaker Voice Clone	(Du et al., 2024a)
`Vevo1.5`	Open-source	Identical Voice Clone	(Zhang et al., 2025c)
	Open-source	Same Speaker Voice Clone	(Zhang et al., 2025c)
`MaskGCT`	Open-source	Identical Voice Clone	(Wang et al., 2025i)
	Open-source	Same Speaker Voice Clone	(Wang et al., 2025i)
`OpusLM-7B`	Open-source	Identical Voice Clone	(Tian et al., 2025)
	Open-source	Same Speaker Voice Clone	(Tian et al., 2025)

5.3.2 In-house Models.

We additionally employed five proprietary in-house TTS systems, each exploring different conditioning strategies for realism-oriented speech synthesis. All five fall under the ground-truth-conditioned family above (i.e., they produce $\tilde{\boldsymbol{S}}_r$ via cloning).

- `InhouseTTS1` (Phone–Cosy2 Pipeline). The input is a phone sequence produced by a graphme-to-phoneme (G2P) front-end. A 0.4B GPT-2 model (Radford et al., 2019) predicts `CosyVoice2` S3-tokens, which are then passed through a flow-matching model to generate mel-spectrograms. A neural vocoder from `CosyVoice2` converts the mel features into audio. Both the GPT-2 and flow-matching modules are augmented with speaker embeddings for speaker control.

- `InhouseTTS2` (BPE–Cosy2 Pipeline). Instead of phones, this model takes Qwen2.5 byte-pair-encoding (BPE) tokens as input (Xu et al., 2025). The same 0.4B GPT-2 predicts `CosyVoice2` tokens, followed by the identical flow-matching and vocoder stages as in `InhouseTTS1`. Speaker embeddings are also injected to maintain consistent speaker style.

- `InhouseTTS3` (Embedding–AudioCode Pipeline). Here, the input is Qwen2.5 BPE embeddings, which a 0.4B GPT-2 model maps directly to 4,096-dimensional audio codes. These codes are decoded via a cascaded architecture: `RepCodec` decoder (Huang et al., 2024b) → content representation (with speaker embedding) → `DDIM` (denoising diffusion implicit model) module (Jeong et al., 2021) → `DAC` latent representation (Kumar et al., 2023) → `DAC` decoder. Conditioning is provided through prefix text embeddings (e.g., instructions expressing emotional states such as "angry, hateful: you betrayed me..."), along with speaker ID. The input format is `[spk_id, BPE embedding, audio codes]`.

- `InhouseTTS4` (Instruction-Conditioned AudioCode Pipeline). This model also employs a 0.4B GPT-2 to predict 4,096-dimensional audio codes, but conditions generation on instruction embeddings. A `CLIPTextEncoder` (Radford et al., 2021) converts instruction text into embeddings, which serve as a prefix, combined with speaker ID and phone sequence. The input format is `[spk_id, instruction embedding sequence, phone sequence, audio codes]`.

- `InhouseTTS5` (Reference-Conditioned AudioCode Pipeline). Similar to `InhouseTTS4`, but instead of text-based instructions, conditioning is achieved through reference audio. A style encoder, following the NANCY++ design (Choi et al., 2023), compresses the reference audio into a fixed 32-dimensional embedding. This embedding is concatenated with the phone sequence and audio codes for generation. The input format is `[32-dim ref embedding, phone sequence, audio codes]`.

5.4 Realism Evaluation Data Summary.

The resulting realism evaluation dataset is constructed from a combination of curated media sources, including internal collections, public datasets (e.g., NCSSD (Liu et al., 2024b)), and original recordings collected for this project.[4] We compile 15k annotated role-play samples: 9k in Mandarin and 5k in English. The dataset is divided into 12k samples for training, 2k mixed samples for the base test set, and 1k real-recording samples for the test set. Splits are constructed at the character level: no character appearing in the training set appears in the basic test set, preventing memorization of character-specific vocal or stylistic patterns from inflating test correlations; the real-recording test set uses actors who are entirely absent from the training data. The first test set supports general evaluation using our proposed pipeline with negative samples, while the second focuses exclusively on human speech performed by both professional and amateur voice actors. For benchmarking, 2,000 scenarios from the realism dataset are designated for realism role-play benchmarking. Detailed statistics and analyses are provided in Appendix I.

[4]All data were obtained under appropriate licenses. To respect the terms of the underlying recordings, the public release distributes *mel-spectrogram features* rather than source waveforms; listenable audio is regenerated locally with an open-source, MIT-licensed neural vocoder (BigVGAN-v2 (Lee et al., 2023)), and pre-rendered vocoder reconstructions are additionally bundled for convenience. Because the released mel discards phase, reconstructions are model-generated approximations of—not copies of—the source recordings. See the Reproducibility and Data Availability statement (Sec. 10) for the full release format.

6 DRAME-EvalBench (Role-play Evaluation Modeling)

6.1 Experimental Setup

Building on the two proposed datasets, we conduct experiments under three settings: zero-shot, few-shot, and finetuning, using pre-trained ALLMs. We formulate the task as a multi-metric estimation problem (Tjandra et al., 2025; Yao et al., 2025; Shi et al., 2025b;a), where a single model is trained or prompted to predict multiple evaluation dimensions.

Zero-shot and Few-shot ALLMs. We first evaluate the zero-shot capabilities of both proprietary and public models, including `Gemini2.5Pro`, `GPT4o-audio`, `KimiAudio`, `Qwen2Audio`, and `Qwen2.5Omni`. All evaluations follow a unified prompting template that provides audio input alongside scene and character context.

For proprietary models, we report the average of five independent runs to reduce variance and obtain more stable predictions. For open-source models, we observed weaker instruction-following ability and limited adaptation to unfamiliar tasks. To mitigate this, we did not request simultaneous prediction of multiple metrics. Instead, we queried one metric at a time. Following Song & Bahri (2025); Lukasik et al. (2024), we request completion in a fixed template and extract the probability distribution over the discrete tokens `[1]`–`[5]`. The final score is then computed from these probabilities, which we found makes open-source models usable to some extent under this setting.

For few-shot experiments, we focus primarily on proprietary ALLMs, which demonstrate stronger instruction-following capabilities and have been widely used as judges in prior work. For each test query, we provide three additional annotated samples from the training set as in-context exemplars.[5]

Finetuning Model: `DRAME-Eval`. In addition to zero-shot and few-shot evaluations, we fine-tune a dedicated evaluation model, denoted as `DRAME-Eval` for archetype and realism role-play evaluation, respectively. We initialize from `Qwen2Audio-7B-Instruct` (Chu et al., 2024), which provides a strong foundation for multi-modal audio-text reasoning. To better leverage the multiple human annotations available in our datasets, we adopt a sampling-based training strategy: at each step, one annotator label is randomly selected, following practices in label distribution learning (Geng, 2016; Wen et al., 2023). This stochastic supervision implicitly exposes the model to the empirical annotation distribution, while during inference we output the full distribution over discrete labels (e.g., `[1]`–`[5]`) and compute the expectation as the final score, consistent with recent decoding-based regression methods (Song & Bahri, 2025; Xiao et al., 2025; Wang et al., 2025f; Lukasik et al., 2024).

Detailed setups and hyperparameters are detailed as follows: we primarily adopted a parameter-efficient strategy using LoRA (Hu et al., 2022), applied to the `Qwen2Audio-7B-Instruct` model (Chu et al., 2024). LoRA was configured with a rank of 16, $\alpha = 32$, and a dropout of 0.1, targeting the major projection layers, and trained for 10,000 steps to ensure stable convergence. The training pipeline employed a per-device batch size of 4 with gradient accumulation of 4 steps (effective batch size 32), AdamW optimization with cosine scheduling, a peak learning rate of 5×10^{-5}, weight decay of 0.01, and a warmup strategy defined by 500 steps. Gradient clipping with a norm of 1.0, `bf16` precision, and gradient checkpointing were enabled to stabilize training and reduce memory overhead. Distributed training was performed on 2xH100 GPUs using DeepSpeed (ZeRO stage-1).

While we utilize various prompts and sample them randomly during the fine-tuning experiments, we provide an example prompt as follows:

Example Prompts in Realism Evaluation (Pitch Dynamics Dimension)

```
You are an expert evaluator of speech delivery and storytelling.
```

[5]Please find our detailed settings such as prompts, decoding algorithms in Appendix J.1 with additional ablation analysis in Appendix J.2.

```
Based on the given information, rate the following SINGLE dimension on a 1-5 scale
(1 = poor, 5 = excellent).
Return ONLY a number between 1 and 5.

Global Profile: {global_profile}
Local Scene: {local_story}
Available Traits: {available_traits}

Dimension: PitchDynamics - Variety and appropriateness of pitch contours.

5: Natural, situation-appropriate melodic movement; subtle, believable inflections
that reveal emotion or emphasis.
4: Generally varied and fitting, but with minor stiffness or occasional flat
spots.
3: Neutral or limited inflection; neither distracting nor especially expressive.
2: Forced, exaggerated, or oddly sudden shifts that feel 'performed' rather than
lived.
1: Monotone or blatantly mis-pitched for the scene (e.g., cheerfully high in a
funeral).
Example output: 4"


Output the answer in <answer> </answer>.
```

We further conduct ablation studies in Sec. 6.3, examining: the choice of different base audio LLMs and full-parameter fine-tuning versus LoRA adaptation. These analyses highlight the robustness of our sampling–expectation training strategy and clarify the trade-offs in model design.

Evaluation Metrics for Role-play Evaluation Model. For most dimensions with 1-5 scale, we report Pearson correlation. For *Content Pass* estimation, we report accuracy.

6.2 DRAME-EvalBench Main Analysis

Archetype Evaluation. The results in Table 7 reveal clear distinctions across evaluation settings. In the zero-shot condition, proprietary ALLMs such as `Gemini2.5Pro` and `GPT4o-audio` substantially outperform public models, achieving strong *content pass* accuracy (∼90%) and moderate correlations with other dimensions (0.43–0.48). By contrast, public models exhibit limited predictive ability, with correlations often below 0.2. Few-shot prompting provides modest gains for proprietary models, improving both accuracy and correlation (e.g., `Gemini2.5Pro` rises from 0.480 to 0.495 average correlation). The most significant improvement arises from finetuning: `DRAME-Eval` achieves an average correlation of 0.629, marking a large margin over zero-shot/few-shot models and demonstrating the effectiveness of task-specific supervision.

Realism Evaluation. Table 8 extends the analysis to fine-grained realism evaluation. Similar to archetype results, proprietary ALLMs lead the zero-shot setting, with `Gemini2.5Pro` obtaining an average correlation of 0.390, compared to 0.284 for `GPT4o-audio`. Public models remain weak baselines. Few-shot prompting consistently improves performance: `Gemini2.5Pro` gains +0.042 and `GPT4o-audio` gains +0.083 over their zero-shot averages, showing the benefit of in-context supervision. Once again, finetuning is most effective: `DRAME-Eval` achieves 0.625 average correlation, higher than the best few-shot results. Importantly, the gains are consistent across all 10 realism dimensions, confirming that the sampling–expectation training strategy generalizes beyond simple classification to nuanced paralinguistic cues.

Real Recording Test Set. The real-recording evaluation in Table 9 is notably more challenging, with all models suffering substantial performance drops. Proprietary models in the zero-shot setting achieve only ∼0.04–0.08 average correlation, and public models often fall below zero. Few-shot prompting yields small but noticeable improvements, especially for `GPT4o-audio` (0.226 avg.), suggesting that exemplars partially

Table 7: Archetype role-play evaluator performance on the in-distribution archetype test set (English+Mandarin pooled). *Content Pass Acc.* is classification accuracy and is reported separately; *Audio Quality*, *Human Likeness*, and *Appropriateness* are Pearson correlations with human ratings. *Avg.* is the unweighted mean of those three correlation columns only and *does not* include the classification accuracy. `DRAME-Eval` is fine-tuned from `Qwen2Audio-7B-Instruct` (Chu et al., 2024); results for this model can therefore be directly compared with the `Qwen2Audio` zero-shot / few-shot rows.

Config.	Model	Content Pass Acc.	Audio Quality	Human Likeness	Appropriateness	Avg.
Zeroshot	`Gemini2.5Pro`	89.3%	0.465	0.483	0.492	0.480
	`GPT4o-audio`	89.9%	0.395	0.483	0.422	0.433
	`KimiAudio`	71.7%	0.089	0.132	0.121	0.114
	`Qwen2Audio`	86.7%	0.078	0.210	0.200	0.163
	`Qwen2.5Omni`	84.1%	-0.023	0.039	0.042	0.019
Fewshot	`Gemini2.5Pro`	91.1%	0.477	0.492	0.517	0.495
	`GPT4o-audio`	89.9%	0.296	0.374	0.377	0.349
Finetune	`DRAME-Eval`	**93.6%**	**0.682**	**0.680**	**0.525**	**0.629**

Table 8: Realism role-play evaluator performance on the *in-distribution* basic test set (English+Mandarin pooled). All ten columns are Pearson correlations between model and human scores. *Avg.* is the unweighted mean across all ten correlations. `DRAME-Eval` is fine-tuned from `Qwen2Audio-7B-Instruct` (Chu et al., 2024); its row can be directly compared with the `Qwen2Audio` zero-shot baseline to quantify the gain from fine-tuning on human-annotated data. Values are pooled across languages and reported without confidence intervals, and should therefore be read as a coarse summary (see the caveats at the end of Sec. 6.2).

Config.	Model	Prosodic Dynamics			Emotional Expressiveness			Character Consistency		Contextual Relevance		Avg.
		Pitch	Rhythm	Stress	Accuracy	Intensity	Transition	Identity	Traits	Local	Global	
Zeroshot	`Gemini2.5Pro`	0.319	0.352	0.336	0.391	0.316	0.336	0.529	0.509	0.349	0.465	0.390
	`GPT4o-audio`	0.132	0.210	0.179	0.323	0.258	0.273	0.360	0.397	0.277	0.440	0.284
	`KimiAudio`	0.045	0.033	0.065	0.018	0.020	0.024	0.072	0.013	0.009	0.026	0.032
	`Qwen2Audio`	0.029	0.030	0.051	-0.006	-0.007	-0.002	0.009	-0.023	0.056	-0.003	0.013
	`Qwen2.5Omni`	0.124	0.073	0.102	0.227	0.208	0.226	0.151	0.291	0.143	0.203	0.175
FewShot	`Gemini2.5Pro`	0.380	0.415	0.406	0.418	0.354	0.354	0.596	0.524	0.376	0.497	0.432
	`GPT4o-audio`	0.236	0.264	0.230	0.417	0.344	0.412	0.422	0.514	0.332	0.498	0.367
Finetune	`DRAME-Eval`	**0.621**	**0.627**	**0.631**	**0.632**	**0.596**	**0.601**	**0.660**	**0.655**	**0.563**	**0.668**	**0.625**

Table 9: Realism role-play evaluator performance on the *real-recording* test set (1k samples, professional + amateur human voice actors; English+Mandarin pooled). All ten columns are Pearson correlations between model and human scores; *Avg.* is their unweighted mean. This test set is intentionally out-of-distribution with respect to the synthesis bank used to construct training data: every sample is real human acting rather than synthesized speech. It is reported as a measurement of the synthetic-to-real transfer gap rather than as a setting on which fine-tuning is claimed to be human-aligned.

Config.	Model	Prosodic Dynamics			Emotional Expressiveness			Character Consistency		Contextual Relevance		Avg.
		Pitch	Rhythm	Stress	Accuracy	Intensity	Transition	Identity	Traits	Local	Global	
Zeroshot	`Gemini2.5Pro`	0.159	0.070	0.103	0.070	0.052	0.096	0.092	0.077	0.075	0.028	0.082
	`GPT4o-audio`	0.067	-0.013	0.035	0.013	0.042	0.036	-0.004	0.130	0.065	0.018	0.039
	`KimiAudio`	-0.115	-0.039	-0.116	-0.020	-0.122	-0.080	**0.114**	-0.032	-0.137	-0.080	-0.063
	`Qwen2.5Omni`	0.078	0.071	0.043	0.074	0.047	0.045	-0.082	0.076	0.102	0.098	0.055
	`Qwen2Audio`	0.055	0.057	0.088	0.072	0.077	0.057	-0.006	-0.040	0.050	0.111	0.052
FewShot	`Gemini2.5Pro`	**0.319**	0.179	0.263	0.111	0.078	0.184	-0.001	0.118	0.216	0.074	0.166
	`GPT4o-audio`	0.299	0.193	0.222	0.121	**0.377**	0.344	-0.013	**0.248**	**0.301**	0.168	0.226
Finetune	`DRAME-Eval`	0.288	**0.270**	**0.362**	**0.277**	0.331	**0.397**	-0.034	0.077	0.217	**0.279**	**0.247**

mitigate the domain mismatch between synthetic and real human speech. Fine-tuning attains the highest overall average (0.247), but the gap to the best few-shot baseline is small ($\sim$0.02 on average, likely within sampling noise absent confidence intervals) and several dimensions remain weak: *Voice-Identity-Matching* is slightly negative (-0.034), and *Trait Embodiment* is near zero (0.077). We do *not* read these numbers as evidence that `DRAME-Eval` is human-aligned on real human speech; the honest reading is that it is a strong in-distribution regressor that does not yet generalize to authentic recordings on exactly the dimensions where realism evaluation is supposed to add value. One plausible contributing factor — which we flag as a hypothesis rather than a measured finding — is that, with every test sample being natural human speech, naturalness differences no longer carry discriminative signal and the judge must rely on subtler acting cues;

the negative correlations on some dimensions, however, indicate genuine distribution shift beyond mere task difficulty. Reporting this real-recording test set is intended to make the synthetic-to-real transfer gap a measured quantity rather than an unstated caveat, so that future work can treat it as a primary optimization target.

Overall. Across both archetype and realism evaluations, three trends are consistent: (i) zero-shot performance is limited, particularly for open-source models, (ii) few-shot prompting offers moderate but consistent gains for proprietary ALLMs, and (iii) fine-tuning with task-specific supervision (`DRAME-Eval`) yields the largest improvements on the in-distribution test sets. The real-recording numbers, however, qualify these in-distribution gains: the next research step is not to push the in-distribution Pearson higher, but to close the synthetic-to-real transfer gap that the same model exposes.

Caveats on the headline numbers. Two reporting choices are worth flagging here. First, all averages in Tabs. 7–9 are pooled across English and Mandarin; given the rater-pool argument in Sec. 5 and the language-specific score gaps observed in App. H, the pooled Pearson should be read as a coarse summary, with per-language tables being the cleaner unit of comparison. Second, for *Emotion Intensity* and *Emotion Transition*, ratings are imputed to 1 when the upstream *Emotion Accuracy* gate triggers `N/A` (Sec. 5.2 and App. I); this inflates the apparent correlation when both the human label and the model's prediction collapse to the floor. Treating gated entries as missing (mask, do not impute) is the cleaner protocol; we retain the imputed convention here for comparability with the annotation protocol, and note it as a known upward bias on those two columns.

6.3 Fine-tuning Ablation Experiments

Tables 10, 11, and 12 present ablation studies across different fine-tuning strategies, including the use of `Qwen2Audio-7B-Base` or `Qwen2Audio-7B-Instruct`; and the use of fully fine-tuning or LoRA fine-tuning. We use `Base-FT`, `Base-LoRA`, `Instruct-FT`, and `Instruct-LoRA` to represents different configuration settings in following experiments.

Archetype evaluation. On the archetype test set, all fine-tuned variants achieve strong *Content Pass* accuracy above 92%. Differences appear mainly in correlation-based metrics. Here, `Base-FT` yields the highest score on *Audio Quality* and *Appropriateness*, while `Instruct-LoRA` delivers the strongest performance in *Human Likeness* and the best overall average (0.629). This suggests that instruction-tuned checkpoints paired with parameter-efficient adaptation can better capture higher-level perceptual traits.

Realism evaluation (basic set). For realism evaluation on the basic test set, `Instruct-FT` consistently achieves the strongest correlations across most dimensions, especially for *Character Identity* and *Traits*, leading to the best average correlation (0.668). `Base-LoRA` also performs competitively (0.630), while `Instruct-LoRA` lags slightly behind (0.625). These results highlight that full fine-tuning may still offer an advantage when the evaluation set is relatively homogeneous and closer to the training distribution.

Realism evaluation (real recordings). A different pattern emerges on the real recording test set. Both `Base-LoRA` and `Instruct-FT` struggle, yielding negative or near-zero correlations across most dimensions. In contrast, `Instruct-LoRA` shows clear improvements, achieving positive correlations in all prosodic and emotional dimensions and producing the highest overall average (0.247). This indicates that LoRA adaptation on top of an instruction-tuned backbone provides better generalization to out-of-distribution, real-world data.

Overall summary. While `Instruct-FT` is competitive on in-domain evaluation, it suffers a substantial drop when applied to real recordings. By contrast, `Instruct-LoRA` offers a more robust balance: it maintains reasonable performance on archetype and basic sets, while significantly outperforming other settings on real-world data. Based on this robustness, we adopt `Instruct-LoRA` as our final configuration for subsequent experiments.

Table 10: Archetype role-play evaluator fine-tuning ablation on the in-distribution test set. *Content Pass Acc.* is classification accuracy and is reported separately; the remaining columns are Pearson correlations with human ratings. *Avg.* is the unweighted mean of those three correlation columns only (Audio Quality, Human Likeness, Appropriateness) and excludes the classification accuracy.

Config.	Content Pass Acc.	Audio Quality	Human Likeness	Appropriateness	Avg.
`Base-FT`	92.1%	**0.688**	0.648	**0.551**	0.616
`Base-LoRA`	92.6%	0.668	0.648	0.545	0.620
`Instruct-FT`	92.6%	0.650	0.684	0.492	0.609
`Instruct-LoRA`	**93.6%**	0.682	**0.680**	0.525	**0.629**

Table 11: Realism role-play evaluator fine-tuning ablation on the in-distribution basic test set. All ten columns are Pearson correlations with human ratings. *Avg.* is the unweighted mean across all ten correlation columns.

Config.	Prosodic Dynamics			Emotional Expressiveness			Character Consistency		Contextual Relevance		Avg.
	Pitch	Rhythm	Stress	Accuracy	Intensity	Transition	Identity	Traits	Local	Global	
`Base-FT`	0.643	0.633	0.655	0.657	0.600	0.599	0.689	0.668	0.611	0.707	0.646
`Base-LoRA`	0.628	0.633	0.643	0.637	0.586	0.590	0.668	0.651	0.578	0.684	0.630
`Instruct-FT`	**0.657**	**0.655**	**0.675**	**0.674**	**0.627**	**0.626**	**0.752**	**0.685**	**0.615**	**0.714**	**0.668**
`Instruct-LoRA`	0.621	0.627	0.631	0.632	0.596	0.601	0.660	0.655	0.563	0.668	0.625

Table 12: Realism role-play evaluator performance (real recording test set). We report Pearson correlation coefficients for other dimensions. The Avg. column is the average of all coefficients.

Config.	Prosodic Dynamics			Emotional Expressiveness			Character Consistency		Contextual Relevance		Avg.
	Pitch	Rhythm	Stress	Accuracy	Intensity	Transition	Identity	Traits	Local	Global	
`Base-FT`	-0.422	-0.119	-0.333	-0.261	-0.330	-0.351	0.059	0.538	-0.3083	-0.153	-0.168
`Base-LoRA`	-0.372	-0.016	-0.304	-0.195	-0.273	-0.249	-0.091	0.458	-0.291	-0.146	-0.148
`Instruct-FT`	-0.269	-0.031	-0.202	-0.208	-0.222	-0.242	**0.076**	**0.543**	-0.219	-0.094	-0.087
`Instruct-LoRA`	**0.288**	**0.270**	**0.362**	**0.277**	**0.331**	**0.397**	-0.034	0.077	0.217	**0.279**	**0.247**

7 Conclusion

We presented **Speech-DRAME**, a unified framework for assessing spoken role-play through two complementary lenses: *Archetype Evaluation* (top-down) and *Realism Evaluation* (bottom-up, human-grounded). Our human-annotated **DRAME-EvalBench** — including contrastive negative samples constructed via the operator $\mathcal{N}$ — enables training and testing of `SEval` models, and our fine-tuned **DRAME-Eval** further improves agreement with human raters, on the in-distribution test sets, over already-capable zero-/few-shot ALLM judges. On a separate real-recording test set composed entirely of human acting, however, correlation remains modest (0.247 average); we report this synthetic-to-real gap as a measured open problem rather than a solved one. Together, these resources are intended as a reproducible, analyzable foundation for spoken role-play assessment, and we view closing the synthetic-to-real gap and using `DRAME-Eval` as an automatic judge for ranking `SFnd` systems as the natural next steps.[6]

8 Limitations and Future Works

Despite the design effort behind Speech-DRAME, several limitations remain. First, while our dual evaluation paradigm combines top-down archetype assessment with bottom-up realism evaluation, both strategies are constrained by the quality and scope of available data. Archetype evaluation inevitably relies on simplified stereotypes, which can overlook subtler paralinguistic and contextual cues. Realism evaluation, though richer, still struggles with domain mismatch between synthetic training data and real human recordings, as evidenced by the performance drop observed on the real-recording test set. A plausible contributing factor (stated here as a hypothesis, not a measured finding) is that when every test sample is natural human speech, naturalness differences no longer carry discriminative signal and judgments must rest on subtler performance cues; the negative correlations on some dimensions, however, point to genuine distribution shift beyond task

[6]Limitations and future works are discussed in Appendix 8.

difficulty alone. This gap highlights how much room remains between benchmark-driven evaluation and the full demands of real-world acted speech.

Second, the present framework emphasizes bilingual evaluation (Mandarin and English), but linguistic and cultural diversity is far broader. Cross-lingual generalization and culturally sensitive evaluation are still open challenges, particularly since we observed annotation discrepancies across language groups. This limits the extent to which Speech-DRAME can claim universal coverage.

Third, while DRAME-Eval substantially outperforms zero-shot and few-shot ALLMs, its reliance on fine-tuning from a strong base model raises questions of accessibility and scalability. Future work should explore lightweight adaptation techniques, semi-supervised training with partially annotated data, and leveraging emergent capabilities of open-source multimodal LLMs.

Fourth, the reported correlations are point estimates, pooled across English and Mandarin, and without confidence intervals or a human–human agreement ceiling. Per-language breakdowns, bootstrap confidence intervals, and standard chance-corrected inter-rater agreement (e.g., ordinal Krippendorff's α) would make the comparisons more rigorous; this matters most on the real-recording test set, where the gaps between the best systems are small and their ranking may not be statistically separable.

Finally, although we validated DRAME-Eval against human annotations, alignment under realism settings remains modest. This suggests that even human-annotated benchmarks are insufficiently fine-grained to capture the full complexity of human perception in speech role-play. Richer annotation protocols, more diverse speaker populations, and integration of perceptual psychology insights could further improve evaluation fidelity.

Looking forward, we see several promising research directions. We intentionally restrict evaluation to single-turn responses for clarity and tractability. However, multi-turn extensions that capture narrative progression, long-term consistency, and dynamic emotion shifts remain a natural next step. Incorporating multilingual and cross-cultural datasets will enhance generality and fairness. On the modeling side, exploring hybrid evaluation strategies that combine discrete rubrics with continuous perceptual signals (e.g., prosody embeddings or affective measures) may better bridge the gap between automatic and human judgments. Finally, integrating Speech-DRAME into reinforcement learning pipelines as a reward model could help drive the next generation of expressive, human-aligned speech role-play systems.

Multi-turn extension of the formulation. The single-turn equations of Sec. 3.1 extend to multi-turn role-play by reintroducing a preceding spoken-dialogue context $\boldsymbol{C}_{\text{speech}}$ alongside the persistent role and scene specifications:

$$\boldsymbol{S}_r \;=\; \text{SFnd}(\boldsymbol{C}_{\text{speech}},\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}), \qquad \boldsymbol{E}_r \;=\; \text{SEval}(\boldsymbol{S}_r,\, \boldsymbol{C}_{\text{speech}},\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}).$$

All other components — the curation operators $\mathcal{D}^{\text{top-down}}$, $\mathcal{D}^{\text{bottom-up}}$, and the negative-sample operator $\mathcal{N}$ — carry over unchanged. Because all data and experiments in this paper are single-turn, $\boldsymbol{C}_{\text{speech}} = \emptyset$ throughout the released benchmark; this is the only difference between the body and the multi-turn formulation.

9 Ethics Statement

All data collection and annotation in this work followed strict ethical standards. Our bilingual datasets were constructed exclusively from legally accessible sources, including licensed media, open-source corpora, and original recordings conducted with informed consent. Annotators were recruited with appropriate compensation and provided with clear instructions, including the right to withdraw at any point. To minimize risks, no personally identifiable information was included in the released data, and all recordings were anonymized before annotation. We ensured cultural and linguistic diversity in annotator pools to mitigate potential bias across Mandarin and English role-play scenarios. While this study did not require formal IRB approval, we adhered to best practices in human-subject research, including transparency in task design and careful consideration of annotator well-being. Full details of our annotation protocol and ethical safeguards are presented in Appendix G.

10 Reproducibility and Data Availability

We emphasize reproducibility as a core design principle of Speech-DRAME, and we release **DRAME-EvalBench** together with the toolchain needed to reproduce our results. Our experiments were conducted on NVIDIA H100, and random seeds are fixed across all runs to ensure consistency.

Released artifacts. The public release comprises both evaluation perspectives — *Archetype* (8,272 annotated records; 6,773 train / 1,499 test) and *Realism* (15,414 annotated records; 10,990 train / 2,000 base test / 2,164 real-recording test / 260 benchmark) — with the full annotation detail used in this paper: per-dimension and per-sub-dimension scores, per-rater label arrays, confidence, and, for Realism, the contrastive-negative metadata produced by the operator $\mathcal{N}$ (`negative_type`, perturbed components, and perturbation severity). Each record is validated against a published JSON Schema, and the release ships per-split manifests, SHA-256 checksums, a datasheet, per-source license notes, and regenerated score-distribution statistics. We additionally release the training recipes and evaluation code for the `SEval` models and the benchmarking pipeline for both cascaded and end-to-end `SFnd` systems, with standardized prompting templates.

Audio as features. To respect the licenses of the underlying recordings, audio is distributed as *mel-spectrogram features* rather than source waveforms. Listenable audio is regenerated locally with an open-source, MIT-licensed neural vocoder (BigVGAN-v2 (Lee et al., 2023)), whose exact analysis front-end is pinned in the release so that the shipped features are directly vocodable; a self-contained recovery script and an offline Griffin–Lim fallback are included. For convenience the release also bundles pre-rendered vocoder reconstructions, so the data can be auditioned without a GPU or a vocoder install. Because the released mel discards phase, every reconstruction is a model-generated approximation of—not a copy of—the source recording, and is flagged as such in the metadata.

Hosting and review. The benchmark is hosted on Hugging Face Datasets behind a host-agnostic layout, and is accompanied by a static documentation site. For the double-blind review period, all artifacts are anonymized: the dataset, the documentation site, and the citation metadata contain no author, affiliation, or other identifying information. The release toolkit and documentation site are available for review at `https://anonymous.4open.science/r/speech_drame_evalbench-0DF3`; the dataset will be made public on Hugging Face Datasets upon de-anonymization. By releasing the datasets, models, and code in this form, we aim to make all major results in this paper fully reproducible while honoring the redistribution terms of every source.

References

Shinjitsu Agatsuma, Reon Ohashi, Kazuya Tsubokura, Yua Nishio, Mai Ishikawa, Niina Ito, Fukuka Ito, Shiori Minami, Nao Takegawa, Riko Nakamura, et al. Building a role-play interactive system using llm for health guidance education. In *2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pp. 1–3. IEEE, 2024.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918, 2024.

Tanja Bänziger and Klaus R Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3-4):252–267, 2005.

Mary Bucholtz and Kira Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614, 2005.

ByteDance. Doubao - ai chat agent, 2025. URL `https://www.doubao.com/chat/`.

Joseph Campbell. *The hero with a thousand faces*, volume 17. New World Library, 2008.

Xuankai Chang, Jiatong Shi, Jinchuan Tian, Yuning Wu, Yuxun Tang, Yihan Wu, Shinji Watanabe, Yossi Adi, Xie Chen, and Qin Jin. The Interspeech 2024 Challenge on Speech Processing Using Discrete Units. In *Interspeech 2024*, pp. 2559–2563, 2024. doi: 10.21437/Interspeech.2024-1878.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Ensiong Chng. Audio large language models can be descriptive speech quality evaluators. In *International Conference on Learning Representations*, volume 2025, pp. 24920–24934, 2025.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*, pp. 3670–3674, 2021.

Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, et al. Mint-bench: A comprehensive multilingual benchmark for instruction-following text-to-speech. *arXiv preprint arXiv:2604.17958*, 2026.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*, 2024a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024b.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, et al. Audio-aware large language models as judges for speaking styles. *arXiv preprint arXiv:2506.05984*, 2025.

Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. NANSY++: Unified voice synthesis with neural analysis and synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=elDEe8LYW7-`.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4): 161–183, 2024.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Yuhao Du, Qianwei Huang, Guo Zhu, Zhanchen Dai, Sunian Chen, Qiming Zhu, Yuhao Zhang, Li Zhou, and Benyou Wang. Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols. *arXiv preprint arXiv:2508.18240*, 2025.

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

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

ElevenLabs. Multilingual v2 — elevenlabs models documentation. `https://elevenlabs.io/docs/models#multilingual-v2`, 2025. ElevenLabs Documentation.

Fish Audio. Openaudio s1 — ai text-to-speech and voice cloning. `https://openaudios1.com/`, 2025. Fish Audio AI TTS model documentation (demo site).

Dongjie Fu, Fangming Feng, Xize Cheng, Linjun Li, Zhou Zhao, and Tao Jin. Character beyond speech: Leveraging role-playing evaluation in audio large language models via reinforcement learning. *arXiv preprint arXiv:2604.13804*, 2026.

Leah Fung and Rory Laing. A proof of concept study on the use of large language models as a client in typed role plays for training therapists. *Discover Psychology*, 4(1):201, 2024.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748, 2016.

Erving Goffman. Presentation of self in everyday life. *American Journal of Sociology*, 1949.

Google Cloud. Gemini-tts — cloud text-to-speech api documentation. `https://cloud.google.com/text-to-speech/docs/gemini-tts`, 2025. Google Cloud — Cloud Text-to-Speech API.

Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

Stella Guldner, Nadine Lavan, Clare Lally, Lisa Wittmann, Frauke Nees, Herta Flor, and Carolyn McGettigan. Human talkers change their voices to elicit specific trait percepts. *Psychonomic Bulletin & Review*, 31(1):209–222, 2024.

John J Gumperz. *Discourse strategies*. Cambridge University Press, 1982.

Ilya Gusev. Pingpong: a benchmark for role-playing language models with user emulation and multi-model evaluation. *arXiv preprint arXiv:2409.06820*, 2024.

Lori L Holt and Andrew J Lotto. Speech perception as categorization. *Attention, Perception, & Psychophysics*, 72(5):1218–1227, 2010.

Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. Sova-bench: Benchmarking the speech conversation ability for llm-based voice assistant. *arXiv preprint arXiv:2506.02457*, 2025a.

Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. SOVA-bench: Benchmarking the speech conversation ability for llm-based voice assistant. In *Proc. Interspeech 2025*, pp. 5713–5717, 2025b.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ailin Huang, Bingxin Li, Bruce Wang, Boyong Wu, Chao Yan, Chengli Feng, Heng Wang, Hongyu Zhou, Hongyuan Wang, Jingbei Li, et al. Step-audio-aqaa: a fully end-to-end expressive large audio language model. *arXiv preprint arXiv:2506.08967*, 2025a.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12136–12140. IEEE, 2024a.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations*, volume 2025, pp. 82908–82973, 2025b.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025c.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. The singing voice conversion challenge 2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Zhichao Huang, Chutong Meng, and Tom Ko. Repcodec: A speech representation codec for speech tokenization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5777–5790, 2024b.

Hume AI. Voice design — hume api documentation, 2025. URL `https://dev.hume.ai/docs/voice/voice-design`.

Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. In *Proc. Interspeech 2021*, pp. 3605–3609, 2021.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Hui Li, Xiaoran Fan, Ming Zhang, Junjie Ye, Shihan Dou, Zhiheng Xi, et al. Speechrole: A large-scale dataset and benchmark for evaluating speech role-playing agents. *arXiv preprint arXiv:2508.02013*, 2025.

Tom Johnstone and Klaus R Scherer. Vocal communication of emotion. *Handbook of emotions*, 2:220–235, 2000.

CG Jung. *The Archetypes and the Collective Unconscious*. Routledge, 2014.

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Eva Székely, and Joakim Gustafson. Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation. In *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Nadine Lavan. The time course of person perception from voices: A behavioral study. *Psychological science*, 34(7):771–783, 2023.

Sang Gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Yejin Lee, Jaehoon Kang, and Kyuhong Shim. P2va: Converting persona descriptions into voice attributes for fair and controllable text-to-speech. *arXiv preprint arXiv:2505.17093*, 2025.

Wenyu Li, Xiaoqi Jiao, Yi Chang, Guangyan Zhang, and Yiwen Guo. Audiorole: An audio dataset for character role-playing in large language models. *arXiv preprint arXiv:2509.23435*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models. *arXiv preprint arXiv:2505.15727*, 2025.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4187–4196, 2024b.

Ruohan Liu, Shukang Yin, Tao Wang, Dong Zhang, Weiji Zhuang, Shuhuai Ren, Ran He, Caifeng Shan, and Chaoyou Fu. Speechparaling-bench: A comprehensive benchmark for paralinguistic-aware speech generation. *arXiv preprint arXiv:2604.20842*, 2026.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. RoleMRC: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21008–21030, 2025.

Michal Lukasik, Harikrishna Narasimhan, Aditya Menon, Felix Yu, and Sanjiv Kumar. Regression aware inference with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13667–13678, 2024.

Siyuan Ma, Weidi Luo, Yu Wang, and Xiaogeng Liu. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773*, 2024.

Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Held, and Diyi Yang. Audiojudge: Understanding what works in large audio model based speech evaluation. *arXiv preprint arXiv:2507.12705*, 2025.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*, 2025.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech 2021*, pp. 2127–2131, 2021.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

OpenAI. Chatgpt — model: Gpt-4o. `https://chatgpt.com/?model=gpt-4o`, 2025a. ChatGPT interface specifying model GPT-4o.

OpenAI. Gpt-4o-realtime-preview — model documentation. `https://platform.openai.com/docs/models/gpt-4o-realtime-preview`, 2025b. OpenAI Platform Docs.

OpenAI. Tts-1 hd — model documentation. `https://platform.openai.com/docs/models/tts-1-hd`, 2025c. OpenAI Platform Docs.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly, and Simon King. Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023. *Computer Speech & Language*, 90:101747, 2025.

Robert J Podesva and Patrick Callier. Voice quality and identity. *Annual review of applied Linguistics*, 35: 173–194, 2015.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, et al. Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 679–686. IEEE, 2021.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. PersonaPlex: Voice and role control for full duplex conversational speech models. *arXiv preprint arXiv:2602.06053*, 2026.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Richard Schechner. *Performance studies: An introduction*. Routledge, 2017.

Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand, Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. MULTIVOX: A benchmark for evaluating voice assistants for multimodal interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28469–28481, 2025.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 255–265, 2017.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*, 2023.

Jiatong Shi, Yifan Cheng, Bo-Hao Su, Hye-jin Shim, Jinchuan Tian, Samuele Cornell, Yiwen Zhao, Siddhant Arora, and Shinji Watanabe. Arecho: Autoregressive evaluation via chain-based hypothesis optimization for speech multi-metric estimation. *arXiv preprint arXiv:2505.24518*, 2025a.

Jiatong Shi, Hye jin Shim, and Shinji Watanabe. Uni-VERSA: Versatile Speech Assessment with a Unified Network. In *Interspeech 2025*, pp. 1798–1802, 2025b. doi: 10.21437/Interspeech.2025-1960.

Yemin Shi, Yu Shu, Siwei Dong, Guangyi Liu, Jaward Sesay, Jingwen Li, and Zhiting Hu. Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play. *arXiv preprint arXiv:2505.02707*, 2025c.

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26312–26332, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1349. URL `https://aclanthology.org/2025.findings-acl.1349/`.

Xingyou Song and Dara Bahri. Decoding-based regression. *arXiv preprint arXiv:2501.19383*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, Huck Yang, and Shinji Watanabe. OpusLM: A Family of Open Unified Speech Language Models. In *Interspeech 2025*, pp. 3259–3263, 2025. doi: 10.21437/Interspeech.2025-1184.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, 2024.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pp. 11836–11850. Association for Computational Linguistics (ACL), 2024.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025a.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, et al. Speechllm-as-judges: Towards general and interpretable speech quality evaluation. *arXiv preprint arXiv:2510.14664*, 2025b.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Voiceassistant-Eval: Benchmarking ai assistants across listening, speaking, and viewing. *arXiv preprint arXiv:2509.22651*, 2025c.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6372–6391, 2025d.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14743–14777, 2024a.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. Benchmarking contextual and paralinguistic reasoning in speech-LLMs: A case study with in-the-wild data. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14133–14148, Suzhou, China, November 2025e. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.760. URL `https://aclanthology.org/2025.findings-emnlp.760/`.

Siyang Wang and Éva Székely. Evaluating text-to-speech synthesis from a large discrete token-based speech language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6464–6474, 2024.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving llm-as-a-judge inference with the judgment distribution. *arXiv preprint arXiv:2503.03064*, 2025f.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*, 2025g.

Xihuai Wang, Ziyi Zhao, Siyu Ren, Shao Zhang, Song Li, Xiaoyu Li, Ziwen Wang, Lin Qiu, Guanglu Wan, Xuezhi Cao, et al. Audio turing test: Benchmarking the human-likeness of large language model-based text-to-speech systems in chinese. *arXiv preprint arXiv:2505.11200*, 2025h.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 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)*, pp. 1840–1873, 2024b.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *The Thirteenth International Conference on Learning Representations*, 2025i.

Brett Welch, Miriam R van Mersbergen, and Leah B Helou. Listeners' perceptions of speaker personality traits based on speech. *Journal of speech, language, and hearing research*, 64(12):4762–4771, 2021.

Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. Ordinal label distribution learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23481–23491, 2023.

Weihao Wu, Liang Cao, Xinyu Wu, Zhiwei Lin, Rui Niu, Jingbei Li, and Zhiyong Wu. Voxrole: A comprehensive benchmark for evaluating speech-based role-playing agents. *arXiv preprint arXiv:2509.03940*, 2025.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. From role-play to drama-interaction: An llm solution. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 3271–3290, 2024.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.730. URL `https://aclanthology.org/2025.findings-emnlp.730/`.

Tim Z. Xiao, Robert Bamler, Bernhard Schölkopf, and Weiyang Liu. Verbalized machine learning: Revisiting machine learning with language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=k3Ab6RuJE9`.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*, 2025b.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, et al. Vstyle: A benchmark for voice style adaptation with spoken instructions. *arXiv preprint arXiv:2509.09716*, 2025.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186. IEEE, 2022.

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. OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26318–26331, July 2025a.

Jian Zhang, Linhao Zhang, Bokai Lei, Chuhan Wu, Wei Jia, and Xiao Zhou. Wildspeech-bench: Benchmarking audio llms in natural speech conversation. *arXiv e-prints*, pp. arXiv–2506, 2025b.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025d.

Haishu Zhao, Aokai Hao, Yuan Ge, Zhenqiang Hong, Tong Xiao, and Jingbo Zhu. Stylebench: Evaluating speech language models on conversational speaking style control. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17542–17546. IEEE, 2026.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. Personaeval: Are LLM evaluators human enough to judge role-play? In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=drdrFhKYjP`.

Appendices

Appendix Contents

Appendix Contents **34**

A Use of Large Language Models **35**

B Extension to Multi-turn Role-play Evaluation **35**

C Implementation Details for Archetype Evaluation Tasks **36**

C.1 Role and Scene Categories . . . 36

C.2 Pipeline Illustration . . . 37

C.3 Annotation Platform for Archetype Evaluation . . . 39

D Annotation Guidelines for Archetype Evaluation **39**

D.1 Initial Assessment & Rejection Instructions . . . 39

D.2 Rubric Rating Instructions . . . 40

D.3 Commentary Guidelines and Confidence Ratings . . . 40

D.4 Full Annotation Guideline (Role-play Archetype Evaluation) . . . 40

E Implementation Details for Realism Evaluation Tasks **43**

E.1 Prompts for Character Profile Creation . . . 43

E.2 Prompts for Local Scene Generation . . . 45

E.3 Prompts for Negative Scenes . . . 46

E.4 Examples of Realism Role-play . . . 51

E.5 Annotation Platform for Realism Evaluation . . . 52

F Annotation Guidelines for Realism Evaluation **53**

F.1 Initial Assessment & Rejection Instructions . . . 53

F.2 Rubric Rating Instructions . . . 53

F.3 Commentary Guidelines and Confidence Ratings . . . 54

F.4 Full Annotation Guideline (Realism Evaluation) . . . 54

G Annotation Ethics **58**

H Detailed Data Analysis (Archetype Evaluation) **59**

I Detailed Data Analysis (Realism Evaluation) **61**

J Evaluation Model Experiments **62**

J.1 Zeroshot/Fewshot Setup for Evaluation Models 62
J.2 Zeroshot/Fewshot Ablation Experiments 68

A Use of Large Language Models

Large language models (LLMs) and audio-capable LLMs were used in four capacities in this work:

1. **Data generation:** LLMs were employed to expand prompts (e.g., character profiles, scene descriptions) and to produce negative samples in realism evaluation. All outputs were screened by human annotators to ensure validity and reduce bias.
2. **Baseline evaluators:** State-of-the-art audio LLMs (e.g., `GPT4o`, `Gemini2.5Pro`, `Qwen2.5Omni`) were included as zero-shot judges in our evaluation benchmark, enabling direct comparison with human-aligned `SEval` models.
3. **Annotation design and task construction:** `Gemini2.5Pro` was used to assist with building instructions and summarizing story contexts. Drafted materials were subsequently refined by the authors and validated by human annotators.
4. **Manuscript preparation:** LLMs were used for polishing, proofreading, and improving clarity of the paper. All conceptual design, analysis, and substantive writing remain the work of the authors, with LLMs serving only as language assistants.

Proprietary LLM outputs were used in a limited way (e.g., instruction building, story summarization, or baseline evaluations) and only for internal research purposes. Released datasets in Speech-DRAME are constructed from open-source models, licensed corpora, and original human recordings.

B Extension to Multi-turn Role-play Evaluation

Our main formulation (Section 3.1) focuses on the *single-turn* setting, where each response $\boldsymbol{S}_r$ is generated and evaluated independently. This simplification enables clearer analysis of speech role-play quality and aligns with our bilingual annotation protocol. Nevertheless, the same formulation extends naturally to multi-turn dialogues.

Multi-turn generation. Consider a dialogue of T turns, with context $\{\boldsymbol{C}^{(t)}_{\text{speech}}\}_{t=1}^{T}$, profile $\boldsymbol{C}_{\text{profile}}$, and scene $\boldsymbol{C}_{\text{scene}}$. At each turn t, the SFM produces

$$\boldsymbol{S}_r^{(t)} \;=\; \text{SFM}\big(\boldsymbol{C}^{(t)}_{\text{speech}},\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}\big). \tag{6}$$

Multi-turn evaluation. Each generated speech segment is scored by an SEM:

$$\boldsymbol{E}_r^{(t)} \;=\; \text{SEM}\big(\boldsymbol{S}_r^{(t)},\, \boldsymbol{C}_{\text{profile}},\, \boldsymbol{C}_{\text{scene}}\big). \tag{7}$$

This yields a sequence of per-turn scores $\{\boldsymbol{E}_r^{(1)}, \ldots, \boldsymbol{E}_r^{(T)}\}$.

Dialogue-level aggregation. A dialogue-level evaluation can be obtained by applying an aggregation operator Ψ over turn-level scores:

$$\bar{\boldsymbol{E}}_r \;=\; \Psi\big(\boldsymbol{E}_r^{(1)}, \ldots, \boldsymbol{E}_r^{(T)}\big), \tag{8}$$

where Ψ may be a simple mean, a recency-weighted average, or a learned temporal model that captures consistency across turns. The choice of Ψ reflects the intended evaluation focus, e.g., stability of persona traits, adaptation to scene progression, or coherence across multiple speakers.

Discussion. While our released dataset and experiments concentrate on single-turn evaluation, the modular nature of the formulation ensures that extending to multi-turn is straightforward. Future work can leverage Ψ to study richer aspects of speech role-play, such as persona persistence, interaction dynamics, and role consistency across extended dialogues.

C Implementation Details for Archetype Evaluation Tasks

This appendix provides concrete examples and extended discussion of the archetype evaluation annotation pipeline that are omitted in the main text for clarity.

C.1 Role and Scene Categories

We constructed archetype evaluation prompts by combining role categories with contextual scenes. The following illustrates the core categories with representative examples.

Roles.

- **Occupation / Functional Role in Modern Daily Life.** Example:
  - *You are a firefighter. Thick smoke fills the building, and you discover a trapped civilian. You say:*
- **Occupation / Functional Role in Fictional or Fantasy Settings.** Example:
  - *You are the leader of a dark cult in a martial arts fantasy world. The righteous alliance has come to challenge you outside your stronghold. You say:*
- **Social Identity.** Example:
  - *You are an overprotective parent. Your child has not come home late at night, and you cannot reach them by phone. You say:*
  - *You are a laid-back "go-with-the-flow" friend. Your colleague excitedly tells you about a highly competitive promotion and urges you to go for it. You say:*
- **Named / Specific Characters.** Example:
  - *You are Sherlock Holmes. In the courtroom, your rival lawyer mocks your deductions in front of the jury. You remain unimpressed. You say:*
- **Personality Traits.** Example:
  - *Your personality is defined as hot-blooded. Your team is losing a game, and in the final moments you must deliver the decisive score. You say:*
  - *Your personality is defined as constantly complaining. You are dining at a restaurant where the food is bland and the service disappointing. You say:*
- **Relative Roles.** Example:
  - *You are Parent. Your child returns home late at night after not answering calls. You say:*
  - *You are Stranger. You see a stranger's bag is unzipped and items are about to fall out, so you decide to alert them. You say:*

Scenes.

- **Everyday External Event / State Change.** Example:
  - *Your boss calls you into the office and tells you that because of your excellent performance, you are being promoted with a raise. You return to your desk, unable to hide a smile. You say:*
  - *You overhear coworkers gossiping and realize you have become the subject of their rumors. Your face turns red. You say:*

Excluded Dimensions. We initially considered more fine-grained control dimensions, but excluded them to preserve the archetype-level focus. Examples of excluded dimensions include: *Emotional State & Expression, Single Emotion Prompts, Explicit User Input / Assistant Output structures, Emotional Transition / Fluctuation, Voice Style & Prosody (pacing, intonation, vocal timbre, expressive style, accent/dialect), and Content / Interaction Patterns (task/topic controls, specific linguistic structures, or speech production challenges).*

These dimensions are more closely aligned with explicit instruction-following or style transfer rather than broad stereotype judgment, and they reduced clarity in human evaluations.

The Choice of Prompts. We also experimented with broader, less specific templates that allowed more freedom for generative models but produced less discriminative evaluation results:

- *You are [role]. You usually say:*
- *You are a {{ role }} in the {{ universe }}. Someone approaches you for help. You say:*

While LLMs often responded fluently to these broad templates, human annotators found judgments to be more ambiguous and tolerant, making scores less sharp. The final design therefore emphasizes more concrete, vivid prompts, as elaborated in the above examples in Appendix C.1.

C.2 Pipeline Illustration

Text generation. We followed a three-step expansion: keyword → scene description → full prompt with template. Some examples are shown as follows:

- Occupation: firefighter → "smoke fills the building, you discover a trapped civilian" → *You are a firefighter. Smoke fills the building, and you discover a trapped civilian. You say:*
- Fictional role: warrior in a fantasy world → "you witness bullies harassing a weaker person" → *You are a warrior in a martial arts fantasy world. You see bullies harassing a weaker person. You say:*
- Event: promotion → "your boss calls you in to announce a raise" → *Your boss calls you into the office and announces you are being promoted with a raise. You return to your desk, smiling uncontrollably. You say:*

All steps are automatically processed by `Gemini2.5Pro` (Comanici et al., 2025), but each output is manually verified to ensure high quality for downstream usage. We demonstrate an example for Roles: Occupation / Functional Role in Fictional or Fantasy as follows:

Step 1: Generate fictional universes

```
Instruction:  Generate {{n_samples}} common fictional or cinematic universes as
keywords, and output them in JSONL format.  The universes should cover a wide range
of genres (e.g., cyberpunk, martial arts wuxia, urban fantasy, high fantasy magic),
avoiding redundancy or over-concentration.

Output format:
{"universe_en": "..."}
{"universe_en": "..."}
```

Step 2: Generate occupations/roles

Instruction: As a rigorous character designer, prepare {{n_samples}} occupation or functional role types for the given universe.

Examples:
- Fantasy Magic: mage, knight, dragon, bard, saint, demon, demon lord, elf, dwarf, goblin
- Science Fiction: interstellar explorer, AI robot, cyborg, genetic engineer, space pirate, colonial governor
- Post-apocalyptic: survivor, scavenger, raider, mutant, dungeon guard, resistance member
- Cosmic Horror: investigator, cultist, occult scholar, apostle of the Old Gods, cursed individual, mad prophet

Output format:

{"role_en":"...", "universe_en":"..."}
{"role_en":"...", "universe_en":"..."}

Step 3: Final prompt generation

Instruction: As a top-tier role-play scenario designer, create concise and effective prompts for evaluating voice-based role-play. Each prompt must directly activate the role's *core personality and speaking style* without explicitly describing mannerisms (e.g., "gently," "decisively").

Requirements:
1. **Conciseness**: Each scenario_en must be minimal and focus on the essential trigger.
2. **Personality activation**: The scene must elicit distinctive identity, social status, and communication style (tone, vocabulary, rhythm).
3. **Immersion**: Scenarios should create immediate role immersion with tension or iconic role traits.
4. **Personality-scenario linkage**: Each prompt must include a personality_analysis summarizing the role's dominant traits, which align with the designed scenario.

Example output:

```
{"role_en": "Demon Lord", "universe_en": "Fantasy Magic",
 "personality_analysis": "arrogant, scornful, controlling",
 "scenario_en": "A captured human hero defiantly
     glares at you before your throne."}

{"role_en": "Demon", "universe_en": "Fantasy Magic",
 "personality_analysis": "cunning, patient, manipulative",
 "scenario_en": "You tempt a desperate mortal,
     promising to grant their wish."}
```

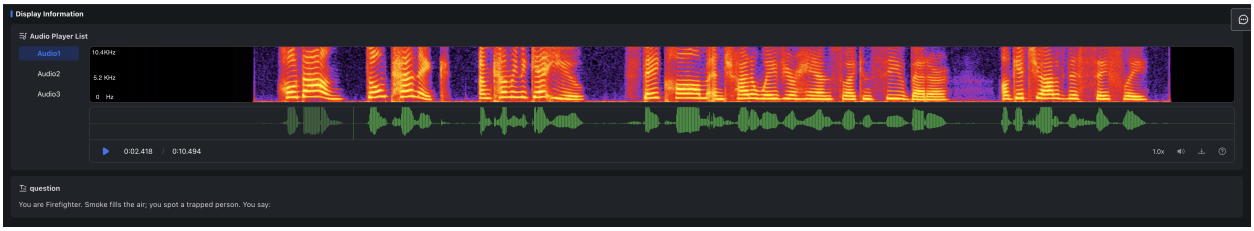
Final expansion for each role keyword ({{n_samples}} scenarios per role):

```
{"role_en": "{{ keyword.role_en }}",
 "universe_en": "{{ keyword.universe_en }}"}
```

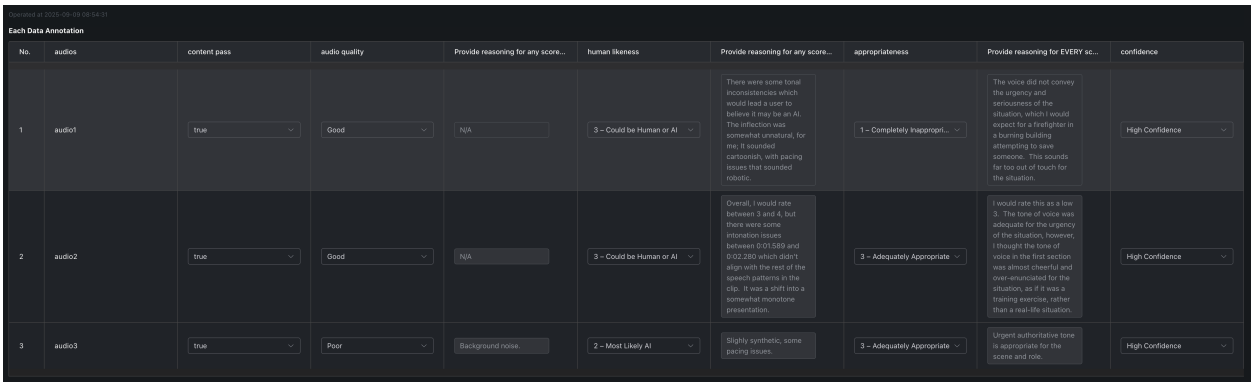
Speech generation. Prompts were rendered into audio with CosyVoice2-0.5B Du et al. (2024b). The prompts are selected from human-verified high-quality samples, where the selection is conducted from open-source datasets, such as LibriSpeech/Gigaspeech for English, and WenetSpeech for Chinese (Panayotov et al., 2015; Chen et al., 2021; Zhang et al., 2022).

C.3 Annotation Platform for Archetype Evaluation

Figure 4 illustrates the internal annotation interface used for *Archetype Evaluation*. The platform provides an integrated environment for both playback and evaluation of model-generated role-play speech samples. Annotators first interact with the audio player interface (subfigure 4a), which visualizes waveforms and spectrograms for multiple versions of the same prompt, enabling side-by-side comparison under a shared role instruction (e.g., firefighter, teacher, or actor). Then, they proceed to the annotation interface (subfigure 4b), where they assign scores for multiple dimensions such as *Audio Quality*, *Human Likeness*, and *Appropriateness*, along with reasoning and confidence inputs. This structured interface supports high-consistency, fine-grained evaluation of role-play behaviors in speech models.



(a) System interface showing the role-play audio playback and visualization components, including waveform and spectrogram views for each model output.



(b) Annotation interface where evaluators rate each sample across multiple perceptual and contextual dimensions with reasoning and confidence entries.

Figure 4: Annotation platform for Archetype Evaluation. Subfigure (a) presents the system's audio visualization and playback interface, while subfigure (b) shows the structured annotation form for rating and reasoning across key perceptual dimensions.

D Annotation Guidelines for Archetype Evaluation

This appendix contains the full human annotation instructions used for archetype evaluation.

D.1 Initial Assessment & Rejection Instructions

- Focus on the first 30 seconds only.
- Reject if the generation is irrelevant, breaks role instructions, or contains severe defects.

- Reject if foreign-language content is included (except for widely recognizable words).
- Reject if the clip is too short (1–2s) or clearly breaks immersion.

Detailed case examples are provided in the full instruction document.

D.2 Rubric Rating Instructions

We adopt three scales, each rated from 1 (worst) to 5 (best):

- **Audio Quality (No Context):** independence from content, focus on distortions or synthesis artifacts.
- **Human Likeness (No Context):** perceived naturalness and plausibility of human delivery.
- **Appropriateness (Context Considered):** tonal fit of the vocal performance to the described role and scene.

The complete rubrics with per-score definitions are reproduced below:

D.3 Commentary Guidelines and Confidence Ratings

Annotators are required to provide short, clear justifications:

- Comments are mandatory for Appropriateness (all scores).
- Comments for Audio Quality and Human Likeness are only required for scores ≤ 3.
- Gender mismatches should be flagged in comments but not factored into rubric scoring.

Annotators also provide a confidence level (high, moderate, low) for their Appropriateness rating, reflecting familiarity with the role and scene. This provides additional granularity in analysis.

D.4 Full Annotation Guideline (Role-play Archetype Evaluation)

Role-play Archetype Evaluation

`Goal. The overall goal of the initial assessment for each audio is to determine if any of the clips generated from each model meet any criteria to receive a REJECTION (referred to as "content pass: fail" in our tasking interface).`

Initial Assessment & Rejection Instructions. Listen to each audio *in full*, paying special attention to the following caveats and rejection criteria.

Caveats

- **30-second focus:** Assessment should focus on the first 30 seconds of the generation ONLY. Do not consider any content past 30s; disregard it entirely from your assessment and rubric ratings. (You may stop listening beyond 30s.)
- **Early breakdowns:** If repetition, restarts, or obvious breakdowns occur *before* 30s (or for total clips shorter than 30s), the clip may be rejected.
- **Excessively short clips:** If a clip is only 1–2 seconds, reject due to insufficient information for reliable rubric ratings.

- **Foreign language content:** If *any* part before 30s contains a foreign language (even a single word/short phrase) in otherwise English audio, reject the entire clip.
  Exception: Extremely common/recognizable borrowed words (e.g., "croissant") may pass. Niche items (e.g., "coq au vin") should be rejected.

Rejection Criteria

1. **Instruction Failure.** Generated content does not adhere to the trait(s) or role described in the prompt ("question").
   - *Example (reject):* Prompt: "Your personality is sharp-tongued. Your friend asks about their new haircut." Audio says: "It looks great on you." ⇒ reject.
   - *Example (pass):* Same prompt; audio says: "It looks like your stylist put you through a wood chipper." ⇒ content pass = true.
   - *Fourth-wall breaks:* Reject any generation acknowledging the scenario or narrating its role (e.g., "I am a lawyer in a courtroom...") or speaking for imagined other parties (two-way dialogue narration).
   - *Stage-direction speak:* Phrases like "pats you on the back reassuringly" are grounds for rejection.
   - *Placeholders:* Very short/abstract placeholders like "nurse" or "concept" are acceptable; longer placeholders like "and you can fill in the details here" are rejectable.
2. **Irrelevance.** Generated content is entirely unrelated to the prompt.
   - *Example (reject):* Prompt: "You are a kindergarten teacher. You are helping a crying child." Audio: "I'd like to thank everyone for this prestigious award." ⇒ reject.
   - *Example (pass):* "Oh no. Are you hurt? Don't worry. We'll get you taken care of, little one." ⇒ content pass = true.

Important: Once an audio receives content pass = false, *do not* rate rubrics; proceed to the next clip.

Rubric Rating Instructions (for content pass = true). We shift focus from *what* was said to *how* it was said (vocal performance). Use gut intuition guided by rubric definitions. Score each audio *in isolation* (no ranking across clips).

Commentary

- After each 1–5 rating, add a short comment explaining your score.
- For **Audio Quality** & **Human Likeness**: comment only if score ≤ 3; if 4 or 5, write "N/A".
- For **Appropriateness**: comment for *all* scores 1–5. Explain *why* the *delivery* was/was not a good match to the in-context traits (role & scene).
- Do *not* critique word choice or logic; past the initial rejection step, only the *performance* matters.
- Prefer brief, clear, non-technical comments; no need to quote transcripts.

Gender Mismatch Flagging (comments only)

If the prompt indicates a specific gender, flag mismatches in comments (e.g., "GENDER MISMATCH"). Do *not* penalize rubric scores for gender mismatch; rate vocal performance as-is.

Confidence

Provide a confidence level (High / Moderate / Low) for your *Appropriateness* rating, reflecting your familiarity with the role/scene.

- Low confidence if unfamiliar with the role/scene (e.g., "Air Traffic Controller").

- High confidence if very familiar (e.g., "Grandmother").
- Lower confidence *will not* reflect poorly on annotators; it adds nuance to analysis.

Rating Rubrics: General Caveats

- Judge within traditional *stereotypes* for the prompt; when in doubt between two scores, choose the *lower*.
- If any portion of the clip is ill-fitting, prefer the lower overall score; do not average across good and bad segments.
- Evaluate the *performance*, not the script content.
- Pronunciation errors: penalize only on **Human Likeness** if they significantly affect comprehension; do not affect **Appropriateness**.

Audio Quality (No Context)

Independently rate technical clarity (ignore content and fit). After scoring Audio Quality, *disregard* its issues for the other two dimensions.

Score	Definition
5 Excellent	Clean, clear, high-quality throughout; no digital noise/glitches/issues.
4 Good	Good overall; minor artifacts detectable on close listen (slight sharpness/hiss, etc.).
3 Fair	Acceptable but obvious artifacts (buzz, echo, synthetic timbre) impacting clarity.
2 Poor	Frequent/consistent glitches or severe issues; distracting metallic/clipping/harsh tones.
1 Very Poor	Extremely poor; layered artifacts dominate; severe clipping/static/failures, hard to understand.

Human Likeness (No Context)

Rate how likely an average listener would believe the clip is human-delivered (ignore Audio Quality issues and contextual fit).

Score	Definition
5 Definitely Human	Indistinguishable from human: rich intonation, natural pacing, consistent identity; organic pauses/breaths, logical inflection.
4 Most Likely Human	Generally convincing; slight stiffness/over-even pacing or minor awkward pauses.
3 Could be Human or AI	Mix of natural and unnatural traits; borderline cases with both good and bad sections.
2 Mostly Likely AI	Clearly artificial patterns: choppy timing, missing pauses, flat/illogical inflections, forced breaths.
1 Definitely AI	Rigid, machine-like delivery throughout; flat/monotone, "read-aloud" style, no natural rhythm.

Appropriateness (Context Considered)

Rate *only* how well the *delivery* fits the role/scene described in the prompt (tonal fit). Ignore Audio Quality and Human Likeness issues while scoring this dimension.

Score / Criteria	Definition
5 Completely Appropriate	Fully fits BOTH role and specific scene; aligns with emotional stance, social dynamics, and situational events; strong immersion.
4 Mostly Appropriate	Minor issues: role slightly off but scene reaction good; or scene slightly off but role good; or both slightly off with no major issues. Strong overall match with small flaws (e.g., intensity/subtlety).
3 Adequately Appropriate	Generally acceptable for BOTH role and scene but lacking depth; noticeable inconsistencies.
2 Slightly Appropriate	EITHER role *or* scene reaction is clearly not acceptable; weak alignment or mismatched/missing elements.
1 Completely Inappropriate	BOTH role and scene are contextually mismatched; tone/urgency/status illogical; heavy misfit with expected alignment.

E Implementation Details for Realism Evaluation Tasks

This appendix starts with a summary of the task preparation pipeline for our proposed realism evaluation. Then, we provides the detailed prompts and system setups for constructing the Realism Evaluation dataset described in Section 5.

E.1 Prompts for Character Profile Creation

We designed prompts to extract or generate concise role descriptions corresponding to characters in TV, film, or natural dialogue sources. These descriptions emphasize identity, relationship, and communicative goals.

Example Prompt: Character Profile (Retrieval)

```
You are a character-description assistant.


Task
I will give you a character name and the name of the show or game they are from.
Your job is to check the internet and return a concise description of the character
that focuses on how they behave when interacting with others.


Input
Character: {speaker}
Context: {show_or_game}


Description guidelines
- Focus specifically on the character's behavior and personality, especially how
they express themselves when talking to others.
- Cover all aspects of how they interact with people, so that their overall
communication style and relationships are clear.
- Do NOT mention the show or game name in the description.
- *Never* mention or speculate about accents.
- The profile should contain NO guessing words like "maybe", "might", "could be",
```

"likely", etc.
- The profile must be internally consistent, include **no contradictions**.
- If the context is a game, do NOT mention anything about interacting with a player. Only describe how the character interacts with other characters.
- Do NOT include any actor or voice actor information. - Lilith's example: Lilith is known for her exceptionally cold and emotionally detached demeanor. Her interactions are often characterized by a flat affect, precise articulation, and a generally humorless approach. She speaks in a measured, almost clinical tone, rarely betraying strong emotions. When interacting with others, she is highly intelligent and articulate, but also critical and often condescending, especially towards those she deems less intellectually gifted. She has a knack for delivering cutting remarks with a dispassionate facade, which can make her intimidating and unsettling. She seems incapable of expressing warmth or affection in a conventional manner, often making others feel uncomfortable or inadequate in her presence.
- Use less than 100 words for profile description.

Return value (strict JSON, minified)
"‘json
"character":"character name","context":"show or game name","full_profile":"full profile"

"‘

Example Prompt: Character Profile (Reasoning)

You are a **character-reasoning assistant**.

Task
Analyze the entire dialogue below and infer a profile of **Speaker {target_speaker}**. Draw on context from every participant plus the emotion/audio captions, but focus your profile solely on Speaker {target_speaker}.

Dialogue:
{dialogue_context}

Emotion caption for Speaker {target_speaker} (if available):
{emotion_caption}

Audio caption for Speaker {target_speaker} (if available):
{audio_caption}

Profile-writing guidelines
- **Never** mention or speculate about accents.
- Please use the name of character indicated from the dialogue context, but if not, you must make a creative guess based on the dialogue. Do this for both speakers.
- Do NOT use generic names like "Speaker 0" or "Speaker 1".

```
- The profile should contain NO guessing words like "maybe", "might", "could be",
"likely", etc.
- The profile must be internally consistent, include no contradictions.
- Ignore or omit any discussion of conversational transitions or turn-taking.
- Where details are missing, make informed, creative inferences to complete the
portrait (see the "Lilith" example you provided for the desired depth and tone).
- Lilith's example:  Lilith is known for her exceptionally cold and emotionally
detached demeanor.  Her interactions are often characterized by a flat affect,
precise articulation, and a generally humorless approach.  She speaks in a
measured, almost clinical tone, rarely betraying strong emotions.  When interacting
with others, she is highly intelligent and articulate, but also critical and often
condescending, especially towards those she deems less intellectually gifted.  She
has a knack for delivering cutting remarks with a dispassionate facade, which can
make her intimidating and unsettling.  She seems incapable of expressing warmth
or affection in a conventional manner, often making others feel uncomfortable or
inadequate in her presence.
- Use less than 100 words for profile description.


Return value (strictly minified JSON)
"`json
{{"full_profile":"your profile text here", "speaker0":  "name of speaker 0",
"speaker1":  "name of speaker 1"}}
"`
```

E.2 Prompts for Local Scene Generation

To provide consistent evaluation context, we generate local scene descriptions summarizing the situational background. These prompts aim to capture the immediate narrative context without introducing extraneous detail.

Example Prompt: Local Scene (Reasoning)

```
Task:
Given the following inputs:
- Episode information from a TV show or movie
- A full transcript of a scene (either a dialogue or a monologue)
- General background information about a specific character
- A selected span of three consecutive sentences spoken by that character

Your goal is to analyze the localized story context surrounding those three lines,
focusing on the character's intent, emotional state, and interaction dynamics.

First, integrate the episode information to understand the broader plot, timeline,
and narrative setup.  Then, use the character's general background to ground your
interpretation of their behavior, personality traits, and emotional expressions.
These two layers of context (episode-level and character-level) should inform your
understanding of the local moment in the transcript.
Try to use the name refer in the character general information instead of speaker
0 or speaker 1.  Even if no information is inlcuded, simple use speaker WIHOUT
number.
```

```
Inputs:  {{Character general information}}
{{Full transcript}}
{Character of interest}
{Three consecutive sentences spoken by that character}


Output Format (in a JSON format, with the following fields):
{{
"Immediate_Situation":  "Describe what is happening in the scene just before and
during these lines.  Include relevant environmental or narrative context.",
"Character_Intention":  "What is the character trying to do or communicate through
these lines?",
"Emotional_Subtext":  "What emotions are being expressed, suppressed, or shifted?",
"Emotion_Subtext_List":  "List the sequence of emotions by their occurrences,
separated by commas.  ONLY use emotions from the following list:  {'Happy', 'Sad',
'Angry', 'Fearful', 'Disgusted', 'Surprised', 'Neutral'}.",
"Character_Dynamics":  "How do these lines reflect or affect the character's
relationship with others in the scene?",
"Overall_Summary":  "The summarization of the above aspects."
}}


The user input is:
{user_input}
```

E.3 Prompts for Negative Scenes

Negative scenes are created by altering or mismatching the original context at varying levels of severity. This section provides the prompting templates and representative outputs.

Example Prompt: Negative Local Scene (Level0)

```
You are given:
- A character profile describing personality, habits, and tone.
- A speech transcript the character says.
- A real story where both the character and the transcript fit naturally.


Your task is to generate a fake local story that maintains surface plausibility but
introduces a controlled mismatch.  Return a JSON object with the following fields:


- 'fake_story':  A new scene where the character and transcript still appear
naturally.
- 'reason':  A short explanation of why the transcript is subtly off in this new
scene.


Constraints:
- Do not modify the character profile or speech transcript.
- Do not repeat the transcript in your fake_scene_description.
- Do not introduce new audio events unless they are essential to the mismatch.
```

- The mismatch should arise from one of the following: **situational logic**, **emotional tone**, or **stylistic clash** -- not from sound cues or transcript changes.

Keep character and transcript fitting in tone, but change the local context so the transcript becomes subtly unnecessary, exaggerated, or slightly illogical.

Inputs:
- Character Profile: {character_global_story}
- Transcript: {local_transcript}
- Real Story: {real_story}

Example Prompt: Negative Local Scene (Level1)

You are given:
- A character profile describing personality, habits, and tone.
- A speech transcript the character says.
- A real story where both the character and the transcript fit naturally.

Your task is to generate a fake local story that maintains surface plausibility but introduces a controlled mismatch. Return a JSON object with the following fields:

- 'fake_story': A new scene where the character and transcript still appear naturally.
- 'reason': A short explanation of why the transcript is subtly off in this new scene.

Constraints:
- Do **not** modify the character profile or speech transcript.
- Do **not** repeat the transcript in your fake_scene_description.
- Do **not** introduce new audio events unless they are essential to the mismatch.
- The mismatch should arise from one of the following: **situational logic**, **emotional tone**, or **stylistic clash** -- not from sound cues or transcript changes.

The transcript and context still logically connect, but the emotional tone is off (e.g., too intense, too calm, too cheerful given the context).

Inputs:
- Character Profile: {character_global_story}
- Transcript: {local_transcript}
- Real Story: {real_story}

Example Prompt: Negative Local Scene (Level2)

You are given:
- A character profile describing personality, habits, and tone.

```
- A speech transcript the character says.
- A real story where both the character and the transcript fit naturally.


Your task is to generate a fake local story that maintains surface plausibility but
introduces a controlled mismatch.  Return a JSON object with the following fields:


- 'fake_story':  A new scene where the character and transcript still appear
naturally.
- 'reason':  A short explanation of why the transcript is subtly off in this new
scene.


Keep the scene and character presence, but make the transcript's style feel
unnatural for the character (e.g., too logical, too emotional, too formal, etc.).


Constraints:
- Do not modify the character profile or speech transcript.
- Do not repeat the transcript in your fake_scene_description.
- Do not introduce new audio events unless they are essential to the mismatch.
- The mismatch should arise from one of the following:  situational logic,
emotional tone, or stylistic clash -- not from sound cues or transcript changes.


Inputs:
- Character Profile:  {character_global_story}
- Transcript:  {local_transcript}
- Real Story:  {real_story}
```

Example Prompt: Negative Global Scene (Level0)

```
You are given:
- A fixed transcript (the exact lines spoken -- do not change this).
- The original character profile (describing personality, tone, style).
- The real story context (scene and situation where the transcript was originally
spoken).


Your task is to generate a fake character profile and a lightly modified scene
context, such that the same transcript is still plausibly spoken, but introduces a
graded mismatch in personality, tone, or identity.


Return a JSON object with the following fields:


- 'fake_profile':  A description of a new character who still plausibly says the
transcript, but whose identity or traits introduce a mismatch.
- 'fake_story':  A version of the real story context adapted to fit this new
character.
```

```
- 'reason':  A short explanation of how the transcript feels off given the new
character profile (superficially or fundamentally).


Change surface traits (e.g., gender, age, species, name) while preserving emotional
and stylistic alignment.  The transcript sounds natural emotionally, but voice or
social expectations may cause slight dissonance.


Constraints (All Levels):
- Do not modify the transcript.
- Keep the character's presence plausible in the story.
- Only alter global identity traits -- avoid local scene logic changes unless
needed to accommodate the new character.
- Do not invent or rely on new audio events unless absolutely essential.


Inputs:
- Character Profile:  {character_global_story}
- Transcript:  {local_transcript}
- Real Story:  {real_story}
```

Example Prompt: Negative Global Scene (Level1)

```
You are given:
- A fixed transcript (the exact lines spoken -- do not change this).
- The original character profile (describing personality, tone, style).
- The real story context (scene and situation where the transcript was originally
spoken).


Your task is to generate a fake character profile and a lightly modified scene
context, such that the same transcript is still plausibly spoken, but introduces a
graded mismatch in personality, tone, or identity.


Return a JSON object with the following fields:


- 'fake_profile':  A description of a new character who still plausibly says the
transcript, but whose identity or traits introduce a mismatch.
- 'fake_story':  A version of the real story context adapted to fit this new
character.
- 'reason':  A short explanation of how the transcript feels off given the new
character profile (superficially or fundamentally).


Introduce a tone or speaking style that clashes with the transcript (e.g., stiff
character says something casual, aloof character says something emotional).  The
mismatch is noticeable but still believable.
```

Constraints (All Levels):
- Do **not** modify the transcript.
- Keep the character's presence plausible in the story.
- Only alter **global identity traits** -- avoid local scene logic changes unless needed to accommodate the new character.
- Do **not** invent or rely on new audio events unless absolutely essential.

Inputs:
- Character Profile: {character_global_story}
- Transcript: {local_transcript}
- Real Story: {real_story}

Example Prompt: Negative Global Scene (Level2)

You are given:
- A fixed transcript (the exact lines spoken -- do not change this).
- The original character profile (describing personality, tone, style).
- The real story context (scene and situation where the transcript was originally spoken).

Your task is to generate a *fake character profile* and a lightly modified scene context, such that the same transcript is still plausibly spoken, but introduces a graded mismatch in personality, tone, or identity.

Return a JSON object with the following fields:

- 'fake_profile': A description of a new character who still plausibly says the transcript, but whose identity or traits introduce a mismatch.
- 'fake_story': A version of the real story context adapted to fit this new character.
- 'reason': A short explanation of how the transcript feels off given the new character profile (superficially or fundamentally).

Create a character whose personality, values, or emotional posture fundamentally contradict the transcript. Their saying it requires unusual circumstances (e.g., parody, breakdown, irony).

Constraints (All Levels):
- Do **not** modify the transcript.
- Keep the character's presence plausible in the story.
- Only alter **global identity traits** -- avoid local scene logic changes unless needed to accommodate the new character.
- Do **not** invent or rely on new audio events unless absolutely essential.

```
Inputs:
- Character Profile:  {character_global_story}
- Transcript:  {local_transcript}
- Real Story:  {real_story}
```

E.4 Examples of Realism Role-play

To make the construction of the realism evaluation task concrete, we present a worked example that begins with a curated local scene and character profile, followed by the same transcript adapted into mismatched contexts. These mismatches appear either at the *local-scene* level or at both the *character profile and scene* level, thereby providing controlled degradations of realism.

We use the character Yunjin from *Genshin Impact* as a running case study, which is also included in our released dataset.[7] This example illustrates how alignment between a speaker's persona, local context, and delivery underpins realism, and how systematically breaking that alignment produces contrastive test cases.

Curated Local Scene. Yunjin is narrating a fierce battle through operatic singing. Her verses depict a hero charging into enemy lines, slaying the enemy commander, and securing peace. The atmosphere is filled with tension and intensity, highlighting the hero's extraordinary bravery and determination.

Character Profile. Yunjin is a passionate opera performer with deep knowledge of traditional singing art. Her lyrics vividly portray battlefield struggles and heroic deeds, conveying admiration for those who stand firm against overwhelming odds. She always seeks to inspire her audience with courage, justice, and hope, channeling her artistry into both entertainment and moral elevation.

Original Transcript (English Translation).

> **Yunjin:** Armored riders charge into the enemy lines, no longer entangled with the blue-clad foes.
> **Yunjin:** Deep amidst the horde of demons, the enemy commander is seized, his red robe cut down by the sword.
> **Yunjin:** A blade of frosty steel strikes fear into their hearts, calming the storms and securing peace.
> **Yunjin:** It is said,
> **Yunjin:** Though thousands of soldiers guard the pass, one hero holds it alone; even with blades upon their body, they remain undaunted.

Negative Local Scenes. To stress realism evaluation, we insert the same transcript into alternative backgrounds. These reveal how context-shifted delivery degrades narrative coherence:

- **Level 0 (Mild Mismatch).** *Fake Scene:* A quiet teahouse where Yunjin casually chats with friends about recent weather. She suddenly delivers the heroic verses in full operatic style. *Reason:* The shift from light conversation to intense battle imagery feels abrupt and unsupported. The mismatch is mild but noticeable.
- **Level 1 (Moderate Mismatch).** *Fake Scene:* A teahouse with elderly patrons enjoying tea in a calm setting. Yunjin, dressed plainly, begins singing passionately. *Reason:* The intense tone clashes with the subdued environment, startling listeners. Her heroic lines sound out of place, generating tonal dissonance instead of inspiration.
- **Level 2 (Severe Mismatch).** *Fake Scene:* A modern café with soft jazz music and casual chatter. Yunjin sips a latte while reading a magazine, then recites the battle verses. *Reason:* The dramatic imagery of warriors and swords is wholly incompatible with the relaxed, contemporary café ambience. The mismatch is extreme, rendering the delivery absurd.

[7] As acknowledged, the use of real-world recordings is conducted under legal and ethical review. All curated data will be released as part of our open-source package.

Negative Character Profiles. We replace Yunjin with new character profiles and scenes. Although the transcript remains the same, realism degrades because the delivery no longer matches the speaker's persona or situation.

- **Level 0 (Mild Mismatch).** *Fake Scene:* At a modern tech company's annual party, a shy programmer is asked to perform. Under pressure, he recites opera lines he once saw online. *Fake Profile:* Zhang Wei, an introverted software engineer with little artistic background, usually immersed in code and algorithms. *Reason:* The heroic verses sound alien coming from someone unfamiliar with opera or traditional culture. The clash produces awkward humor but still feels like a coping attempt.
- **Level 1 (Moderate Mismatch).** *Fake Scene:* In a futuristic laboratory, a robotic engineer celebrates a successful combat simulation. Overwhelmed with excitement, he blurts out the opera verses. *Fake Profile:* Xiao A, a calm and rational roboticist, normally concise and analytical, not prone to poetic expression. *Reason:* The dramatic, metaphorical verses conflict with his logical persona and sterile environment. The mismatch is sharper: the heroic rhetoric feels uncharacteristic and jarring.
- **Level 2 (Severe Mismatch).** *Fake Scene:* On a modern city street, a rebellious street performer in jeans and a rock-band T-shirt strums an electric guitar, suddenly shouting the verses in parody. *Fake Profile:* Mark, an anti-traditionalist performer who mocks classical art and heroic ideals, favoring satire and irony. *Reason:* The verses, originally meant to inspire reverence, become ironic and absurd when delivered by someone who rejects those values. The clash is extreme, generating intentional ridicule and total loss of realism.

Discussion: Realism Example vs. Archetype Examples.

The Yunjin case study illustrates a fundamental difference between realism evaluation and archetype evaluation. Archetype prompts, as described in Appendix C.1, are constructed by pairing abstract role categories with synthetic contextual scenes. These cover a broad space of possible situations, ranging from modern occupations (e.g., firefighters) to fictional personas (e.g., cult leaders in martial arts fantasy) or social identities (e.g., overprotective parents, laid-back friends). While such prompts provide controlled and diverse coverage, they are ultimately template-based and rely on stereotypes to elicit role-play behaviors.

In contrast, the Yunjin example is drawn from authentic performance, where the character's persona (an opera singer) and local scene (a battlefield narration) are naturally aligned through artistic delivery. The realism task does not merely test whether a model can follow a prompt, but whether it can preserve coherence between a speaker's global identity, immediate context, and expressive prosody. When Yunjin's transcript is transplanted into mismatched settings or paired with incongruent speaker profiles, the breakdown in realism becomes apparent: the same words shift from inspiring to awkward, jarring, or even absurd.

This distinction highlights the complementary strengths of the two evaluation paradigms. Archetype evaluation systematically explores a wide combinatorial space of roles and situations, but may lack the paralinguistic and contextual richness of real speech. Realism evaluation, on the other hand, anchors assessment in authentic delivery and contextual fit, providing a finer-grained lens on coherence and naturalness. Together, they ensure that role-play evaluation spans both broad stereotypical coverage and nuanced real-world plausibility.

E.5 Annotation Platform for Realism Evaluation

The annotation interface for *Realism Evaluation* focuses on assessing contextual and narrative coherence in model-generated speech. Unlike the archetype-oriented evaluation, which emphasizes stylistic and emotional fidelity, this platform examines how well the voice performance aligns with a character's personality, dialogue situation, and scene progression.

Figure 5 shows the core playback interface. Annotators can preview each audio sample with synchronized spectrogram and waveform displays, while reviewing contextual metadata including the *Local Scene*, *Character Style*, and *Character Profile*. These descriptions help situate the listening task, providing the background necessary for judging realism in the character's vocal and behavioral expression.

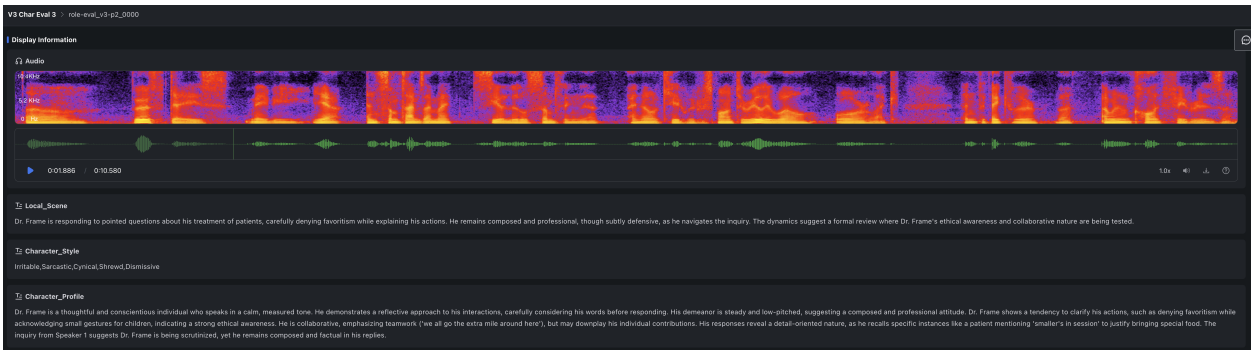


Figure 5: System interface for Realism Evaluation. Annotators can visualize spectrograms and waveforms while reviewing contextual metadata such as the scene description, character style, and profile to guide their evaluation.

Once familiar with the scene, annotators proceed to the detailed evaluation form shown in Figure 6. This interface is divided into two major sections: prosodic/emotional quality, higher-level coherence and semantic alignment.

F Annotation Guidelines for Realism Evaluation

This appendix contains the full human annotation instructions used for realism evaluation.

F.1 Initial Assessment & Rejection Instructions

- Read the **Character Profile** (long-term identity, style) and **Local Scene** (immediate context, goal/emotion) before listening.
- Focus evaluation on the *spoken performance of the target speaker only*; ignore background music, ambient noise, and non-speech events.
- If the audio's speaker is *clearly not* the intended target identity, assign a score of **1** to all Profile-dependent dimensions (e.g., Voice Identity Matching, Trait Embodiment) and proceed with other dimensions as applicable.
- If the prompt information is **contradictory** or **insufficient** to evaluate (e.g., audio too short, traits mutually incompatible), mark the prompt quality accordingly and *do not* force granular ratings that cannot be judged.

Detailed case examples and progressive gating for emotional dimensions are provided in the full instruction document.

F.2 Rubric Rating Instructions

We adopt five dimensions, each rated from 1 (worst) to 5 (best). Emotional Expressiveness uses *progressive gating*.

- **Prosodic Dynamics** (Pitch Variation, Rhythmic Naturalness, Stress & Emphasis): naturalness of patterns and rhythms.
- **Emotional Expressiveness** (Progressive gating):
  1. Emotion Accuracy (required),
  2. Emotion Intensity Control (rate only if $2.1 \geq 3$),

3. Dynamic Range (rate only if $2.2 \geq 3$).

- **Character Consistency** (Voice Identity Matching, Trait Embodiment): alignment with Character Profile.
- **Contextual Relevance** (Local Scene Fit, Global Story Coherence): fit to immediate scene and long-term identity.
- **Semantic Match** (spoken content vs. Local Scene goal/emotion/action).

The complete rubrics with per-score definitions and examples are reproduced below.

F.3 Commentary Guidelines and Confidence Ratings

Annotators should provide short, clear justifications:

- Use the comment box to note contradictions, missing information, vagueness, or multi-speaker issues.
- Keep comments brief and non-technical; do not quote transcripts unless necessary.
- Do not reward theatricality/technical polish unless it naturally fits the character *and* scene.

Annotators also provide a **confidence** level (1–5) reflecting familiarity with the role/scene and clarity of audible cues. Low confidence *does not* reflect poorly on annotators; it adds nuance for downstream analysis.

F.4 Full Annotation Guideline (Realism Evaluation)

Realism Evaluation

`Goal. Rate how well the given speech (audio and transcript) aligns with the character's personality, traits, speaking style, and intent, based on:`

- `Character Profile: the character's long-term personality and speaking habits.`
- `Local Scene: the current scene's context and the character's immediate goal or emotion.`

`Each subcategory is scored on a 1-5 scale. Always read the Character Profile and Local Scene before listening; use both the audio and transcript (reference only) to guide your score.`

Initial Assessment & Pre-Checks. Before rating any speech sample, follow this checklist to ensure consistency and fairness.

Caveats

- **Read first:** Carefully read the *Character Profile* and *Local Scene* to understand the long-term identity and immediate goal/emotion.
- **Use both sources:** Play the audio and skim the transcript together to assess delivery quality, expression, and alignment with character intent (do not penalize transcript mismatches).
- **Focus only on human speech:** Do not factor in background music/soundtracks, environmental or ambient noise, or non-speech audio events; these may reflect the scene but should not influence scoring of the spoken performance.

- **Target speaker only:** If multiple speakers are present, evaluate only the target character's lines (consider others only as context for turn-taking).
- **Identity mismatch:** If the clip's speaker is clearly not the intended target speaker, assign a score of **1** for every rating dimension that relies on the Character Profile (e.g., Character Style Traits, Trait Embodiment, Character Consistency).

Prompt Quality (Pre-Check)

Rate the prompt before annotation:

1. **Prompt contains contradictory information.** Conflicts make accurate annotation difficult or impossible. Examples: incompatible accent specifications; Local Scene drastically at odds with intended emotion (see *Semantic Match*); Local and global scenes inconsistent; traits described as "evolving over time" without a timeline.
2. **Prompt is insufficient to evaluate the voice sample.** Audio too short to gauge the dimensions; or the prompt lacks specific characteristics/is too vague or multi-directional to know which version of the character is present. (It is OK if audio does not match transcript exactly; if too much is missing to annotate, mark as insufficient.)
3. **Prompt is sufficient to evaluate the voice sample.** No material issues; proceed.

Overall Speaker Focus and Character Style Traits

- **Checkbox 1: Multi-Speaker Influence.** Check *Yes* if other speakers were present and made it harder to isolate/judge the target character (check *Yes* if uncertain). Otherwise check *No*.
- **Question 2: Observed Character Style Traits.** Based on the Character Profile, are the defined Style Traits clearly present in the delivery?
  - Keep trait tags if a trait is expressed (positively or clearly in the opposite direction); this supports high/low embodiment scoring.
  - Delete incorrect pre-populated traits; you do not need to replace them. If you delete many traits, *Trait Embodiment* should likely be low.
  - If no traits are expressed, enter "N/A". If unsure, leave original traits unedited.
- *Example (customizable):* `"Character_style":"Playful, Charismatic, Confident, Shrewd"`.

Rubric Rating Instructions. Score each subcriterion on a 1–5 scale. Use the detailed rubrics below and the examples to maintain calibration across characters, tones, and emotional styles. Prioritize in-character believability; do not reward theatricality, emotional intensity, or technical polish unless they naturally fit the character *and* scene.

Guideline 1. Prosodic Dynamics (Patterns and Rhythms)

Focus: pitch, rhythm, and emphasis in speech. **Guideline 1.1 Pitch Variation (check Local Scene, Character Profile, and audio).**

Score	Description / Example
5	Pitch moves naturally/believably, subtly capturing emotional state & speech pattern (e.g., gentle rise when a calm-but-curious character asks a question).
4	Slight variation that fits character; somewhat restrained/inconsistent.
3	Neutral pitch/non-distracting but lacks expressive intent.
2	Forced/theatrical or mismatched pitch shifts.
1	Monotone or misaligned with scene/character.

Guideline 1.2 Rhythmic Naturalness (check Local Scene, Character Profile, and audio).

Score	Description / Example
5	Smooth, believable rhythm reflecting speech habit & scene pacing (e.g., nervous character in fast but coherent bursts).
4	Minor deviations but still natural.
3	Even pace without clear rhythmic personality.
2	Disjointed, overly rigid, or awkward pacing (odd pause placements).
1	Completely unnatural/robotic timing that interrupts flow.

Guideline 1.3 Stress and Emphasis (check Local Scene, Character Profile, and audio).

Score	Description / Example
5	Emphasis highlights key meanings/emotions in realistic ways (e.g., stress on "*can't*" in "We can't leave them behind!").
4	Generally helpful emphasis with minor misalignment.
3	Mild/generic stress patterns (not distracting, limited impact).
2	Over-emphasis or misplaced focus (e.g., stressing function words).
1	Flat or confusing emphasis (all equal or no variation).

Guideline 2. Emotional Expressiveness (Progressive Gating)

Focus: alignment of expressed emotion with character style and situational demand.
Gating: Only rate 2.2 if 2.1 $\geq$ 3; only rate 2.3 if 2.2 $\geq$ 3.

Guideline 2.1 Emotion Accuracy (Required; check Local Scene, Character Profile, and audio).

Score	Description / Example
5	Emotion matches personality and situation with subtlety (e.g., hesitant pause before apologizing for a reserved character).
4	Emotionally grounded with light misalignment.
3	Too neutral/safe for the scene.
2	Out of sync with character or situation (e.g., cheerful during betrayal).
1	Jarring/inappropriate (e.g., laughing/yelling without motivation).

If you select 1 or 2, stop here: set 2.2 and 2.3 = N/A.

Guideline 2.2 Emotion Intensity Control (Only if 2.1 $\geq$ 3; check Local Scene, Character Profile, and audio).

Score	Description / Example
5	Intensity calibrated perfectly (e.g., whispered anger in a tense standoff).
4	Mostly appropriate (slight over-/under-shoot).
3	Slightly exaggerated or dull.
2	Overacted or underplayed (e.g., screaming in a reflective moment).
1	Wildly off-scale (e.g., hysterical laughter in a death scene).

If you select 1 or 2, stop here: set 2.3 = N/A.

Guideline 2.3 Dynamic Range (Only if 2.2 $\geq$ 3; check Local Scene, Character Profile, and audio).

Score	Description / Example
5	Emotional shifts, if present, are natural and in-character (e.g., mentor calmly grows firmer while warning a student).
4	Appropriate shifts present but mild/slow to emerge.
3	Mostly steady; slight modulation; no strong arc.
2	Flat/monotone or unnaturally varied with unjustified jumps.
1	Significant out-of-character fluctuations without triggers.

Guideline 3. Character Consistency

Focus: how well the voice and language align with the character's core identity.

Guideline 3.1 Voice Identity Matching (Character Profile, and audio).

Score	Description / Example
5	Voice clearly matches the expected identity (e.g., youthful, curious tone for energetic sidekick).
4	Acceptable match with minor mismatches.
3	Neutral/not character-specific.
2	Somewhat inconsistent (e.g., deep voice for high-pitched character).
1	Clearly mismatched identity (e.g., gritty adult voice for bubbly child).

Guideline 3.2 Trait Embodiment (Character Profile, and audio).

Score	Description / Example
5	Defining traits subtly but clearly conveyed (e.g., insecure character hesitates/lowers voice).
4	Most traits conveyed; minor inconsistencies.
3	Traits weak but not disruptive.
2	Missing or conflicting traits (e.g., arrogant tone for a humble character).
1	No identifiable traits or completely off.

Guideline 4. Contextual Relevance

Focus: appropriateness in the current scene and long-term coherence.

Guideline 4.1 Local Scene Fit (check Local Scene, and audio).

Score	Description / Example
5	Matches the scene's emotional/narrative context (e.g., urgent whisper during stealth).
4	Mostly coherent with minor mismatch (e.g., calm tone while others panic).
3	Neutral: neither reinforces nor harms the scene.
2	Confusing or slightly inappropriate tone (e.g., laughing during a tense standoff).
1	Directly contradicts scene intent (e.g., casual joke while a character is dying).

Guideline 4.2 Global Story (Character Profile) Coherence (check Character Profile, and audio).

Score	Description / Example
5	Perfectly consistent with background and arc (e.g., wise mentor advice aligned with past scenes).
4	Slightly off-style but plausible (e.g., cheerful tone for usually sarcastic character).
3	Generic/filler delivery (*could be anyone*).
2	Contradicts growth or habits (e.g., arrogance after humility arc).
1	Breaks established identity (e.g., cold anger from consistently empathetic healer).

Guideline 5. Semantic Match

Focus: how well the spoken content semantically aligns with the Local Scene's stated goal, emotion, or action.

5.1 Semantic Matchness (Audio, Local Scene).

Score	Description / Example
5	Content perfectly advances the scene goal/emotion (e.g., rescue: "Secure the exit; I'll get the hostages.").
4	Minor nuance mismatch but still scene-appropriate.
3	Neutral filler that neither helps nor contradicts (plausible bridge).
2	Partially conflicts or confuses context; requires *stretch* explanation.
1	Directly contradicts/derails scene intent; no reasonable bridge.

Confidence

Provide a confidence rating (1–5) reflecting your certainty that you:

- Understood the character and scene clearly,
- Heard sufficient emotional cues in the audio, and
- Interpreted speech alignment with intended personality and context.

Guidance: Rate low confidence for noisy/ambiguous audio, vague/contradictory traits, multiple dominant speakers, or hard-to-distinguish emotion/tone.

Comments (Optional)

Use the comment box to note contradictions, missing information, vague descriptions, or multi-speaker issues that affected your evaluation. Keep comments brief, clear, and non-technical; no transcript quotes are required.

G Annotation Ethics

All human annotation in this work was conducted in accordance with standard ethical practices for research involving human annotators. Annotators were recruited via public hiring for Mandarin and internal lab hiring for English. Prior to participation, all annotators were provided with a clear description of the task, the type of data they would encounter, and their rights as participants, including the ability to opt out at any time without penalty.

Annotators gave informed consent before starting the tasks. They were compensated at a rate from $20 to $30 per hour, which corresponds to at least the local minimum wage and was applied uniformly regardless of whether their contributions were later filtered for quality.

The data presented to annotators consisted of synthetic speech samples, anonymized human speech, or recordings from public corpora, and contained no personally identifiable or sensitive information. To mitigate potential risks, annotators were allowed to skip any item that they found uncomfortable or inappropriate.

To ensure annotation reliability, annotators were provided with written guidelines, examples, and trial tasks before beginning. Quality control procedures included inter-annotator agreement checks, and comprehensive screening tests. Each speech sample is evaluated by at least 3 annotators. This study did not involve vulnerable populations and posed minimal risk to participants.

H Detailed Data Analysis (Archetype Evaluation)

The archetype evaluation provides a detailed view of how speech foundation models (`SFnd`) perform when tasked with role-play scenarios grounded in stereotypical characters and situations. Human annotations cover four dimensions, *Appropriateness*, *Human Likeness*, *Audio Quality*, and *Content Pass*, and the results reveal important trends, as well as challenges in interpreting scores across languages.

Annotation Consistency and Cross-Language Comparisons. A critical observation is that annotations in Mandarin and English cannot be directly compared. The annotation groups for the two languages differed in their linguistic backgrounds and cultural interpretations, which leads to systematic differences in scoring. Our bilingual pilot further confirmed that annotators often interpret role-play appropriateness and expressiveness differently depending on their cultural frame of reference. For example, Mandarin raters tend to reward delivery styles that align with culturally expected prosody and politeness strategies, while English raters prioritize spontaneity and naturalness. These discrepancies parallel the findings of the singing voice conversion challenge (SVCC2023) from Huang et al. (2023), where native and non-native listeners provided significantly different ratings for the same samples, particularly for naturalness and speaker similarity. Consequently, while absolute scores differ substantially across Mandarin and English, only within-language comparisons are meaningful. Cross-lingual ranking is not valid given the different annotation bases.

Mandarin Results. Within Mandarin from Table 13, the relative rankings of models are clear. `Doubao` stands out as the strongest performer across all four dimensions, reaching near-ceiling levels in both *Audio Quality* (4.37) and *Content Pass* (0.98). It also leads in role-related dimensions such as *Appropriateness* and *Human Likeness*, indicating balanced competence across linguistic accuracy, delivery, and expressiveness. `G-CosyVoice2` and `Q-CosyVoice2` follow closely, especially in audio quality, though they lag behind in role appropriateness. `KimiAudio` and `GLM4-Voice` achieve mid-range results, reflecting higher variance across samples, while `GPT-4o-realtime` shows competitive content understanding but lower naturalness and appropriateness. `Qwen2.5Omni` consistently ranks the lowest, with large variance, underscoring instability and weaker alignment with archetypal role expectations.

English Results. The picture shifts in English, elaborated in Table 13. Here, `ChatGPT-4o` emerges as the most effective model by a wide margin, achieving top scores in both *Appropriateness* (3.46) and *Human Likeness* (3.49), as well as high content pass rates (0.97). This reflects its strong grounding in English discourse norms and role adherence. In contrast, `CosyVoice2` variants perform well in *Audio Quality* but are weaker in role alignment, highlighting a trade-off between acoustic fidelity and role-play embodiment. Other models such as `Doubao`, `GLM4-Voice`, and `GPT-4o-realtime` achieve moderate results, while `KimiAudio` and `Qwen2.5Omni` trail far behind, showing limited robustness in English scenarios. Taken together, the English results reveal both a sharper separation across models and a wider variance in appropriateness scores compared to Mandarin.

Domain-Level Trends. Table 14 and Figures 7–8 further highlight differences across domains. In Mandarin, performance is relatively consistent across categories, with means clustering around 3.1–3.3. The highest scores appear in `SocialIdentity` and `DailyOccupation`, suggesting that models handle culturally grounded archetypes more reliably. In English, however, domain variation is much greater. Categories such as `FantasyOccupation` and `NamedCharacters` receive notably lower scores (means below 2.0), reflecting the difficulty of producing expressive, role-appropriate speech in imaginative or culturally specific contexts.

Table 13: Overall human annotation results on Mandarin and English archetype evaluation dimensions (mean ± std). Best per column in **bold**.

Model (Mandarin)	Appropriateness	Human Likeness	Audio Quality	Content Pass
Doubao	**3.8236 ± 0.6658**	**3.6572 ± 0.6427**	**4.3743 ± 0.6991**	**0.9848 ± 0.1224**
G-CosyVoice2	3.4243 ± 0.7389	3.3156 ± 0.7215	4.0406 ± 0.7298	0.9725 ± 0.1637
Q-CosyVoice2	3.4228 ± 0.7788	3.2750 ± 0.7335	4.0236 ± 0.7661	0.9674 ± 0.1776
KimiAudio	3.3866 ± 1.1086	3.3391 ± 1.0591	3.9525 ± 1.2801	0.8623 ± 0.3446
GLM4-Voice	3.2022 ± 0.9597	3.0275 ± 0.8687	3.9246 ± 1.1105	0.8949 ± 0.3067
GPT-4o-realtime	3.0228 ± 0.8581	2.6536 ± 0.7151	3.7297 ± 1.0418	0.9098 ± 0.2865
Qwen2.5Omni	2.4942 ± 1.4242	2.3830 ± 1.3201	2.8685 ± 1.7456	0.5482 ± 0.4978
Model (English)				
ChatGPT-4o	**3.4638 ± 0.9671**	**3.4873 ± 0.7857**	1.9620 ± 0.6885	**0.9662 ± 0.1808**
Doubao	2.3798 ± 1.0573	2.2355 ± 0.8918	2.9553 ± 1.1589	0.7995 ± 0.4005
GPT-4o-realtime	2.2736 ± 0.9606	2.0906 ± 0.8124	2.5127 ± 0.9702	0.8521 ± 0.3552
GLM4-Voice	2.2355 ± 1.0354	2.0779 ± 0.9095	2.5694 ± 1.1366	0.7482 ± 0.4342
G-CosyVoice2	2.2168 ± 0.9916	2.0109 ± 0.8585	**3.2530 ± 0.9637**	0.9293 ± 0.2563
Q-CosyVoice2	2.1673 ± 0.9642	1.8068 ± 0.8011	3.2476 ± 0.9113	0.9614 ± 0.1928
Qwen2.5Omni	2.0079 ± 1.0420	1.9408 ± 0.9489	2.4161 ± 1.2583	0.6159 ± 0.4865
KimiAudio	1.5531 ± 0.8716	1.7579 ± 0.9386	2.0097 ± 1.1788	0.4861 ± 0.5000

Table 14: Domain-level human annotation results for Appropriateness dimension of Mandarin and English archetype evaluation (mean ± std). We report results of two languages with inter-model variance separately.

Domain	Mandarin		English	
	Score (mean ± std)	Inter-model Var.	Score (mean ± std)	Inter-model Var.
NamedCharacters	3.0343 ± 1.1087	1.1087	1.9896 ± 1.0409	1.0409
FantasyOccupation	3.1629 ± 0.9661	0.9661	1.9444 ± 0.9809	0.9809
Traits	3.1670 ± 1.2109	1.2109	2.1518 ± 1.0572	1.0572
Events	3.2535 ± 1.1523	1.1523	2.2608 ± 1.1280	1.1280
DailyOccupation	3.2760 ± 0.8652	0.8652	2.5129 ± 1.1781	1.1781
RelativeRoles	3.2762 ± 1.0255	1.0255	2.3594 ± 1.1298	1.1298
SocialIdentity	3.3263 ± 0.9802	0.9802	2.3256 ± 1.0912	1.0912

Occupational roles remain relatively easier for models in both languages, likely because these scenarios are closer to training data distributions.

Interpretation and Broader Implications. Although cross-lingual comparisons are invalid, this does not diminish the utility of archetype evaluation for role-play benchmarking. In fact, the observed inconsistencies reinforce a central point: successful role-play requires models to adapt to the cultural and linguistic norms of the target language community. A system that performs well in Mandarin but poorly in English cannot be considered universally capable, since end users will evaluate appropriateness and naturalness according to their native expectations. Thus, archetype evaluation highlights not only model-specific strengths and weaknesses but also the need for culturally grounded evaluation pipelines.

Another notable pattern is the trade-off between acoustic quality and role embodiment. CosyVoice2 variants achieve strong audio quality but weaker human likeness and appropriateness, whereas ChatGPT-4o prioritizes role coherence at the expense of audio fidelity. This suggests that future systems must better integrate high-quality synthesis with nuanced role adherence. Similarly, the consistently lower scores for imaginative and identity-driven roles underscore the limitations of current SFMs in capturing subtle sociocultural expectations and narrative context.

Summary. In sum, archetype evaluation reveals three major findings: (1) Mandarin and English annotations cannot be directly compared, due to linguistic and cultural annotation inconsistencies, consistent with prior evidence from SVCC 2023. (2) Within each language, clear model rankings emerge, `Doubao` dominates in Mandarin, while `ChatGPT-4o` excels in English. (3) Domain sensitivity remains a challenge, particularly for imaginative or culturally rich roles, where scores are substantially lower and inter-model variance is higher.

I Detailed Data Analysis (Realism Evaluation)

Overview. Although realism evaluation targets nuanced paralinguistic behaviors, human annotators achieve *consistently high agreement* once the rubric disentangles the aspects to be judged. As shown in the agreement summary in Fig. 9, most dimensions fall in the 0.86–0.88 range, with the lowest agreement on *Traits Embodiment* (≈ 0.83) and the highest on *Local Scene Fit / Semantic Match* (≈ 0.88). Emotion-related, progressive metrics (Intensity and Transition) are slightly lower (≈ 0.85), reflecting the added difficulty of judging graded expressivity.

Agreement definition. Let R denote the possible score range (e.g., $R = 4$ for a 1–5 scale). For a sample i with annotator scores $\{s_{ij}\}_j$, we compute the sample standard deviation σ_i and define the per-sample agreement

$$a_i \;=\; \max\big(0,\; \min(1,\; 1 - \sigma_i/R)\big).$$

We then report the overall agreement $A = 1 - \bar{\sigma}/R$, where $\bar{\sigma}$ is the mean of $\{\sigma_i\}$ over samples with at least two annotations. We also clamp A to $[0, 1]$ and ignore single-annotator cases. A minimal, reproducible implementation is:

Caveats and planned upgrades. The metric $A = 1 - \bar{\sigma}/R$ above is a deliberately simple, scale-aware summary; it is *not* Krippendorff's α, Fleiss' κ, or any standard ordinal inter-rater agreement (IAA). Because A collapses any constant-offset disagreement into the same number as a tight cluster on a single bin, it should be read as a coarse *lower bound* on rater consistency rather than as a chance-corrected agreement. We acknowledge two specific consequences here so that downstream tables in Sec. 6 can be interpreted properly: (i) the gating choice for *Emotion Intensity/Emotion Transition* (where `N/A` is mapped to 1) artificially inflates A on those two dimensions because annotators who all hit the floor look perfectly consistent under σ-based agreement; (ii) without a per-dimension human *upper bound* (annotator–annotator Pearson under the same protocol used for model–human comparisons), the model–human Pearson values reported in Tabs. 7–15 cannot be compared to a ceiling. Krippendorff's α with ordinal weights per dimension, together with annotator–annotator Pearson per dimension as a human upper bound on the model–human correlations, would be the appropriate standard complements to A; we flag their absence as a limitation of the current agreement reporting. The A values are provided as a simple reading of the rubric's internal consistency.

Listing 1: Agreement computation (concise).

```
def compute_agreement(values_by_sample, max_possible_range):
    stds, per_sample = [], {}
    for i, scores in enumerate(values_by_sample):
        x = [float(s) for s in scores if s is not None]
        if len(x) < 2: continue
        x = np.asarray(x, dtype=float)
        s = x.std(ddof=1)
        a = 1 - (s / max_possible_range) if max_possible_range > 0 else 1
        a = max(0, min(1, a))
        per_sample[f"sample_{i}"] = {
            "scores": x.tolist(), "mean": round(x.mean(), 3),
            "std": round(s, 3), "range": round(x.max()-x.min(), 3),
            "agreement_score": round(a, 3), "annotator_count": len(x)
        }
        stds.append(s)
```

```
mean_std = np.mean(stds) if stds else np.nan
A = 1 - (mean_std / max_possible_range) if stds else np.nan
if not np.isnan(A): A = max(0, min(1, A))
return {"mean_intra_sample_std": mean_std,
        "sample_agreements": per_sample,
        "overall_agreement": A}
```

Distributional characteristics. Fig. 10–11 visualize the score distributions per dimension. We highlight three patterns: (i) **Progressive/gated emotion metrics.** For *Emotion Intensity* and *Emotion Transition*, we assign `1` when the scene is marked *N/A* (i.e., emotion accuracy is insufficient to meaningfully judge intensity/transition). This produces a visible mass at 1, a sparse bin at 2, and slightly lower overall agreement; downstream analyses conditioning on adequate *Emotion Accuracy* (e.g., mean ≥ 3) recover more bell-shaped distributions. (ii) **Traits Embodiment.** Scores tend to avoid the middle (~ 3), concentrating on low values when traits clearly mismatch and high values when traits are strongly realized. This yields the lowest inter-rater agreement among dimensions, reflecting the inherently compositional and role-specific nature of trait judgments. (iii) **Other dimensions.** *Pitch Variation*, *Rhythmic Naturalness*, *Stress & Emphasis*, *Voice Identity Matching*, *Local Scene Fit*, *Global Story Coherence*, and *Semantic Match* are approximately unimodal with modes at 3–4, consistent with moderate–strong realization in the dataset and with the high agreement reported.

Professional vs. Amateur voice actors. The paired histograms in Figs. 12–22 compare professional (left) and amateur (right) recordings:

- **Prosody control (Pitch, Rhythm, Stress).** Professional speech is shifted toward higher bins with tighter spread, indicating steadier control and fewer low-end tails.
- **Emotion rendering (Accuracy, Intensity, Transition).** Professionals show higher centers and reduced mass at 1 (fewer gated `N/A`s), suggesting more consistent emotional grounding and smoother dynamics; amateurs exhibit wider variance and more mass at the extremes.
- **Identity and traits (Voice Identity, Traits Embodiment).** Professionals present clearer peaks in upper bins, while amateurs are broader or bimodal, reflecting inconsistent character anchoring.
- **Scene and semantics (Local Scene Fit, Global Coherence, Semantic Match).** Differences are present but smaller; both groups cluster in the mid–high range, with professionals exhibiting slightly narrower distributions.

Overall, professionals outperform amateurs on expressivity and control, while narrative/semantic alignment shows a milder gap.

Implications for modeling. First, the consistently high agreement confirms that the rubric isolates perceptually coherent factors, so learned evaluators can target them reliably. Second, the gated behavior of *Intensity/Transition* motivates a *hierarchical* evaluator: verify *Emotion Accuracy* before scoring graded expressivity, or model the joint structure with conditional heads. Third, *Traits Embodiment* remains challenging; models may benefit from explicit character-trait priors, long-term context aggregation, or contrastive supervision between matched and mismatched traits. Finally, prosody-focused objectives (e.g., pitch/rhythm/stress control) are likely to yield the largest quality gains for amateur-style speech.

J Evaluation Model Experiments

J.1 Zeroshot/Fewshot Setup for Evaluation Models

For both zeroshot and fewshot experiments, we adopt a version of the annotation guidelines to ensure consistency with the human annotation process. The exact rubric prompts for archetype and realism evaluation used are provided below:

Rubrics Prompts (Archetype Evaluation)

```
Follow the rubric EXACTLY and return ONLY a minified JSON object (valid for
json.loads) - no extra text.
Set content_pass to false (i.e., reject) if ANY apply:
- Clip ∼1-2s (too short to judge).
- Obvious breakdown/loop/restart.
- Any multilingual word/phrase (except extremely common one-offs like "croissant").
- Irrelevant to prompt's role/scene.
- Instruction failure:  breaks persona/scene, 4th-wall narration, multi-party
narration, long placeholder scaffolding.
Otherwise:  PASS.


WHAT TO DO
1) Apply the rubrics below strictly.  If torn between two scores, choose the LOWER
one.
2) Return ONLY a minified JSON object (valid for json.loads).  Integers 1-5.  No
extra text.
3) If not rejected, judge performance (HOW it's said), not wording.
4) If rejected, set all scores to 1.


0) Content-Pass
If rejected, set to false, otherwise set to true.


1) AudioQuality (Context-Free)
Focus ONLY on digital/synthetic artifacts (noise, hiss, metallic timbre, clipping,
glitchy transitions, pitch collapse, static).
After you record this score, PRETEND these issues do not exist for later
dimensions.
Scale:
- 5 (Excellent):  Clean/clear; no noticeable artifacts.
- 4 (Good):  Minor imperfections (slight hiss/sharpness) under close listen.
- 3 (Fair):  Noticeable artifacts (buzz/echo/synthetic texture) that impact
clarity.
- 2 (Poor):  Frequent glitches/clipping/metallic resonance/harsh transitions.
- 1 (Very Poor):  Dominated by artifacts (static/crackling/synthesis failure/severe
clipping or pitch collapse).


2) HumanLikeness (Context-Free)
How human-like is the delivery?  Ignore AudioQuality issues (already captured) and
ignore contextual fit.
Scale:
- 5 (Definitely Human):  Indistinguishable; rich intonation, natural real-time
pacing, consistent identity; organic breaths/pauses.
- 4 (Most Likely Human):  Generally convincing; minor stiffness or overly even
timing/pauses.
- 3 (Could be Human or AI): Mixed traits; some mechanical pauses/inflection, other
parts human-like.
- 2 (Mostly Likely AI): Clearly artificial; choppy rhythm, flat/illogical
```

inflections, awkward or forced breaths.
- 1 (Definitely AI): Rigid read-aloud monotony; no natural rhythm/pauses; entirely machine-like.

3) Appropriateness (Context-Dependent)
Tonal fit to the prompt's role & scene (stereotype-based). If ANY portion before 30s is ill-fitting, choose the LOWER score.
Ignore AudioQuality and generic naturalness here-judge only contextual tone/demeanor/status/urgency/emotion.
Scale:
- 5 (Completely Appropriate): Fits both role AND scene perfectly; fully aligned with emotion, social dynamics, status, and events; immersive and consistent.
- 4 (Mostly Appropriate): Strong match with minor flaws (e.g., slight intensity/timing/subtlety issues) that do not break alignment.
- 3 (Adequately Appropriate): Broad fit to role/scene but lacks depth or shows mild inconsistencies in key elements.
- 2 (Slightly Appropriate): Only role OR scene fit is acceptable; superficial/partial alignment; missing/mismatched cues.
- 1 (Completely Inappropriate): Fails role AND scene; tone/status/urgency clearly mismatched.

OUTPUT SCHEMA (minified JSON)
{"content_pass":true/false,"audio_quality":1-5,
"human_likeness":1-5,"appropriateness":1-5}

Rubrics Prompts (Realism Evaluation)

WHAT TO DO
1. **Read first:** Global Profile ⇒ Local Scene.
2. **Play** the audio **while skimming** the transcript (transcript may be imperfect).
3. **Focus only on the target speaker.**
   - Ignore background SFX / music / other voices except where they influence the target character's timing or tone.
4. **Apply the rubric below EXACTLY** and return **only** a minified JSON object - no prose, line-breaks, or extra keys (valid for json.loads).
5. **Scale is 1 (poor) ... 5 (excellent).** Use integers only.

DIMENSIONS & RULES
For every dimension, 5 = best possible, 1 = worst. Use these anchor descriptions to choose the closest integer.

1) PitchDynamics
5: Natural, situation-appropriate melodic movement; subtle, believable inflections.
4: Generally varied and fitting, minor stiffness or flat spots.
3: Neutral or limited inflection; neither distracting nor expressive.
2: Forced or sudden shifts that feel "performed."
1: Monotone or blatantly mis-pitched for the scene.

```
2) RhythmicNaturalness
5:  Flowing, lifelike rhythm; natural pauses.
4:  Mostly smooth; slight hiccups.
3:  Even pacing; read-aloud style.
2:  Halting, mechanical, broken phrases.
1:  Robotic timing, destroys immersion.


3) StressEmphasis
5:  Perfect stress on key words; enhances meaning.
4:  Generally correct with few mistakes.
3:  Mostly flat; generic sentential stress.
2:  Over-emphasis or misplaced stress.
1:  Flat OR every word equally stressed.


4) EmotionAccuracy
5:  Emotion fits precisely; subtle, authentic.
4:  Appropriate but slightly under/overstated.
3:  Neutral, misses clear opportunity.
2:  Noticeably off (e.g., joking tone in crisis).
1:  Jarring or nonsensical emotion.


5) EmotionIntensity (only if 4 ≥ 3)
5:  Perfectly calibrated.
4:  Minor overshoot/undershoot.
3:  Mildly too strong/weak.
2:  Clearly over- or under-acted.
1:  Wildly inappropriate.


6) EmotionalDynamicRange (only if 5 ≥ 3)
5:  Nuanced, organic shifts.
4:  Variation present, somewhat constrained.
3:  Mostly steady; slight modulation.
2:  Flat OR abrupt jumps.
1:  Chaotic, inconsistent emotion.


7) VoiceIdentityMatching
5:  Perfectly matches profile.
4:  Acceptable, minor mismatch.
3:  Generic voice.
2:  Noticeable mismatch.
1:  Obviously wrong identity.


8) TraitEmbodiment
5:  Core traits audible, believable.
4:  Most traits present, one weak.
```

```
3:  Traits faint, neutral delivery.
2:  Traits clash or unclear.
1:  No traits or fully off.


9) LocalSceneFit
5:  Perfect alignment with objective & mood.
4:  Minor mismatch.
3:  Neutral fit.
2:  Somewhat confusing.
1:  Contradicts scene.


10) GlobalStoryFit
5:  Consistent with character arc.
4:  Slightly atypical but plausible.
3:  Generic filler.
2:  Conflicts with backstory.
1:  Breaks identity entirely.


11) SemanticMatchness
5:  Perfectly advances scene.
4:  Minor mismatch but acceptable.
3:  Neutral filler.
2:  Partially conflicts with scene.
1:  Directly contradicts/derails.


Penalty Rule:  If the speaker is clearly the wrong person, set
voice_identity_matching = 1 and trait_embodiment = 1.


OUTPUT SCHEMA (minified JSON)
{ "pitch_dynamics":1-5, "rhythmic_naturalness":1-5, "stress_emphasis":1-5,
"emotion_accuracy":1-5, "emotion_intensity":1-5, "emotional_dynamic_range":1-5,
"voice_identity_matching":1-5, "trait_embodiment":1-5, "local_scene_fit":1-5,
"global_story_fit":1-5, "semantic_matchness":1-5 }
```

For zeroshot and fewshot settings, we condition the model with the rubric prompt and then append few-shot demonstration triplets (input, audio, gold JSON) to mirror the human annotation protocol. As shown in Listing 2, We use `Gemini2.5Pro` for these few-shot exemplars and require a minified JSON output consistent with the rubric schema.

Listing 2: Prompt construction for zeroshot and fewshot experiments (Gemini 2.5 Pro).

```python
import json
from typing import List, Dict, Literal

Mode = Literal["archetype", "realism"]

def _minify_json(d: Dict) -> str:
    return json.dumps(d, separators=(",", ":"))

def _format_input_desc(sample: Dict, mode: Mode) -> str:
```

```
    if mode == "archetype":
        q = sample.get("question", "")
        return f"Prompt's role and scene: {q}"
    gp = sample.get("char_profile", "")
    ls = sample.get("local_scene", "")
    cs = sample.get("char_style", "")
    return f"Global Profile: {gp}; Local Scene: {ls}; Traits: {cs}"

def _audio_part_from_wav(path: str, wav_to_base64) -> Dict:
    b64 = wav_to_base64(path)
    return {
        "type": "input_audio",
        "input_audio": {"data": "data:audio/wav; base64," + b64, "format": "wav"},
    }

def build_prompt_with_examples(
    rubric_prompt: str,
    few_shot_examples: List[Dict],
    current_item: Dict,
    mode: Mode,
    wav_to_base64,
    model_name: str = "gemini-2.5-pro",
) -> List[Dict]:
    """Constructs rubric + few-shot + current eval into a single prompt."""
    parts: List[Dict] = []

    # 1) Rubric
    parts.append({"type": "text", "text": rubric_prompt.strip() + "\n\n"})

    # 2) Few-shot demonstrations
    for i, ex in enumerate(few_shot_examples, start=1):
        parts.append({"type": "text", "text": f"**Few-shot Example {i}**\n\n"})
        parts.append({"type": "text", "text": f"**Example Input:** {_format_input_desc(ex, mode)}\n\n"})
        parts.append(_audio_part_from_wav(ex["wav_path"], wav_to_base64))
        if "annotation_prompt" in ex:
            gold_json = _minify_json(ex["annotation_prompt"])
            parts.append({"type": "text", "text": f"\n\n**Example Output:** {gold_json}\n\n"})

    # 3) Current evaluation
    parts.append({"type": "text", "text": "**Current Evaluation**\n\n"})
    parts.append({"type": "text", "text": f"**Current Input:** {_format_input_desc(current_item, mode)}\n\n"})
    parts.append(_audio_part_from_wav(current_item["wav_path"], wav_to_base64))
    parts.append({"type": "text",
                  "text": "\n\nPlease evaluate the current audio and return ONLY a minified JSON."})

    return [
        {"role": "system", "content": f"You are a helpful assistant using {model_name}."},
        {"role": "user", "content": parts},
    ]
```

Table 15: Realism role-play evaluator performance with `Gemini2.5Pro` in different fewshot conditions. Here, "1run" stands for only conduct generation for once.

Config.	Prosodic Dynamics			Emotional Expressiveness			Character Consistency		Contextual Relevance		Avg.
	Pitch	Rhythm	Stress	Accuracy	Intensity	Transition	Identity	Traits	Local	Global	
`Zeroshot-1run`	0.277	0.325	0.303	0.355	0.286	0.295	0.480	0.475	0.308	0.424	0.353
`Zeroshot`	0.319	0.352	0.336	0.391	0.316	0.336	0.529	0.509	0.349	0.465	0.390
`3-Shot-1run`	0.316	0.335	0.323	0.358	0.279	0.294	0.530	0.476	0.335	0.428	0.366
`3-Shot`	0.400	0.450	0.432	0.414	0.344	0.353	0.584	0.505	0.374	0.477	0.433
`6-Shot`	0.419	0.431	0.427	0.455	0.393	0.391	0.606	0.540	0.404	0.519	0.459
`9-Shot`	**0.419**	**0.453**	**0.434**	**0.464**	**0.404**	**0.416**	**0.613**	**0.551**	**0.407**	**0.524**	**0.469**

J.2 Zeroshot/Fewshot Ablation Experiments

Table 15 shows how different prompting strategies affect realism evaluation performance with `Gemini2.5Pro`. Several observations can be drawn:

- Effect of repeated runs. When comparing single-run decoding (`Zeroshot-1run`) against averaging over five runs (`Zeroshot`), we see consistent gains across nearly all dimensions (average score improves from 0.353 to 0.390). This suggests that sampling variance plays a substantial role in the reliability of large language model (LLM) evaluators, and averaging multiple outputs reduces this variance, producing more stable correlation with human annotations.
- Impact of few-shot conditioning. Introducing even three examples (`3-Shot`) yields a clear improvement over zeroshot (0.433 vs. 0.390). This indicates that providing explicit annotated demonstrations helps the model better align with the rubric. Interestingly, when using only a single run with few-shot demonstrations (`3-Shot-1run`), the performance is only marginally better than pure zeroshot, highlighting that both averaging and demonstrations are complementary: demonstrations improve grounding, while multiple runs mitigate stochasticity.
- Scaling the number of demonstrations. Increasing the few-shot set size from three to six and nine further boosts performance (0.459 and 0.469, respectively). The improvements are most pronounced in prosodic and emotional categories (e.g., *Pitch*, *Accuracy*, *Intensity*), suggesting that more examples provide stronger guidance in modeling subtle paralinguistic cues. Gains in *Character Consistency* and *Contextual Relevance* are also noticeable, although smaller in magnitude, which may indicate that these higher-level judgments are less sensitive to prompt scaling once a few demonstrations are provided.

In general, the results demonstrate two complementary strategies for stabilizing LLM-based evaluators: (i) averaging multiple runs to reduce randomness, and (ii) enriching prompts with more few-shot exemplars to improve adherence to rubric-based evaluation. Combining both leads to the strongest overall performance, with the nine-shot condition achieving the highest average correlation (0.469).

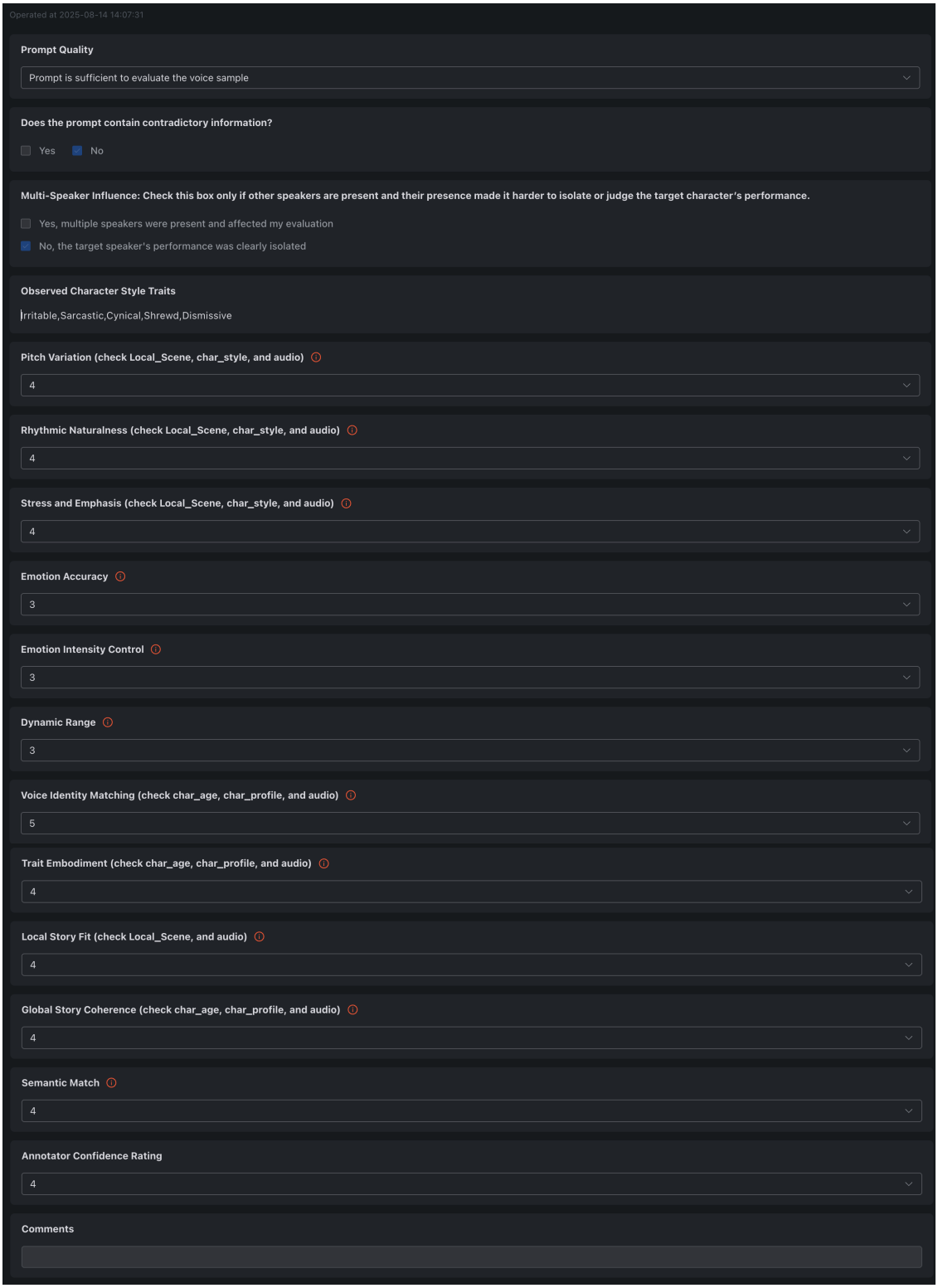
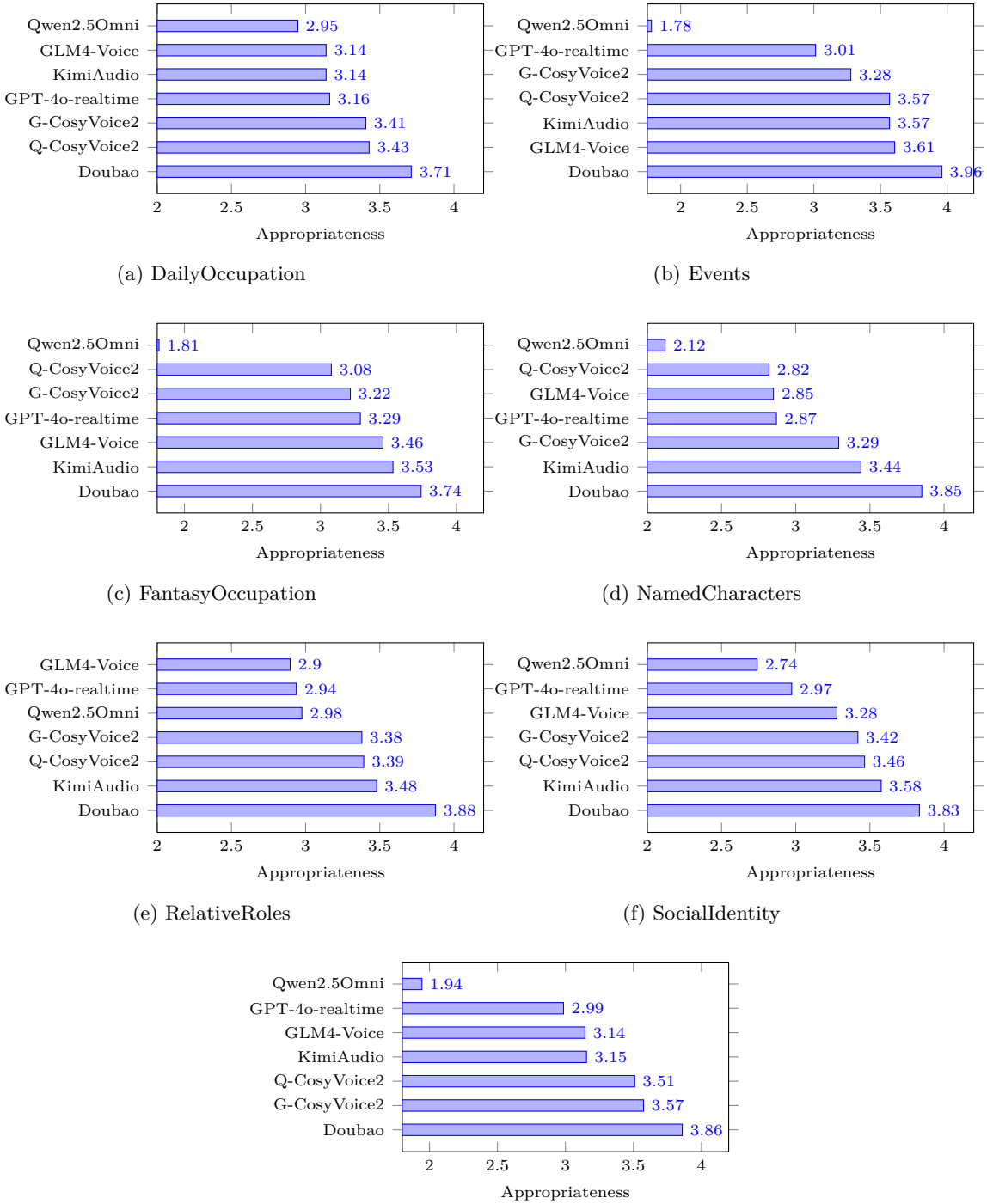


Figure 6: Annotation interface for Realism Evaluation. The form guides annotators through fine-grained perceptual judgments, connecting low-level speech delivery features with high-level narrative and character consistency.



(a) DailyOccupation (b) Events (c) FantasyOccupation (d) NamedCharacters (e) RelativeRoles (f) SocialIdentity (g) Traits

Figure 7: Model rankings within domains for Appropriateness dimension of **Mandarin** archetype evaluation (higher is better). Each bar shows the mean score for that domain.

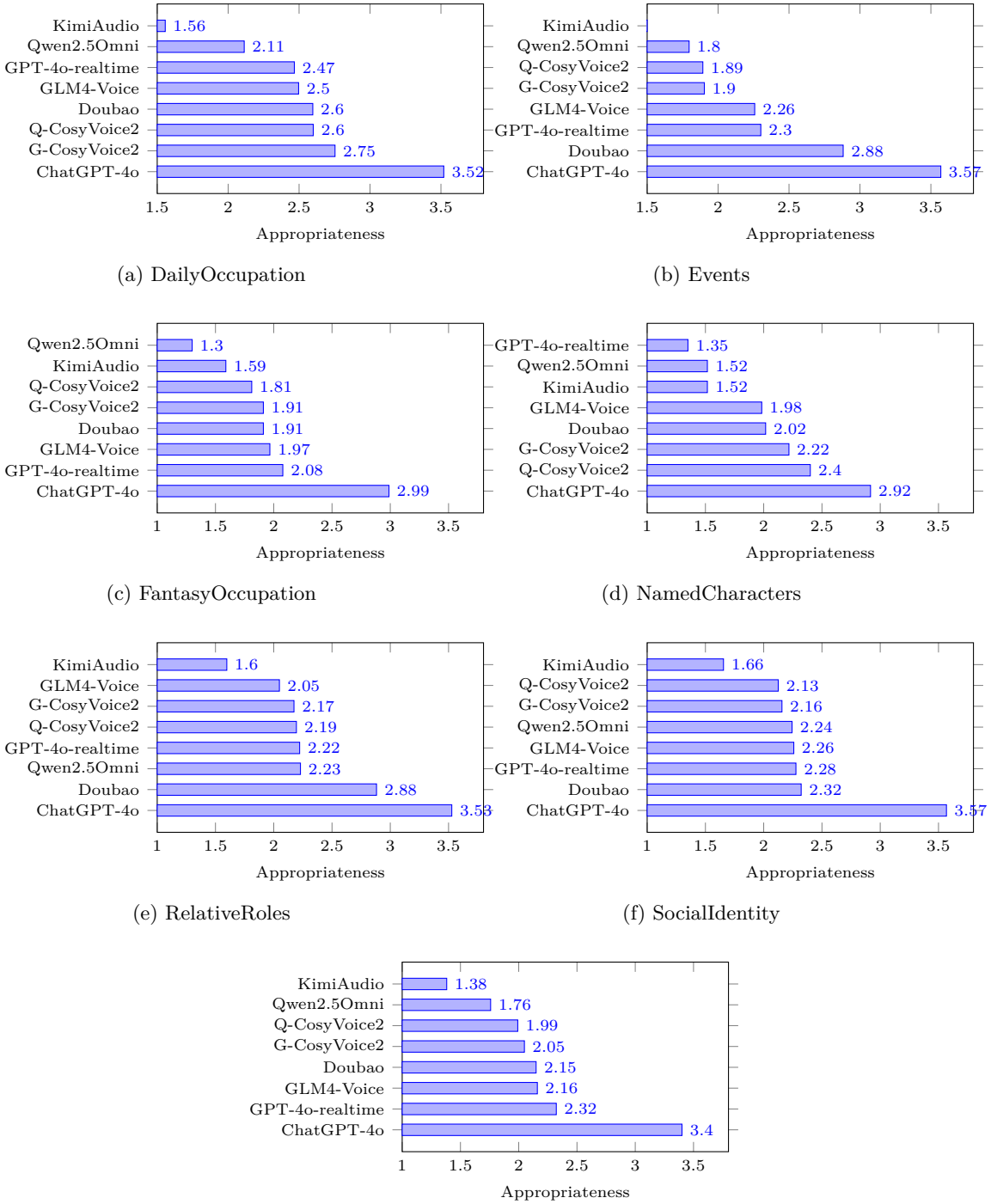


Figure 8: Model rankings within domains for Appropriateness dimension of **English** archetype evaluation (higher is better). Each bar shows the mean score for that domain.

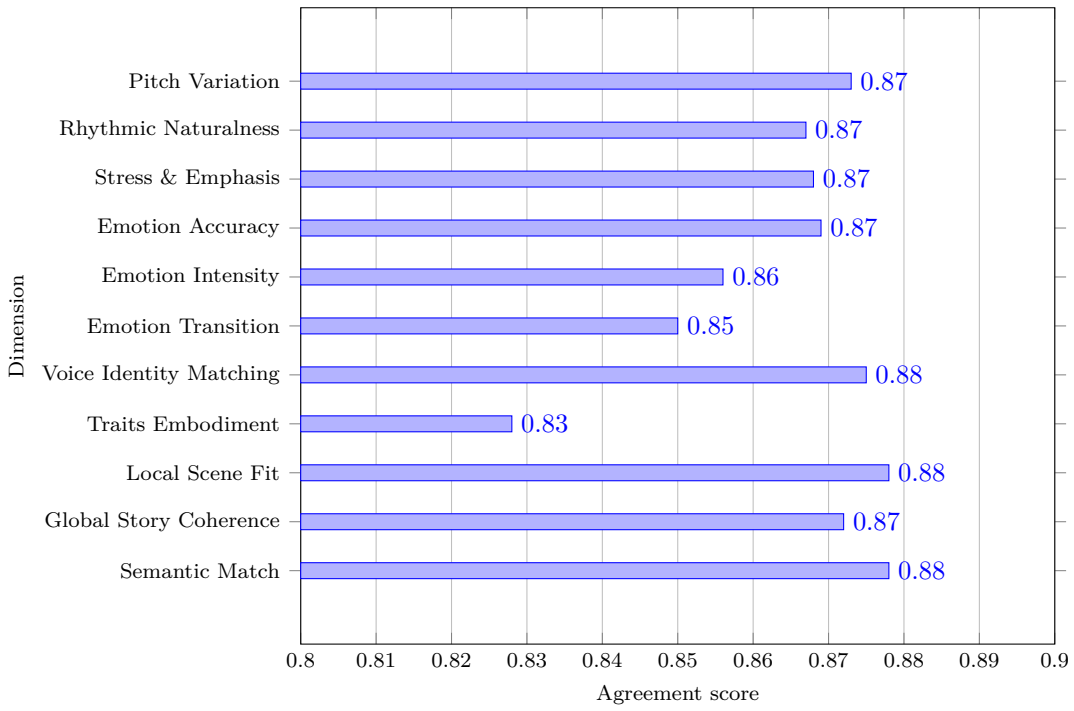
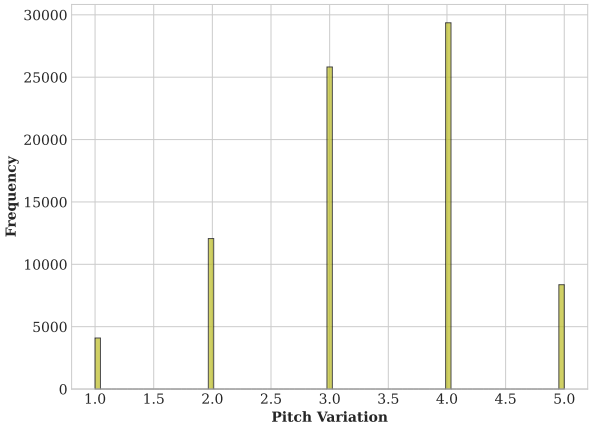
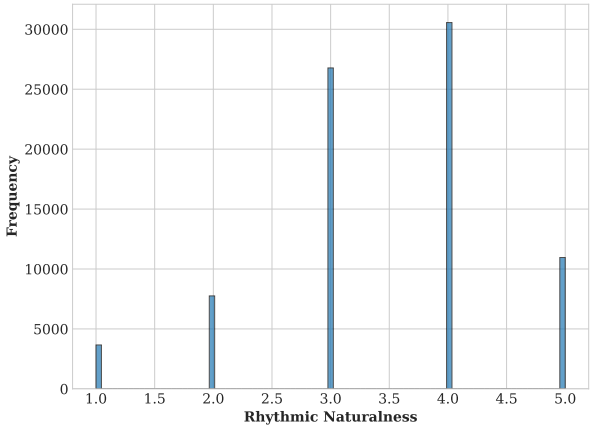


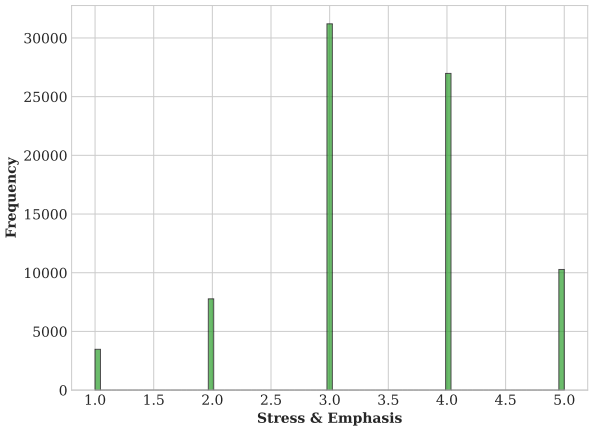
Figure 9: Agreement scores across evaluation dimensions. Higher values indicate stronger annotator consistency.



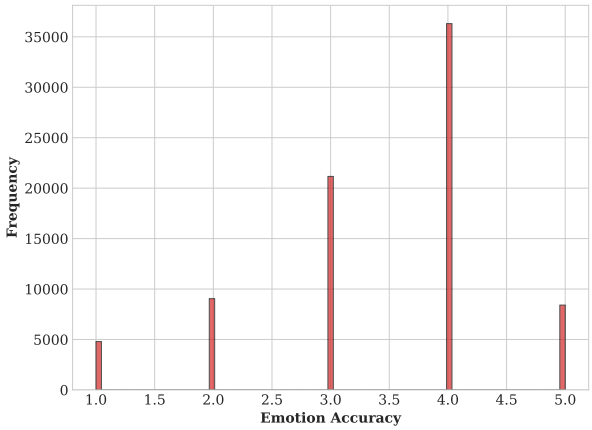
(a) Distribution of Pitch Variation



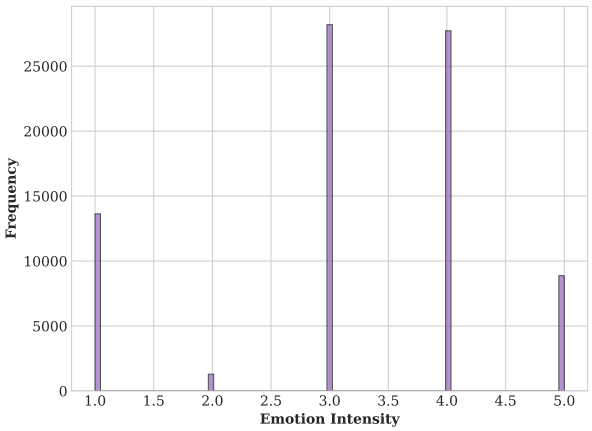
(b) Distribution of Rhythmic Naturalness



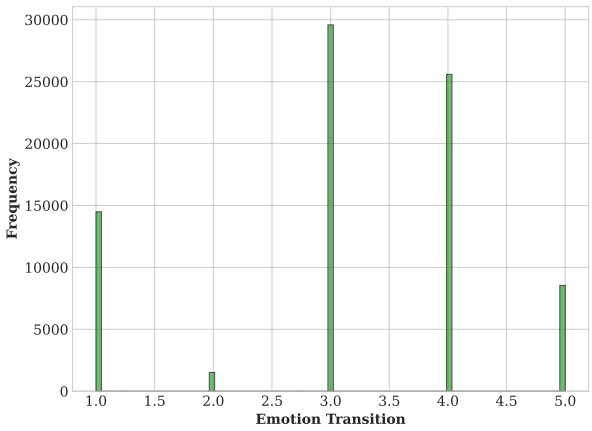
(c) Distribution of Stress & Emphasis



(d) Distribution of Emotion Accuracy

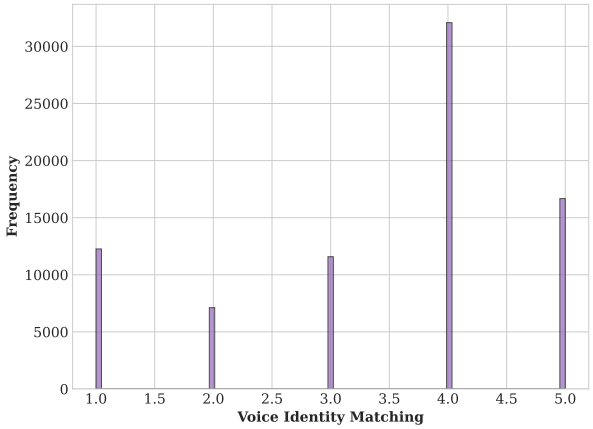


(e) Distribution of Emotion Intensity

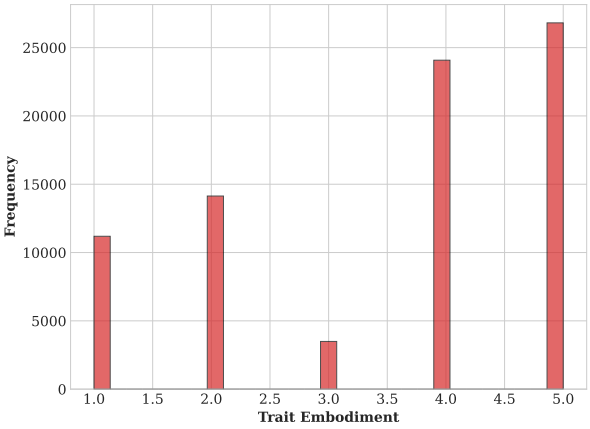


(f) Distribution of Emotion Transition

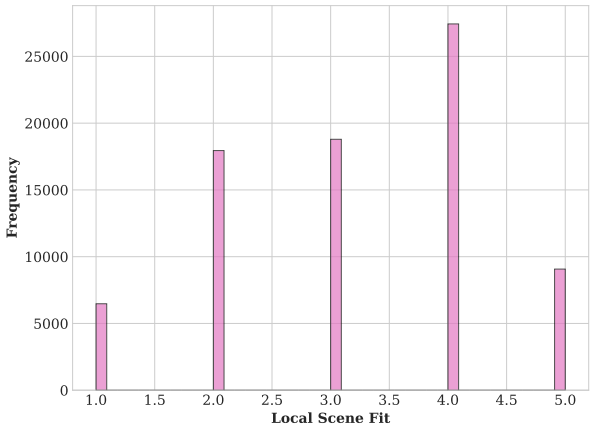
Figure 10: Distributions of evaluation dimensions in Realism Evaluation (1).



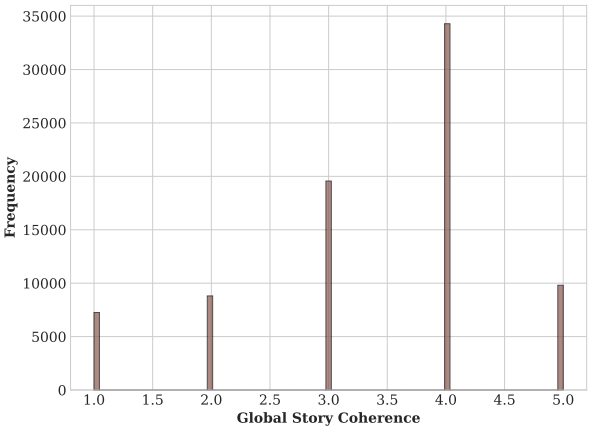
(a) Distribution of Voice Identity Matching



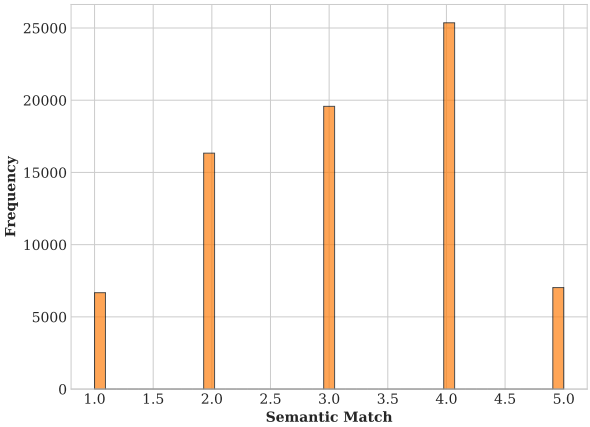
(b) Distribution of Trait Embodiment



(c) Distribution of Local Scene Fit

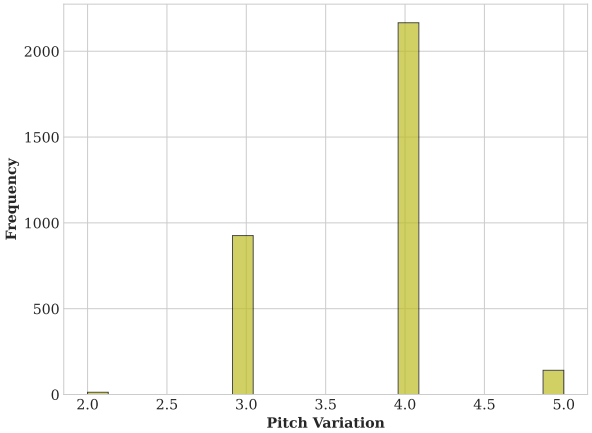


(d) Distribution of Global Story Coherence

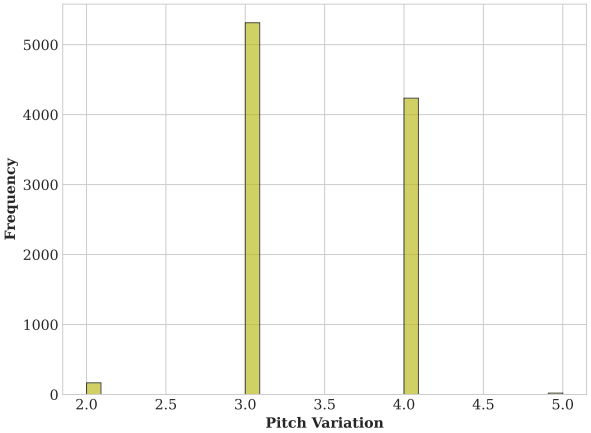


(e) Distribution of Semantic Match

Figure 11: Distributions of evaluation dimensions in Realism Evaluation (2).

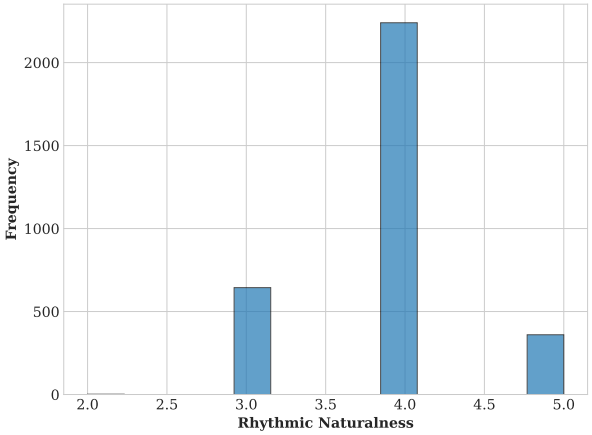


(a) Professional Voice Actor

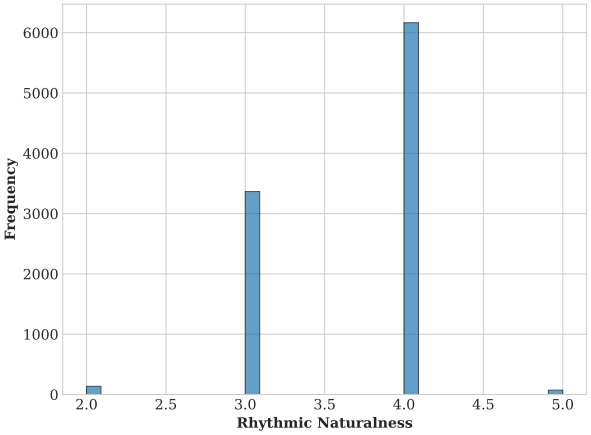


(b) Amateur Voice Actor

Figure 12: Distribution of Pitch Variation in Realism Evaluation: Professional vs. Amateur.

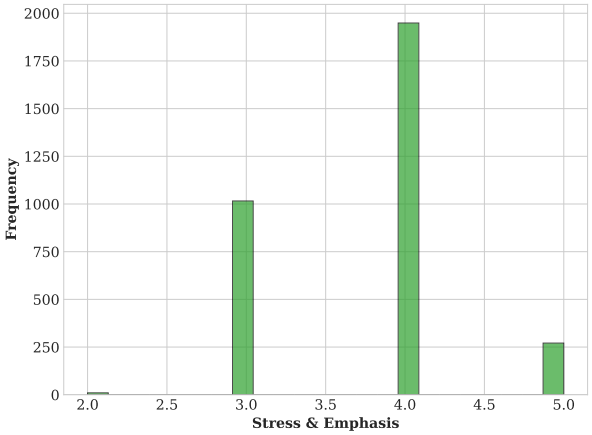


(a) Professional Voice Actor

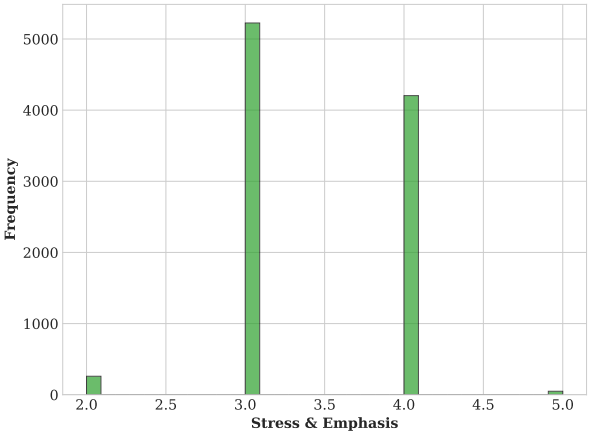


(b) Amateur Voice Actor

Figure 13: Distribution of Rhythmic Naturalness in Realism Evaluation: Professional vs. Amateur.

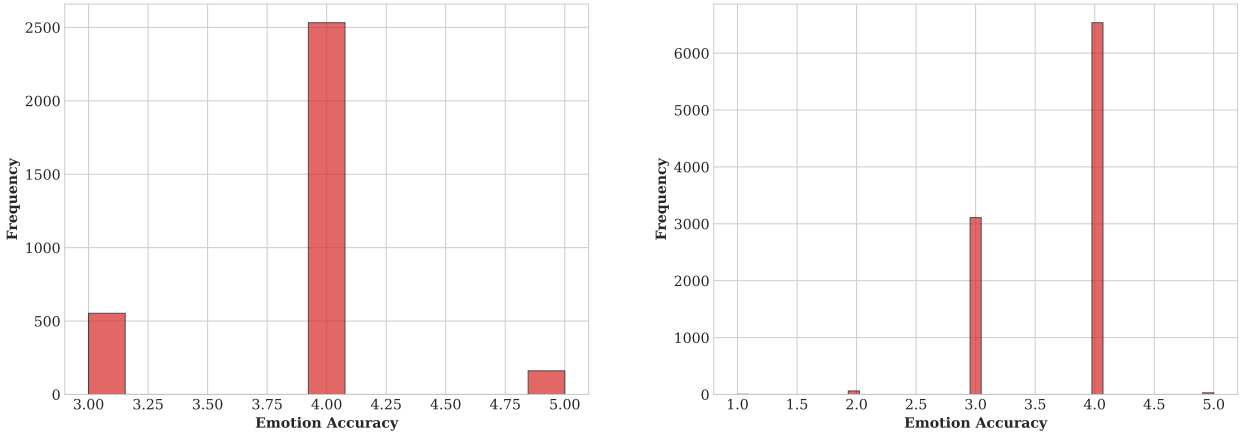


(a) Professional Voice Actor



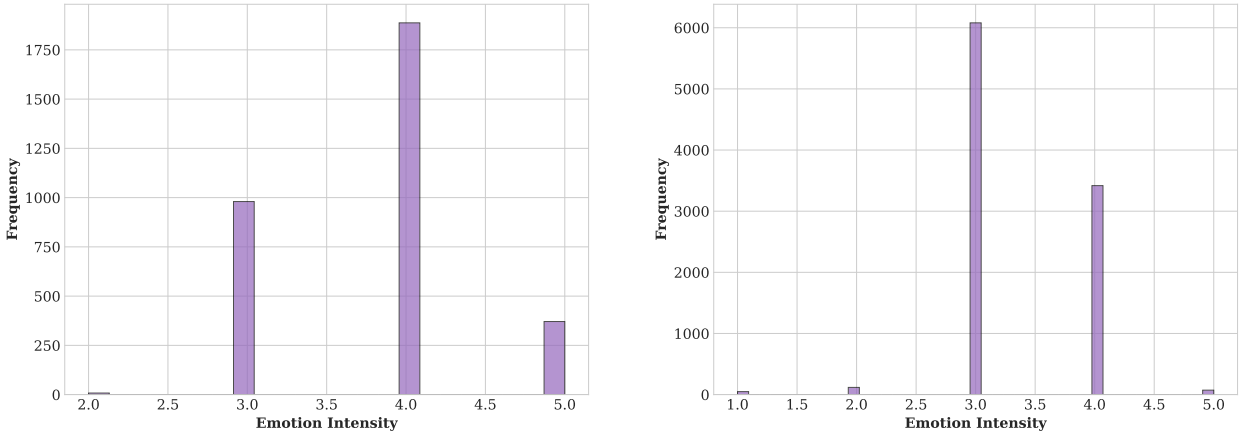
(b) Amateur Voice Actor

Figure 14: Distribution of Stress & Emphasis in Realism Evaluation: Professional vs. Amateur.



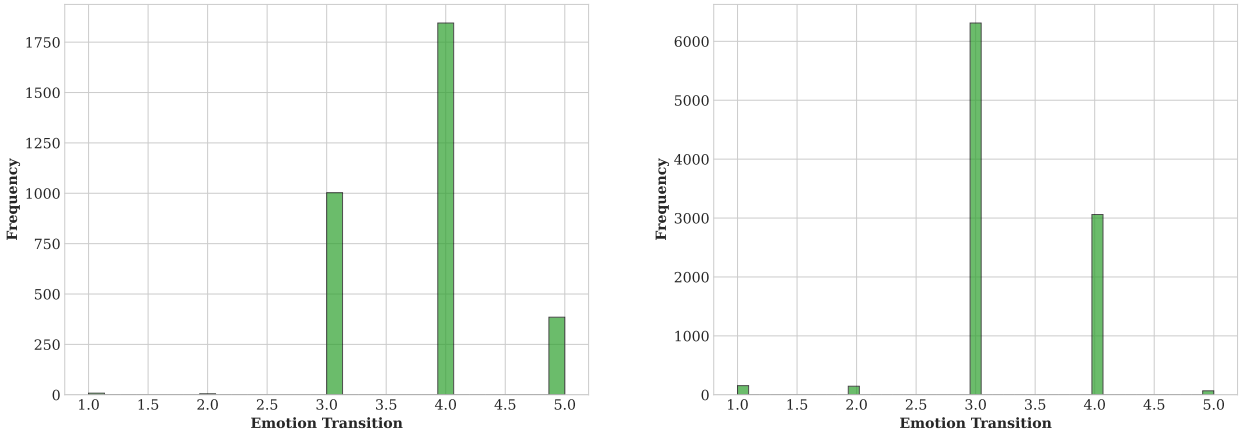
(a) Professional Voice Actor (b) Amateur Voice Actor

Figure 15: Distribution of Emotion Accuracy in Realism Evaluation: Professional vs. Amateur.



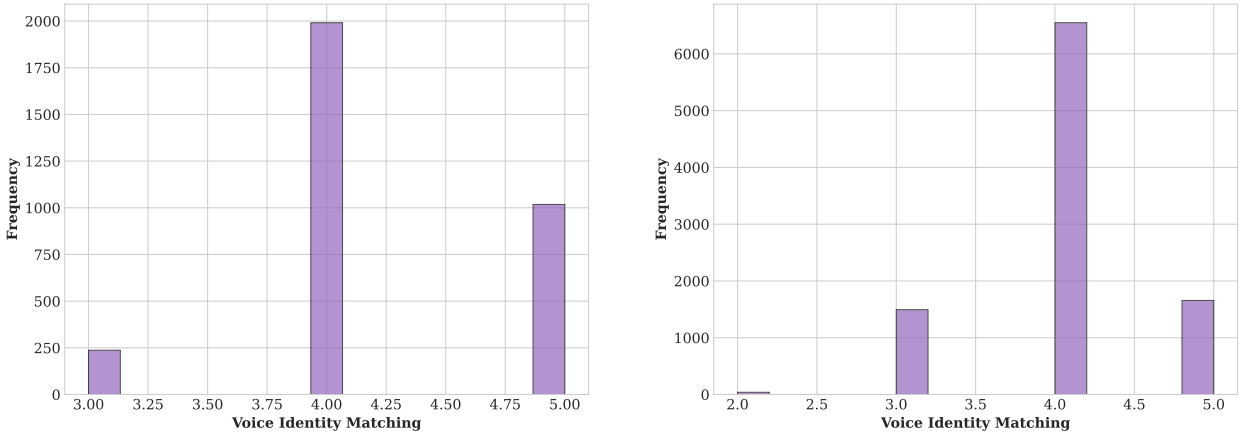
(a) Professional Voice Actor (b) Amateur Voice Actor

Figure 16: Distribution of Emotion Intensity in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor (b) Amateur Voice Actor

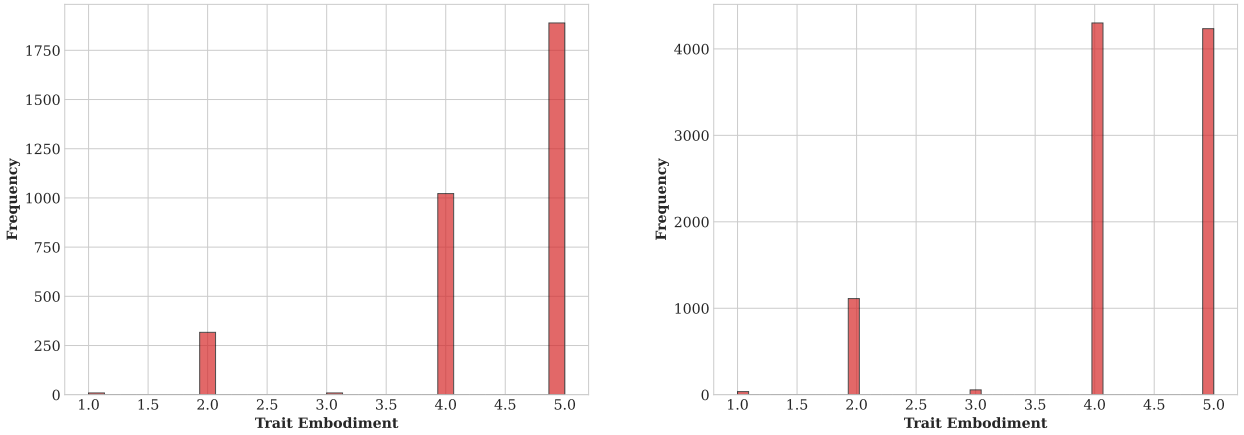
Figure 17: Distribution of Emotion Transition in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor

(b) Amateur Voice Actor

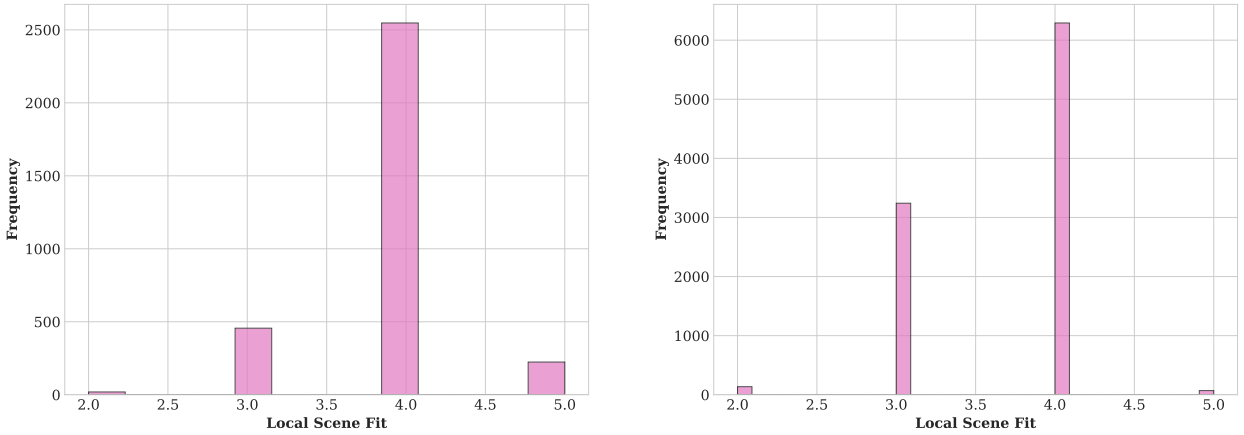
Figure 18: Distribution of Voice Identity Matching in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor

(b) Amateur Voice Actor

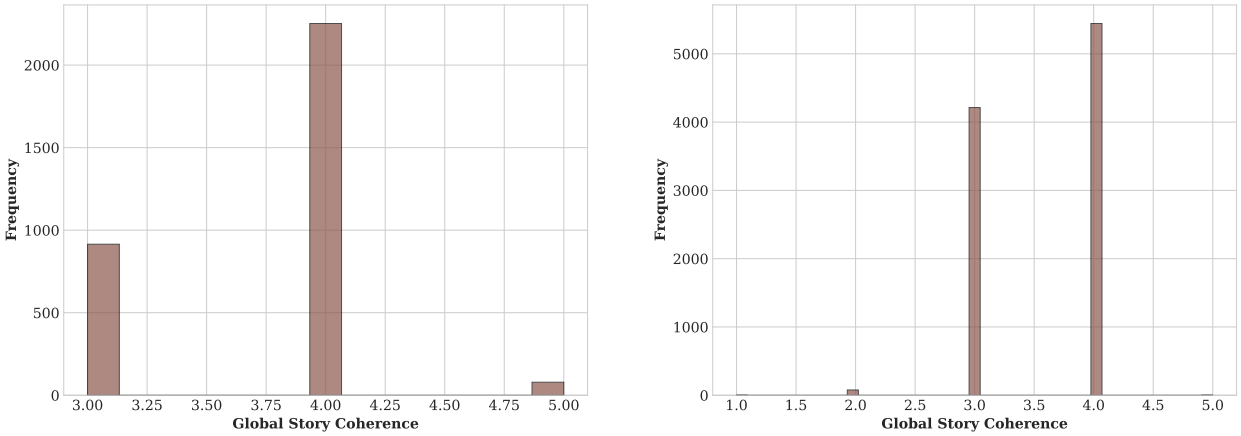
Figure 19: Distribution of Traits Embodiment in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor

(b) Amateur Voice Actor

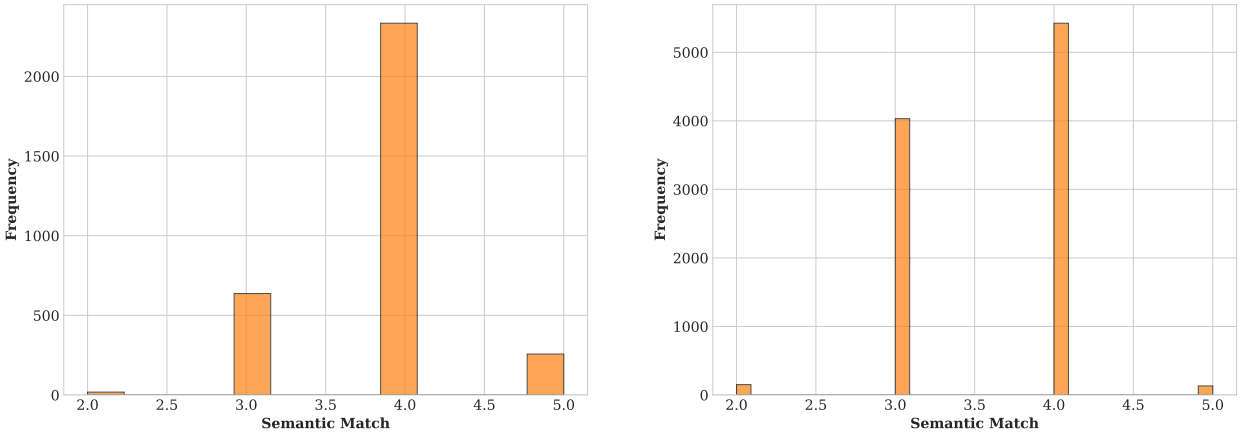
Figure 20: Distribution of Local Scene Fit in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor

(b) Amateur Voice Actor

Figure 21: Distribution of Global Story Coherence in Realism Evaluation: Professional vs. Amateur.



(a) Professional Voice Actor

(b) Amateur Voice Actor

Figure 22: Distribution of Semantic Match in Realism Evaluation: Professional vs. Amateur.