SpeechLLM-as-Judges: Towards General and Interpretable Speech Quality Evaluation

Hui Wang[1*] Jinghua Zhao[1] Yifan Yang[2] Shujie Liu[2†] Junyang Chen[1] Yanzhe Zhang[1]
Shiwan Zhao[1] Jinyu Li[2†] Jiaming Zhou[1] Haoqin Sun[1] Yan Lu[2] Yong Qin[1†]
[1]College of Computer Science, Nankai University
[2]Microsoft Corporation
Correspondence: wanghui_hlt@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

Generative speech technologies are progressing rapidly, but evaluating the perceptual quality of synthetic speech remains a core challenge. Existing methods typically rely on scalar scores or binary decisions, which lack interpretability and generalization across tasks and languages. We present SpeechLLM-as-Judges, a new paradigm for enabling large language models (LLMs) to conduct structured and explanation-based speech quality evaluation. To support this direction, we introduce SpeechEval, a large-scale dataset containing 32,207 multilingual speech clips and 128,754 annotations spanning four tasks: quality assessment, pairwise comparison, improvement suggestion, and deepfake detection. Based on this resource, we develop SQ-LLM, a speech-quality-aware LLM trained with chain-of-thought reasoning and reward optimization to improve capability. Experimental results show that SQ-LLM delivers strong performance across tasks and languages, revealing the potential of this paradigm for advancing speech quality evaluation. The relevant code, models, and data are publicly available at `https://github.com/NKU-HLT/SpeechLLM-as-Judges`.

1 Introduction

Recent advances in generative modeling, including neural text-to-speech (TTS) (Chen et al., 2025b; Wang et al., 2025a), speech-to-speech translation (Barrault et al., 2023), and large-scale spoken dialogue systems (Cheng et al., 2025), have profoundly transformed the field of human-computer interaction. Modern applications such as voice assistants and conversational AI increasingly rely on the ability to generate natural, intelligible, and high-quality speech. In this context, accurately assessing the perceptual quality of the generated speech is essential to ensure the reliability of the system and guide the development of the model (Vinay and Lerch, 2022; Yang et al., 2025c). However, achieving such evaluation remains challenging due to the complexity of perceptual quality factors, the diversity of speech generation tasks, and the lack of standardized, interpretable, and scalable assessment methods.

One of the primary challenges is the lack of interpretability in existing methods. Standard evaluation protocols such as Mean Opinion Score (MOS) and AB preference tests typically provide scalar scores or categorical judgments that reflect overall quality perception, but fail to offer insights into the specific factors that influence those judgments. Similarly, commonly used objective metrics such as Mel Cepstral Distortion (MCD) (Kubichek, 1993) are designed to approximate signal similarity but not offer clear explanations related to human perception (Vinay and Lerch, 2022). The absence of aspect-level attribution and structured reasoning limits their utility for targeted improvement and quality control, which in turn constrains the development of speech generation systems.

Another limitation lies in their limited ability to generalize across both data and task dimensions. On the data side, existing models are typically trained on narrowly scoped datasets, often restricted to a single language, a limited range of sources, and relatively simple annotations. Such constraints lead to weak robustness when models are deployed in multilingual, cross-domain, or real-world scenarios (Wang et al., 2023, 2024). On the task side, most evaluation frameworks are designed to perform a single function, such as absolute quality scoring or preference classification, and thus cannot simultaneously address other critical objectives, including generating improvement suggestions or detecting synthetic speech. Although recent speech foundation models provide architectural flexibility for multi-task evaluation (Chu et al.,

[*]Work done during an internship at Microsoft Research Asia.
[†]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7675–7700

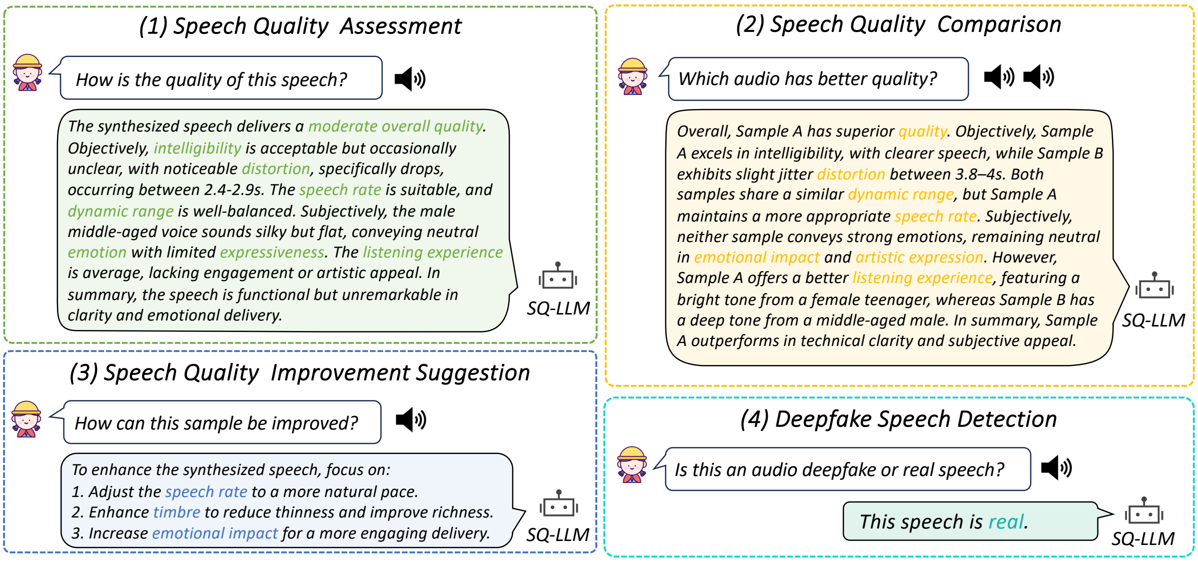


Figure 1: Example interactions showcasing the core capabilities of SpeechLLM-as-Judges. The model supports: speech quality assessment, comparison, improvement suggestion, and deepfake detection.

2024; Dinkel et al., 2025), they frequently underperform in quality assessment, primarily due to insufficient fine-grained, low-level perceptual reasoning (Wang et al., 2025c).

To address these challenges, we propose **SpeechLLM-as-Judges**, a new paradigm that enables large language models (LLMs) to perform general-purpose, interpretable speech quality evaluation. This approach equips existing speech large language models with the ability to reason about, explain, and compare perceptual aspects of speech quality across diverse tasks, as illustrated in Figure 1. To support this goal, we construct SpeechEval, a large-scale multilingual dataset comprising over 30,000 speech clips and 128,000 annotations, spanning four languages and covering multiple evaluation tasks, including descriptive speech quality assessment, pairwise comparison, quality-improvement suggestion, and deepfake speech detection. Building on SpeechEval, we design a speech-quality–aware large language model using instruction tuning with Chain-of-Thought (CoT) reasoning and reward optimization by Policy Gradient Optimization (GRPO) learning to enhance its capabilities. Comprehensive experiments not only validate the effectiveness and feasibility of this paradigm but also highlight its promising potential for advancing speech quality evaluation.

Our work makes the following key contributions:

- We present SpeechEval, a large-scale dataset for speech quality evaluation, containing 128,754 annotations across languages, domains, and evaluation types.
- We propose SQ-LLM, a speech-quality-aware model trained in two stages: instruction tuning with chain-of-thought reasoning and reward learning via GRPO. This approach enables the model to perform interpretable and general-purpose quality evaluation.
- We conduct extensive experiments to evaluate the effectiveness of the SpeechLLM-as-Judges paradigm, demonstrating its advantages in accuracy, interpretability, and generalization across tasks and domains.

2 Related Work

Conventional Speech Quality Evaluation Traditional methods for evaluating speech quality can be broadly categorized into subjective and objective approaches. Subjective protocols rely on human judgment and typically yield scalar or categorical outcomes. While these methods offer valuable perceptual insights, they are inherently time-consuming and labor-intensive (Liu et al., 2025). In contrast, objective metrics like the STOI and MCD (Kubichek, 1993) aim to approximate perceptual quality through signal-based heuristics or learned mappings. Although these techniques are efficient, reproducible, and scalable, they often fall short in providing human-aligned explanations for their assessments (Vinay and Lerch, 2022). These limitations collectively hinder the scalability, efficiency, and accuracy of speech quality evaluation in practical scenarios.

Dataset	# Label	Languages	Quality Assessment		Quality Comparison		Quality Suggestion	Deepfake Detection
			Form	# Dim	Form	# Dim		
ASVspoof2019-LA (Wang et al., 2020)	121,461	English	✗	—	✗	—	✗	✓
VCC2018 (Lorenzo-Trueba et al., 2018)	113,168*	English	Scores	2	✗	—	✗	✗
BC2019 (Wu et al., 2019)	812	Chinese	Scores	1	✗	—	✗	✗
BVCC (Cooper and Yamagishi, 2021)	7,106	English	Scores	1	✗	—	✗	✗
NISQA (Mittag et al., 2021)	14,672	English	Scores	5	✗	—	✗	✗
QualiSpeech (Wang et al., 2025c)	14,577	English	Natural-language (Human-annotated)	7+4	✗	—	✗	✗
ALLD-dataset (Chen et al., 2025a)	25,680	English	Natural-language (LLM-generated)	5	Natural-language (LLM-generated)	5	✗	✗
SpeechEval(OURS)	**128,754**	**Chinese, English, Japanese & French**	**Natural-language** (Human-annotated)	**8+3+5**	**Natural-language** (Human-annotated)	**8+3+5**	✓	✓

* The VCC2018 dataset comprises 113,168 human annotations, including 82,304 naturalness assessments and 30,864 speaker similarity assessments.

Table 1: Overview of existing speech quality evaluation datasets and task coverage.

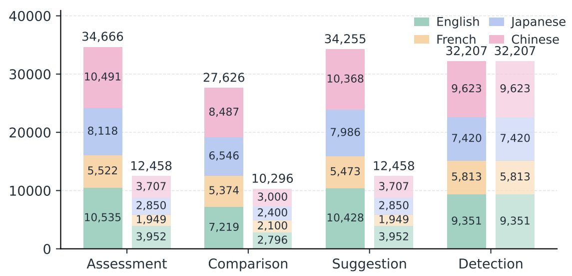


Figure 2: Per-task statistics across four languages, highlighting both the number of audio items (light color) and annotations (solid color).

Task-Specific Models for Speech Quality Prediction Recent efforts have explored learning-based approaches for automatic speech quality prediction, including models such as MOSNet (Lo et al., 2019), UTMOS (Saeki et al., 2022), Audiobox Aesthetics (Tjandra et al., 2025), and RAMP (Wang et al., 2023, 2025b). These systems are typically trained to regress MOS scores or classify quality labels. While such models improve over traditional metrics in terms of data-driven performance, they remain task-specific, language-constrained, and limited to scalar output formats. Few of them support multi-task inference or offer interpretable rationales for their predictions. As a result, their utility for generative speech workflows remains limited.

LLMs for Speech Quality Understanding Recent advances in LLMs have motivated extending evaluation beyond text to speech (Zhang et al., 2025). However, prior studies find that speech LLMs perform poorly on speech-quality benchmarks in direct evaluation and typically require task-specific tuning to produce reliable feedback (Chen et al., 2025a; Wang et al., 2025c,d). Nevertheless, as illustrated in Table 1, their effectiveness remains limited by several factors: most models operate only on English, rely on semi-automatically generated or constrained datasets, and support only a narrow range of tasks. These limitations suggest that current speech LLMs have not yet acquired generalized quality understanding capabilities. In contrast, our work equips speech LLMs with multi-task capabilities through supervised training on large-scale data, enabling structured and interpretable judgments across languages, domains, and evaluation tasks.

3 SpeechEval Dataset

3.1 Task Formulation

To move beyond simple scalar scores and limited task types, we propose four tasks that cover key real-world needs. The first task, **Speech Quality Assessment (SQA)**, focuses on generating natural language descriptions for a single utterance. The model is expected to articulate perceptual impressions across multiple dimensions. The second task, **Speech Quality Comparison (SQC)**, requires the model to compare two utterances and determine which one is of higher quality, along with a justification. This aligns with practical evaluation workflows like A/B testing. The third task, **Speech Quality Improvement Suggestion (SQI)**, introduces a corrective angle: given a suboptimal utterance, the model must suggest actionable modifications that could improve its quality. Lastly, **Deepfake Speech Detection (DSD)** concerns the ability to distinguish human speech from synthetic or manipulated speech. This task is closely tied to perceptual realism and quality degradation, making it a relevant subtask for quality-oriented systems.

3.2 Dataset Overview

The SpeechEval dataset contains 32,207 unique utterances and 128,754 annotations across four major tasks. Each task includes speech samples in English, Chinese, Japanese, and French, covering a broad range of speakers, speaking styles, voice

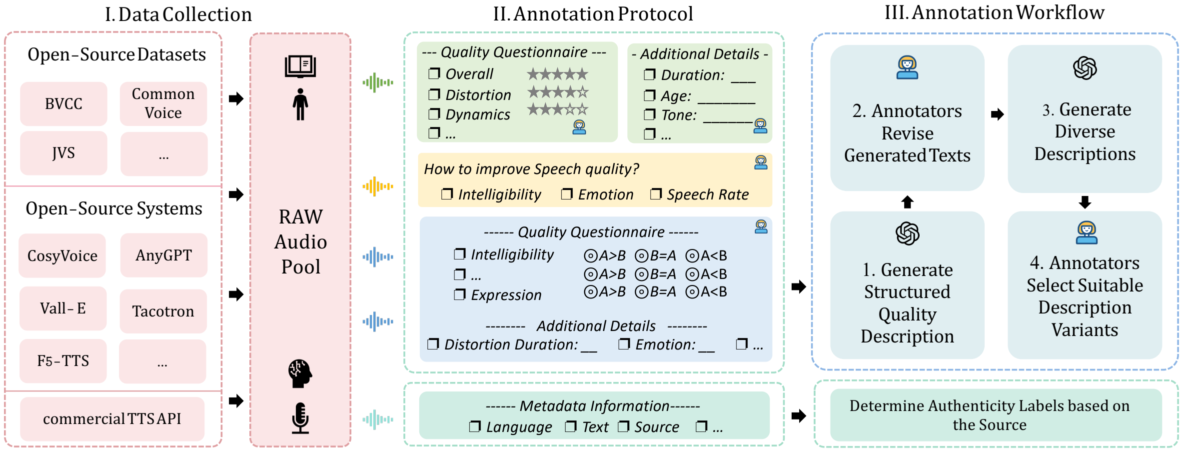


Figure 3: Overview of the SpeechEval data construction process, including data collection (left), task-specific annotation protocols (middle), and a human-in-the-loop annotation workflow with LLM assistance (right).

qualities, and acoustic conditions. As shown in Figure 2, the data is distributed relatively evenly across tasks and languages. Chinese and English together account for the majority of samples, while Japanese and French are also well represented, ensuring multilingual coverage. All annotations are paired with structured labels and natural language explanations. A demonstration of the data, along with a detailed breakdown of the annotation and the associated metadata, can be found in Appendix A.

3.3 Dataset Construction

Data Collection We collect speech samples from multiple sources to ensure diversity in speaker characteristics, content, and quality. Specifically, we include real speech from public corpora (Ardila et al., 2020; Lorenzo-Trueba et al., 2018; Cooper and Yamagishi, 2021; Panayotov et al., 2015; Shi et al., 2021; BZNSYP, 2020; Wu et al., 2019; Sonobe et al., 2017; Takamichi et al., 2019; Honnet et al., 2017), along with synthetic speech generated using open-source speech generation systems (Li et al., 2019; Kim et al., 2021; Wang et al., 2017; Chien et al., 2021; Zhan et al., 2024; Zhang et al., 2023; Wang et al., 2025e; Chen et al., 2024; Suno, 2024; ChatTTS, 2024; Du et al., 2024; Chen et al., 2025b). Commercial TTS engines, such as Aliyun[1], Volcengine[2], and Microsoft TTS[3], are also included.

Annotation Protocol Each sample is annotated using a structured schema that captures speech quality. As shown in Table 2, we define three high-level aspects: overall rating, objective production quality, and subjective content enjoyment. These aspects are decomposed into eight subdimensions, including intelligence, distortion, speech rate, dynamic range, tone balance, emotional impact, artistic expression, and subjective experience. Depending on the task, annotators provide either ratings or pairwise comparisons for these dimensions. In addition, we collect three types of categorical metadata: distortion type, emotion type, and speaker gender. We also include five open-ended fields covering distortion duration, distortion severity, perceptual description, speaker age, and speaking tone. This annotation schema supports both structured supervision and natural language supervision.

Aspects	Sub-dimensions
Overall	(1) Overall Quality
Production Quality	(2) Intelligibility (3) Distortion (4) Speech Rate (5) Dynamic Range
Content Enjoyment	(6) Emotional Impact (7) Artistic Expression (8) Subjective Experience

Table 2: Structured speech quality annotation protocol. See Appendix A.3 for details.

Annotation Workflow To balance annotation quality with scalability, we adopt a human-in-the-loop pipeline as illustrated in Figure 3. For the Speech Quality Assessment, Comparison, and Suggestion tasks, annotators begin by completing a

[1]Aliyun TTS: https://ai.aliyun.com/nls/tts
[2]Volcengine TTS: https://console.volcengine.com
[3]Microsoft TTS: https://speech.microsoft.com/portal/voicegallery

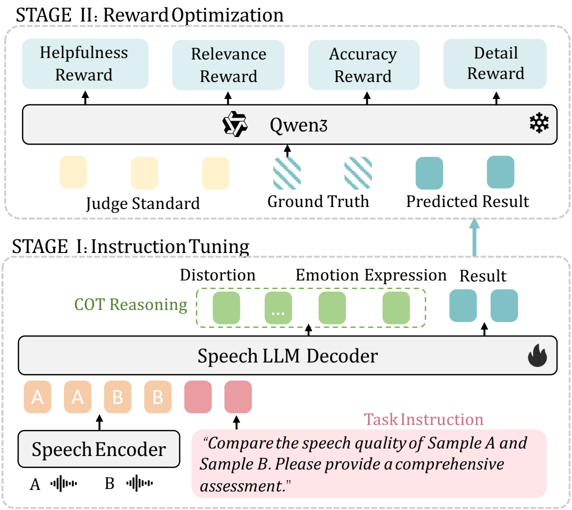


Figure 4: Overview of SQ-LLM training. Stage I uses instruction tuning with dimension-wise CoT reasoning. Stage II applies multi-aspect feedback for refinement.

structured quality questionnaire that captures both judgments and additional information. The responses are passed to an LLM that generates initial textual descriptions. Annotators review and revise these drafts to ensure accuracy and clarity. The revised texts are subsequently passed back to the LLM to produce a set of diverse candidates, from which annotators select appropriate variants. This semi-automated process enables efficient production of high-quality and linguistically diverse annotations and is adaptable across both tasks and languages. For the Deepfake Speech Detection task, authenticity labels are derived based on metadata about the audio's origin. These labels are assigned by tracing the source of the sample.

4 SQ-LLM

Figure 4 presents SQ-LLM, a unified speech evaluation model. This section outlines its architecture and two-stage training; see Appendix B for details.

4.1 Model Architecture

We design SQ-LLM to provide a unified solution for four distinct speech quality evaluation tasks: quality assessment, comparison, improvement suggestion, and deepfake detection. Rather than treating each task separately, SQ-LLM formulates them within a single instruction-based framework, where all tasks are handled through natural language prompts and structured outputs.

The model consists of a speech encoder and a speech-aware language decoder, both built on top of the Qwen2.5-Omni (Xu et al., 2025). The encoder transforms input speech signals into continuous representations that capture relevant acoustic features. These are combined with the textual task instruction and passed to the decoder, which generates outputs such as quality dimension scores, free-form rationales, or classification results, depending on the task. This architecture allows SQ-LLM to generalize across tasks and languages while maintaining consistent, interpretable outputs.

4.2 Instruction Tuning with CoT Reasoning

To improve interpretability and consistency in generation, we adopt a reasoning-oriented instruction tuning strategy for the assessment, comparison, and suggestion tasks. Given a task instruction and one or more input utterances, the model first generates coarse predictions over a set of $N = 8$ predefined quality dimensions. These dimension-wise predictions are used as intermediate signals that guide the final result. Since these annotations are already collected during human labeling, they serve as natural supervision targets for structured reasoning.

We define a joint training objective that encourages the model to produce accurate intermediate scores and coherent explanations. Formally, the overall loss is defined as:

$$\mathcal{L} = \lambda \sum_{i=1}^{N} \cdot \mathcal{L}_{\text{dim}}^{(i)} + \mathcal{L}_{\text{ans}}, \tag{1}$$

where $\mathcal{L}_{\text{dim}}^{(i)}$ denotes the loss for the i-th quality dimension, and $\mathcal{L}_{\text{ans}}$ is the cross-entropy loss for the final answer. λ controls the contribution of intermediate reasoning.

4.3 Reward Optimization via GRPO

To further enhance the output alignment with human preferences, we apply Generalized Policy Gradient Optimization (DeepSeek-AI, 2025) to fine-tune SQ-LLM based on task-specific reward functions. Instead of relying on preference pairs, we design automatic reward evaluators for each task that score the model along four dimensions: Helpfulness, Relevance, Accuracy, and Level of Detail.

These rewards are produced by a single frozen evaluator, instantiated as Qwen3 (Yang et al., 2025a). Let p be the input prompt, y be the model output, g be the ground-truth metadata, and $t \in \mathcal{T} = \{\text{Assessment}, \text{Comparison}, \text{Suggestion}\}$. For each dimension, referred to as $d \in \mathcal{D} = \{\text{Helpfulness}, \text{Relevance}, \text{Accuracy}, \text{Detail}\}$, evaluator first returns an integer score $\mathcal{E}_d \in$

Task	Train	Val.	Test	Total
Assessment	23,769	5,392	5,505	34,666
Comparison	19,260	4,069	4,297	27,626
Suggestion	23,494	5,316	5,445	34,255
Detection	6,600	5,724	19,883	32,207
Total	73,123	20,501	35,130	128,754

Table 3: The data split across different tasks.

$\{0, 1, 2 \ldots, 10\}$, which we normalize and aggregate with fixed weights. We define normalized rewards $r_d \in [0, 1]$ and the total reward R_{total}:

$$r_d(p, y, g; t) = \begin{cases} \mathcal{E}_d(p, y, g)/10, & \text{if } t \in \mathcal{T}, \\ \mathbf{1}\{y = g\}, & \text{if } t \notin \mathcal{T}, \end{cases} \quad (2)$$

$$R_{\text{total}} = \sum_{d \in \mathcal{D}} \lambda_d \, r_d. \quad (3)$$

For instance, a quality analysis is rewarded when it accurately identifies the type and impact of distortion, while a comparative response is recognized for justifying its choice with coherent and well-reasoned arguments. The aggregated reward R_{total} serves as the learning signal within the GRPO framework to update the policy, improving output quality without sacrificing controllability.

5 Experimental Setup

5.1 Implementation Details

Dataset Split Table 3 shows the data split. For the first three tasks, the audio data is divided into training, validation, and test sets in roughly a 70%, 15%, and 15%. We ensure that validation and test splits contain unseen speakers, systems, and text content to evaluate generalization. For the detection task, we adopt a protocol based on prior work (Wang et al., 2020). The dataset includes both seen and unseen spoof sources for in-distribution and zero-shot evaluation. The splits are imbalanced by design to reflect real-world distribution. Moreover, we remove cross-task data leakage. Full details of the splitting strategy are in Appendix C.1.

Model Configuration We fine-tune SQ-LLM based on Qwen2.5-Omni-7B using the Swift framework (Zhao et al., 2025) with LoRA and the speech encoder frozen. The model is trained for 8 epochs, a batch size of 4, and a learning rate of 1e-4. and $\lambda = 0.3$ in the instruction tuning stage. We further apply GRPO with a batch size of 1, sampling 4 candidate generations per prompt, and train the LoRA adapters with a learning rate of 1e-6. $\lambda_{\text{Helpfulness}}$, $\lambda_{\text{Relevance}}$, $\lambda_{\text{Accuracy}}$ and λ_{Detail} are set to 1, 1, 2 and 0.5, respectively. Overall, training SQ-LLM requires approximately 43 A100 GPU-hours, including about 12 GPU-hours for SFT and 31 GPU-hours for GRPO.

5.2 Metrics

For generation tasks (Assessment, Comparison, and Suggestion), we adopt a combination of automatic metrics, LLM-based, and human evaluation. Specifically, (i) the automatic metrics include SBERT and FENSE (Zhang et al., 2019; Zhou et al., 2022). (ii) We use LLM Score (LScore) to assess overall performance, and leverage an LLM to extract dimension-level predictions from generated texts: in the Assessment task, we report Pearson correlation coefficients (PCC) over 7 quality dimensions, together with accuracy (ACC) for speech-rate classification; in the Comparison task, we report ACC for dimension-wise preference. (iii) We conduct human evaluation to verify the models: we recruit 8 raters to score model outputs on a 1–5 scale (higher is better). For each model, we sample 60 instances, and each instance is independently rated by at least three raters.

For the Deepfake Speech Detection task, we report equal error rate (EER), minimum detection cost function (minDCF), and accuracy. The calculation details are given in Appendix C.2 and C.3.

5.3 Baselines

We consider three categories of baseline models to evaluate the effectiveness of SQ-LLM. First, we directly benchmark recent multimodal LLMs (Chu et al., 2024; Dinkel et al., 2025; Xu et al., 2025) without any task-specific tuning to assess their general capability for speech quality understanding. Second, we fine-tune existing open-source audio LLMs and our custom-constructed models by coupling advanced speech encoders (e.g., Whisper, WavLM, and AudioBox Aesthetics encoder (AES-E) (Radford et al., 2023; Chen et al., 2022; Tjandra et al., 2025)) with an LLM backbone (Yang et al., 2025b). Third, we include task-specific conventional expert models as strong non-LLM baselines; these systems are designed for a single task. Concretely, for the speech quality assessment and comparison tasks, we evaluate MOSNet, UTMOS, and AudioBox Aesthetics as representative expert baselines, while for the deepfake speech detection task,

MODEL	I. Quality Assessment			II. Quality Comparison			III. Improvement Suggestion		
	LScore	SBERT	FENSE	LScore	SBERT	FENSE	LScore	SBERT	FENSE
Qwen2-Audio-7B	4.161	0.580	0.311	4.591	0.703	0.580	5.176	0.548	0.460
Qwen2.5-Omni-7B	4.610	0.593	0.593	3.602	0.704	0.437	6.426	0.537	0.364
MiDashengLM-7B	5.536	0.699	0.635	4.042	0.665	0.444	6.653	0.600	0.490
Qwen3-8B + Whisper	6.422	**0.803**	**0.699**	5.122	0.818	0.732	**7.478**	0.583	0.574
Qwen2.5-7B + AES-E	6.533	0.772	0.634	5.270	0.793	0.589	6.891	0.530	0.529
Qwen3-4B + WavLM	6.163	0.736	0.585	5.138	0.796	0.618	7.209	0.539	0.537
FT Qwen2-Audio-7B	6.440	0.800	0.559	5.648	0.853	0.686	7.263	0.708	0.708
SQ-LLM (OURS)	**6.833**	**0.803**	0.614	**6.434**	**0.864**	**0.749**	7.420	**0.735**	**0.735**

Table 4: Performance of different models on the Assessment, Comparison, and Improvement Suggestion tasks.

MODEL	Assessment	Comparison
Qwen2-Audio-7B	0.073	0.310
Qwen2.5-Omni-7B	0.070	0.203
MiDashengLM-7B	0.158	0.265
MOSNet	0.333	0.726
UTMOS	0.375	0.741
Audiobox Aesthetics	0.464	0.737
Qwen3-8B + Whisper	0.424	0.591
Qwen2.5-7B + AES-E	0.457	0.577
Qwen3-4B + WavLM	0.422	0.563
FT Qwen2-Audio-7B	0.244	0.587
SQ-LLM (OURS)	**0.520**	**0.751**

Table 5: Results report PCC for Assessment and ACC for Comparison, both on the Overall Quality dimension.

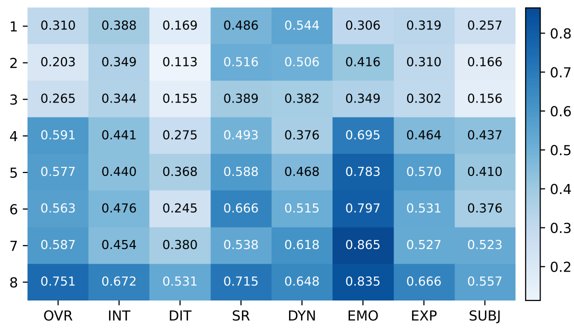


Figure 5: Heatmap of dimension-level accuracy across models on the Speech Quality Comparison task over eight quality dimensions. The 1–8 indices follow the same top-to-bottom order as in Table 4.

we adopt RawNet2, AASIST, and AASIST2. Detailed descriptions are provided in Appendix C.4.

6 Results and Analysis

6.1 Comparison Studies

Speech Quality Assessment Table 4-I presents the assessment results. In the direct-evaluation setting, models generally struggle to capture speech quality; MiDashengLM attains the highest LLM score among them, yet alignment with human ratings and text-based metrics remains weak. When we build custom-constructed pipelines or fine-tune audio LLMs, both the LLM score and the correlation with humans improve, accompanied by consistent gains on traditional metrics. SQ-LLM performs best overall, exhibiting the strongest agreement with human ratings and robust text metrics.

Table 5 shows that recent multimodal LLMs exhibit very low PCC on the Overall Quality dimension, indicating weak alignment with human. Expert MOS predictors markedly improve correlation, while SQ-LLM achieves the highest PCC, surpassing the best expert baseline and LLM variants.

Speech Quality Comparison On the comparison task (Table 4-II), directly evaluated multimodal LLMs show limited ability to discriminate fine-grained quality differences: Qwen2-Audio is the strongest, whereas Qwen2.5-Omni and MiDashengLM lag behind. Equipping models with custom-constructed pipelines or fine-tuning audio LLMs delivers consistent gains. SQ-LLM further advances all dimensions, achieving the best agreement with human. From a correlation-oriented perspective in Table 5, models that better capture the monotonic quality ordering tend to yield higher comparison accuracy. Consistent with this trend, SQ-LLM attains the best ACC, outperforming expert baselines and LLM-based alternatives.

Figure 5 visualizes performance across eight quality dimensions. From a dimension-wise perspective, distortion and subjective experience are consistently the most challenging dimensions, with noticeably lower scores for most baselines, while speech rate and dynamic range are comparatively easier and show smaller performance gaps. Notably, SQ-LLM exhibits the most balanced strength, achieving clear gains on harder dimensions.

MODEL	EER(%)	minDCF	ACC(%)
Qwen2-Audio-7B	–	–	21.220
Qwen2.5-Omni-7B	–	–	27.250
MiDashengLM-7B	–	–	67.480
RawNet2	15.836	0.325	72.036
AASIST	18.158	0.450	71.599
AASIST2	18.633	0.414	72.001
Qwen3-8B + Whisper	16.502	0.411	78.977
Qwen2.5-7B + AES-E	45.241	0.992	67.741
Qwen3-4B + WavLM	44.300	1.000	20.837
FT Qwen2-Audio-7B	8.593	0.194	89.312
SQ-LLM (OURS)	**6.249**	**0.142**	**89.358**

Table 6: Results on deepfake speech detection.

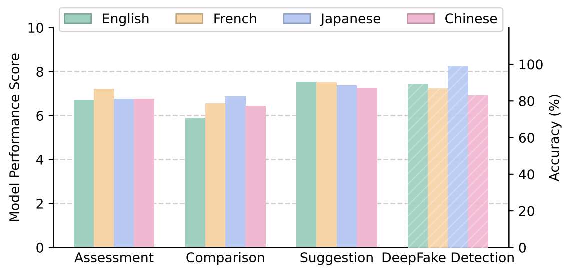


Figure 6: Performance of SQ-LLM across tasks and languages. Left y-axis reports LLM Scores; right y-axis shows accuracy for Deepfake Detection.

Speech Quality Improvement Suggestion Table 4-III reports the results for generating actionable quality-improvement suggestions. Under direct evaluation, systems often produce vague or generic advice, MiDashengLM leads that group on the LLM score, but traditional metrics remain weak. Training markedly helps: both custom-constructed pipelines and fine-tuned audio LLMs improve lexical overlap and informativeness. Notably, Whisper encoder attains the strongest LLM score among the trained baselines. SQ-LLM offers the best overall balance of LLM scores and traditional metrics.

Deepfake Speech Detection For detection task, results are summarized in Table 6. Expert systems provide solid, well-calibrated baselines. In contrast, untuned multimodal LLMs underperform and do not yield calibrated metrics, indicating weak out-of-the-box reliability. Custom-constructed model narrows the gap but is front-end sensitive: Whisper becomes competitive across all metrics, whereas other pipelines show marked degradation. Supervised fine-tuning helps substantially. While SQ-LLM achieves the best results overall, setting the strongest EER, minDCF, and accuracy.

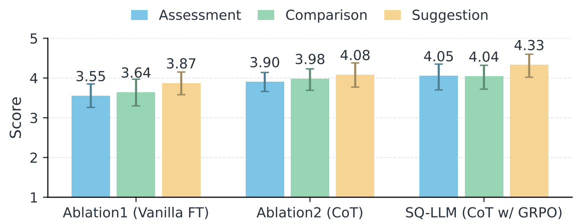


Figure 7: Human-evaluated ablation results on the Assessment, Comparison, and Suggestion tasks with 95% item-level bootstrap confidence intervals.

CoT	GRPO	SQA ↑	SQC ↑	SQI ↑	DSD ↓
✓	✓	**6.833**	**6.434**	**7.420**	**6.249**
✓	✗	6.804	6.420	7.018	6.264
✗	✗	6.657	6.316	6.733	8.574

Table 7: Ablation results across four tasks. Higher LLM Score is better for SQA, SQC, and SQI; EER (%) is reported for DSD.

6.2 Performance Analysis

Figure 6 presents the performance of SQ-LLM across four tasks and four languages. The model demonstrates consistent and strong results in the Assessment, Comparison, and Suggestion tasks, with scores generally above 6.5 across all languages. For the Detection task, the model achieves high accuracy in Japanese, while performance in Chinese and French is relatively lower. These results highlight the model's robust multilingual capability, and a more detailed analysis of the deepfake detection task is provided in Appendix D.

In Figure 7 and Table 7, we perform an ablation study to quantify the effects of CoT tuning and GRPO. Across both presentations, we observe consistent improvements as stronger components are introduced. Enabling CoT yields clear gains on the generation-style tasks, indicating that structured reasoning leads to more accurate and helpful outputs. Adding GRPO on top of CoT further improves performance, with the largest benefit on the Suggestion task, suggesting that preference optimization is particularly effective for open-ended recommendation generation. The deepfake detection results follow the same trend, where CoT substantially enhances robustness and the CoT-with-GRPO variant achieves the best overall performance. See Appendix D for a detailed analysis.

7 Conclusion

We present SpeechLLM-as-Judges, a unified framework for interpretable and generalizable speech quality evaluation. Supported by the SpeechEval dataset, SQ-LLM is trained to handle diverse evaluation tasks through instruction tuning with reasoning and reward optimization. Experiments demonstrate its ability to produce human-aligned, explanatory outputs. This work highlights the potential of speech LLM as a reliable and versatile evaluator in speech generation systems.

Limitations

While our work provides a unified framework for speech quality evaluation, it also has limitations. The current version of SpeechEval focuses on four languages and a fixed set of tasks. Expanding to more low-resource or code-switched languages, as well as incorporating additional evaluation scenarios such as emotional expressiveness or speaker consistency, would further enhance the model's coverage and applicability. We leave these directions for future work.

Ethics Statement

All data used in this work were collected and processed in accordance with relevant ethical guidelines and licensing terms. The speech samples are sourced from publicly available or properly licensed datasets. Human annotations were conducted by trained annotators who received fair compensation. No personally identifiable information was collected during the annotation process. In this work, LLMs are employed for semi-automated annotation assistance and model evaluation, enhancing annotation efficiency and providing more accurate evaluation results.

Acknowledgments

This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

BZNSYP. 2020. data-baker official website.

ChatTTS. 2024. Chattts official website.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025a. Audio large language models can be descriptive speech quality evaluators. *arXiv preprint arXiv:2501.17202*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2025b. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *ArXiv*, abs/2410.06885.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, and 1 others. 2025. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*.

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592.

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

Erica Cooper and Junichi Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? *arXiv preprint arXiv:2105.02373*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2025. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv: 2508.03983*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2019. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662.

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. 2025. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. *arXiv preprint arXiv:2505.11079*.

Pierre-Edouard Honnet, Alexandros Lazaridis, Philip N Garner, and Junichi Yamagishi. 2017. The siwis french speech synthesis database? design and recording of a high quality french database for speech synthesis. *Online Database*.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning-based objective assessment for voice conversion. In *Interspeech 2019*, pages 1541–1545.

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

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

Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*.

Suno. 2024. Suno official website.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. 2019. Jvs corpus: free japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, and 1 others. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Ashvala Vinay and Alexander Lerch. 2022. Evaluating generative audio systems and their metrics. *arXiv preprint arXiv:2209.00130*.

Hui Wang, Shujie Liu, Lingwei Meng, Jinyu Li, Yifan Yang, Shiwan Zhao, Haiyang Sun, Yanqing Liu, Haoqin Sun, Jiaming Zhou, and 1 others. 2025a. Felle: Autoregressive speech synthesis with token-wise coarse-to-fine flow matching. *arXiv preprint arXiv:2502.11128*.

Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023. Ramp: Retrieval-augmented mos prediction via confidence-based dynamic weighting. In *Interspeech 2023*, pages 1095–1099.

Hui Wang, Shiwan Zhao, Xiguang Zheng, Jiaming Zhou, Xuechen Wang, and Yong Qin. 2025b. Ramp+: Retrieval-augmented mos prediction with prior knowledge integration. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1520–1534.

Hui Wang, Shiwan Zhao, Jiaming Zhou, Xiguang Zheng, Haoqin Sun, Xuechen Wang, and Yong Qin. 2024. Uncertainty-Aware Mean Opinion Score Prediction. In *Interspeech 2024*, pages 1215–1219.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025c. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23588–23609, Vienna, Austria. Association for Computational Linguistics.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, and Chao Zhang. 2025d. Enabling auditory large language models for automatic speech quality evaluation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

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

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017*, pages 4006–4010.

Zhizheng Wu, Zhihang Xie, and Simon King. 2019. The blizzard challenge 2019. In *Proc. Blizzard Challenge Workshop*, volume 2019, pages 1–24.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv: 2503.20215*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025b. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yifan Yang, Hui Wang, Bing Han, Shujie Liu, Jinyu Li, Yong Qin, and Xie Chen. 2025c. Towards responsible evaluation for text-to-speech. *arXiv preprint arXiv:2510.06927*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, and 1 others. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and 1 others. 2025. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*.

Yuxiang Zhang, Jingze Lu, Zengqiang Shang, Wenchao Wang, and Pengyuan Zhang. 2024. Improving short utterance anti-spoofing with aasist2. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11636–11640. IEEE.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985. IEEE.

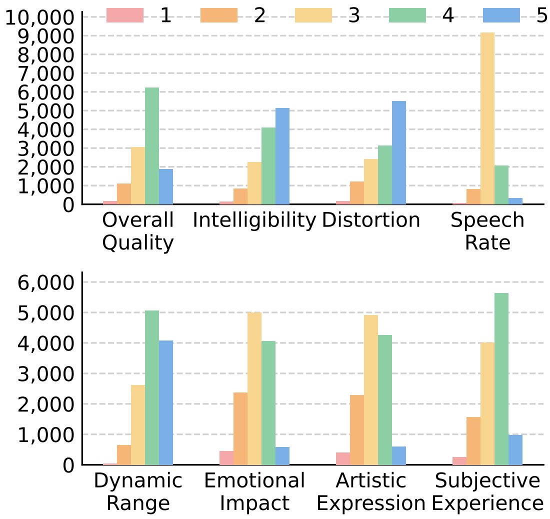
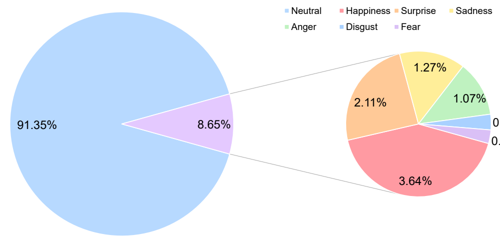
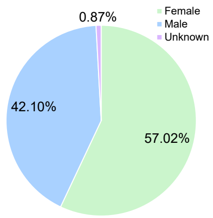


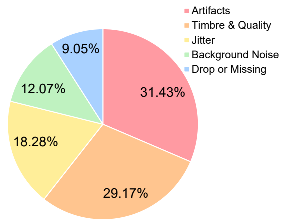
Figure 8: Distribution of assessment scores across the eight dimensions. The speech rate dimension is rated from 1 (too slow) to 5 (too fast), while other dimensions use a 1–5 scale representing perceptual quality.



(a) Emotion distribution: overall on the left, and the breakdown of non-neutral emotions on the right.



(b) Gender distribution.



(c) Distortion type distribution.

Figure 9: Categorical metadata statistics in the SpeechEval assessment data: (a) Emotion distribution, with non-neutral emotions detailed on the right; (b) Gender distribution; (c) Distortion type distribution.

A Dataset

A.1 Dataset Distribution

Figure 8 presents the distribution of assessment scores across the eight annotated dimensions in the SpeechEval dataset. Most dimensions, including Overall Quality, Intelligibility, Distortion, and Subjective Experience, exhibit a reasonably balanced score distribution across the five-point scale, which reflects the diversity and coverage of our collected speech samples. Notably, the Speech Rate dimension shows a skewed distribution towards the extreme values, with a significant peak at score 5. This is due to the annotation scale for Speech Rate ranging from 1 (too slow) to 5 (too fast), where deviations from the optimal speed tend to be penalized more heavily. The score distribution indicates that our dataset provides rich and varied supervision signals, which are crucial for training and evaluating speech models across different perceptual aspects.

In Figure 9, we further analyze the categorical metadata statistics within the speech quality assessment task. The majority of samples are labeled as Neutral in terms of emotion, but non-neutral emotions are also represented, enhancing the emotional diversity of the dataset. Gender labels are balanced between male and female speakers, with a small portion marked as unknown. In terms of distortion types, the dataset includes a mix of artifacts, timbre issues, jitter, background noise, and missing segments, with artifacts being the most prevalent. These attribute distributions demonstrate that SpeechEval provides a comprehensive foundation for evaluating model robustness across realistic speech variations.

A.2 Dataset Examples

Speech Quality Assessment A full evaluation for a low-quality speech sample is presented in Figure 10. Overall and intelligibility scores are low; continuous background noise and timbre/quality distortion are present across 0–5 s, described as

1. Overall Quality
Score: 2
2. Production Quality
Intelligibility Score: 2
Distortion Score: 2 *Distortion Type: background noise;timbre & quality*
Distortion Duration: appeared between 0 s - 5 s (across entire recording)
Distortion Degree: Noticeable *Distortion Description: rough and hissing*
Speech Rate: suitable
Dynamic Range Score: 3
3. Content Enjoyment
Emotional Impact Score: 2 *Emotional Type: Neutral*
Artistic Expression Score:2
Subjective Experience Score:2 *Gender: female* *Age: middle-aged* *Tone Description: The voice sounds rough and hissing*
4. Detailed Description
The speech has significant quality issues, falling short in clarity and naturalness. Objectively, intelligibility is poor, with noticeable background noise and timbre distortion throughout the 0-5 s duration, creating a rough, hissing effect. The speech rate is suitable, and dynamics are moderately consistent. Subjectively, the female middle-aged voice lacks emotional engagement and expressiveness, sounding neutral and unappealing. The overall listening experience is unpleasant due to the voice's rough, distorted quality.
5. Alternative Description
1) *The speech exhibits considerable quality shortcomings, primarily in clarity and natural delivery. From an objective standpoint, intelligibility is subpar, with audible background noise and timbre distortion spanning the 0-5 second range, producing a coarse, sibilant effect. The speech rate is well-paced, and dynamic consistency is moderate. Subjectively, the middle-aged female voice fails to convey emotion, sounding monotonous and uninspiring. The rough, distorted vocal quality detracts significantly from the listening experience.*
2) *The speech has clear quality issues, struggling with both clarity and natural flow. Objectively, it's hard to understand due to constant background noise and a distorted tone (0-5s), which adds a rough, hissing quality. The speed is good, and the volume changes are fairly even. Subjectively, the middle-aged female voice sounds flat and unemotional, failing to hold interest. The overall effect is unpleasant because of the voice's scratchy, uneven sound.*

Figure 10: Assessment example for a low-quality speech sample.

1. Objective Evaluation
Score: 5
2. Production Quality
Intelligibility Score: 5
Distortion Score:5 *Distortion Type: None*
Distortion Duration: None
Distortion Degree: None *Distortion Description: None*
Speech Rate: suitable
Dynamic Range Score: 4
3. Content Enjoyment
Emotional Impact Score: 4 *Emotional Type: Surprise*
Artistic Expression Score:4
Subjective Experience Score:5 *Gender: male* *Age: middle-aged* *Tone Description: The voice sounds bright and full*
4. Detailed Description
The speech demonstrates excellent overall quality, delivering a highly polished performance. Objectively, it excels in intelligibility and lacks any distortion, with a suitable speech rate and smooth dynamics, though volume consistency could be slightly improved. From a subjective perspective, the male middle-aged speaker conveys a bright, full tone with a sense of surprise, enhancing artistic expression and providing an engaging listening experience. The speech is expressive and well-balanced, making it highly enjoyable.
5. Alternative Description
1) *The speech exhibits outstanding quality, presenting a refined and professional delivery. Objectively, it achieves high intelligibility with no distortion, maintaining an appropriate pace and fluid dynamics, though minor volume fluctuations could be addressed. Subjectively, the middle-aged male speaker projects a vibrant, resonant tone infused with a hint of surprise, enriching artistic expression and ensuring an immersive experience. The expressive yet balanced nature of the speech makes it highly appealing.*
2) *This speech is of exceptional quality, showcasing a polished and articulate performance. On a technical level, clarity is excellent, with no audible distortion, and the pacing and flow are well-managed, though slight volume adjustments could enhance consistency. From a listener's perspective, the speaker—a middle-aged man—delivers a warm, rich tone with an undercurrent of surprise, heightening expressiveness and engagement. The overall delivery is dynamic and harmonious, making it a pleasure to hear.*
3) *This speech demonstrates remarkable quality, featuring a highly professional and engaging delivery. Technically, it offers excellent intelligibility without distortion, with a well-modulated speech rate and smooth dynamics, though volume uniformity could be fine-tuned. From a subjective standpoint, the middle-aged male speaker's lively, well-rounded tone, infused with subtle surprise, enhances expressiveness and captivates the audience. The balanced and expressive nature of the speech makes it highly satisfying.*

Figure 11: Assessment example for a high-quality speech sample.

rough and hissing. Rate and dynamics are acceptable, but emotional engagement is weak, and the tone sounds scratchy. The detailed and alternative descriptions explain how these distortions degrade clarity and listening experience.

As a counterpoint, Figure 11 depicts a high-quality recording. Intelligibility is excellent, with a suitable speech rate and well-balanced dynamics; minor volume consistency could be improved. Subjectively, the voice is bright and full with a subtle "surprise" effect, yielding a highly pleasant listening experience.

Speech Quality Comparison In Figure 12, the two samples differ markedly. Sample A is clearly superior in intelligibility and distortion control, while speech rate and dynamics are similar. Subjectively, A sounds crisp and bright with a livelier effect; B is mellow and resonant. The conclusion and alternatives emphasize that A's technical clarity drives the overall preference.

Figure 13 compares two recordings that are effectively on par across the core axes: intelligibility, dynamic range, and overall technical execution. Yet they diverge slightly across sub-dimensions: noise signature, timbre, and expressiveness. These micro-differences yield a modest preference for A, though both remain essentially comparable.

Speech Quality Improvement Suggestion For actionable remediation, Figure 14 provides targeted recommendations for a low-quality recording. The proposed improvements emphasize cleaning the signal, normalizing pacing, and strengthening expressiveness to raise intelligibility and listener appeal. The detailed and alternative descriptions translate these goals into concrete, implementation-ready guidance.

Figure 15 targets high-quality recordings where only subtle refinements are needed. It recommends light enhancements to emotional expressiveness, gentle tonal rebalancing, and fine-tuning of articulation, with alternative phrasings provided for direct implementation. The figure also includes a standardized "no improvement needed" case for samples that already perform optimally across all dimensions, ensuring consistent annotation when refinement is unnecessary.

A.3 Detailed Annotation Protocol

Annotation schema. To support interpretability and task diversity, we design a unified annotation framework covering three aspects: overall rating, objective production quality, and subjective content enjoyment. These aspects are operationalized into eight subdimensions, including intelligibility, distortion, speech rate, dynamic range, tone balance, emotional impact, artistic expression, and subjective experience. Depending on the evaluation task, annotators provide either five-point scalar ratings or pairwise comparisons. The complete scoring criteria are reported in Tables 8–10.

Beyond the primary ratings, we collect categorical metadata for distortion type, emotion type, and speaker gender. We additionally include five open-ended fields: distortion duration, distortion severity, perceptual description, speaker age, and speaking tone, as summarized in Table 11 and Table 12.

Annotators, compensation, and quality control. All annotators had relevant language backgrounds, with English proficiency at least at the College English Test Band 4 (CET-4) level (including TEM-8 holders and overseas students), and $\geq$1 year of annotation experience. A total of 56 professional annotators participated in this study, including 11 Chinese, 13 Japanese, 12 French, and 20 English annotators. The annotator pool comprised 40 female and 16 male annotators, with a relatively consistent age distribution across languages (mean age: 25.3–25.9 years; overall range: 23–31 years). For each target language, annotations were performed by native speakers or professionally trained annotators with domain expertise; among them, professionals in English, Japanese, and French accounted for 50% of the workforce.

All annotators completed training and passed a qualification test before participating in formal labeling. Annotators were fairly compensated: for Chinese, Japanese, and English speech, the payment was 0.62 USD per single-sample annotation and 0.83 USD per comparison task; for French speech, the rates were 0.69 USD and 0.90 USD, respectively. We employed an end-to-end platform-based pipeline with 100% quality inspection; ambiguous cases were escalated to a third annotator for adjudication. We further integrated process-level quality-control tools, including lexicon checks for structured attributes (e.g., speaker gender and age) and consistency checks between distractor-item scores and selected distractor categories, to reduce annotation errors.

1. Overall Quality
A is better than B
2. Production quality
Intelligibility: A is better than B
Distortion: A is better than B

A	*Type: None*	*Duration: None*
	Degree: None	*Description: None*
B	*Type: timbre & quality*	*Duration: 0 -3.7s (throughout the recording)*
	Degree: extremely slight	*Description: Extremely slight rough.*

Speech Rate: A and B are similar in this aspect
Dynamic Range: A and B are similar in this aspect
3. Content Enjoyment
Emotional Impact: A and B are similar in this aspect ***A*** *Type: Surprise* ***B*** *Type: Neutral*
Artistic Expression: A and B are similar in this aspect
Subjective Experience: A and B are similar in this aspect

A	*Gender: female*	*Age: middle-aged*	*Tone Description: crisp and bright*
B	*Gender: male*	*Age: middle-aged*	*Tone Description: mellow and full*

4. Detailed Description
Overall, Sample A demonstrates superior quality compared to Sample B. Objectively, Sample A excels in intelligibility, with clearer speech and no reported distortions, while Sample B exhibits an extremely slight timbre and quality distortion (0–3.7 s), perceived as a faint roughness. Both samples share similar speech rates and dynamic ranges. Subjectively, they perform comparably in emotional impact (Sample A conveys surprise, B is neutral), artistic expression, and listener experience. Sample A features a crisp, bright female middle-aged voice, whereas Sample B has a mellow, full male middle-aged tone. In summary, Sample A's technical clarity and neutral distortions make it the better choice, though both are expressive and engaging.
5. Alternative Description
1) *In this comparison, Sample A emerges as the stronger performer. From an objective standpoint, Sample A delivers clearer speech with no distortions, whereas Sample B has a subtle roughness in timbre (0–3.7s). Both samples maintain similar speech rates and dynamic ranges. Subjectively, their emotional impact (Sample A: surprise; Sample B: neutral), artistic expression, and listener appeal are closely matched. Vocally, Sample A presents a bright, articulate female voice, while Sample B offers a warm, resonant male tone. While both are compelling, Sample A's distortion-free clarity gives it the edge.*

Figure 12: Comparison example with a large quality gap between samples.

1. Overall Quality
A and B are similar in this aspect
2. Production quality
Intelligibility: A and B are similar in this aspect
Distortion: A and B are similar in this aspect

A	*Type: artifacts*	*Duration: 0 -5s(throughout)*
	Degree: slight	*Description: The sound is a little distorted.*
B	*Type: background noise*	*Duration: 0 -5s(throughout)*
	Degree: slight	*Description: The sound has a slight background noise.*

Speech Rate: A is better than B
Dynamic Range: A and B are similar in this aspect
3. Content Enjoyment
Emotional Impact: A is better than B ***A*** *Type: Neutral* ***B*** *Type: Neutral*
Artistic Expression: A and B are similar in this aspect
Subjective Experience: A is better than B

A	*Gender: female*	*Age: 25s*	*Tone Description: clear and light.*
B	*Gender: female*	*Age: 22s*	*Tone Description: low and hollow.*

4. Detailed Description
Overall, Samples A and B demonstrate comparable quality, though subtle differences exist in technical and subjective aspects. Objectively, both samples exhibit similar intelligibility and dynamic range, but Sample A has a slight artifact distortion (0-5 s, slight degree), slightly affecting clarity, while Sample B has faint background noise (0-5 s, slight degree). Sample A's speech rate is more appropriate, enhancing comprehension. Subjectively, Sample A outperforms in emotional impact (neutral but more engaging) and subjective experience, likely due to its clearer, lighter tone (female, 25s) compared to Sample B's low, hollow tone (female, 22s). Artistic expression is similar in both. In summary, while technically close, Sample A offers a marginally better listening experience.
5. Alternative Description
1) *While Samples A and B share many technical similarities, subtle distinctions influence their overall quality. Both exhibit strong intelligibility and dynamic range, but Sample A has minimal artifact distortion (0-5 s), slightly reducing clarity, whereas Sample B contains faint background noise (0-5 s). Sample A's well-paced speech enhances understanding. On a subjective level, Sample A delivers better emotional impact—its neutral yet engaging tone (female, 25s) contrasts with Sample B's deeper, less resonant tone (female, 22s). Artistic expression is comparable. In conclusion, though closely matched, Sample A edges out with a more pleasant listening experience.*
2) *A comparative analysis reveals that Samples A and B are closely matched in quality, differing only in subtle ways. Objectively, intelligibility and dynamic range are equivalent, though Sample A has faint artifact distortion (0-5 s), and Sample B contains minimal background noise (0-5 s). Sample A's optimal speech rate improves listener comprehension. Subjectively, Sample A excels in emotional engagement—its clear, bright tone (female, 25s) contrasts with Sample B's deeper, less expressive tone (female, 22s). Artistic execution is comparable. In summary, while both perform well, Sample A offers a marginally better experience.*

Figure 13: Comparison example with a small quality gap between samples.

1. Improvement Dimensions
expressiveness;speech rate
2. Detailed Description
Adjust the speech rate to a more natural pace to enhance clarity. Improve expressiveness by incorporating emotional variation and tonal dynamics.
3. Alternative Description
1) *Moderate the speech rate for better naturalness and comprehension. Enhance expressiveness through varied intonation and emotional inflection. Minimize distortion to achieve cleaner, more lifelike speech output.*
2) *Fine-tune the speech rate to improve naturalness and listener comfort. Incorporate more dynamic vocal expression to enhance engagement. Eliminate distortion artifacts for a cleaner and more polished output.*
3) *Adjust the speech rate to a more natural and listener-friendly tempo. Enhance vocal expressiveness through varied pitch and emotional modulation. Reduce distortion to ensure smoother and more intelligible speech.*

1. Improvement Dimensions
expressiveness;emotional impact;noise reduction
2. Detailed Description
Enhance emotional expressiveness to make the delivery more engaging. Reduce background noise for cleaner audio quality. Adjust speech rate to a more natural pace while maintaining clarity.
3. Alternative Description
1) *Incorporate greater emotional variation to enhance engagement. Eliminate background noise for improved clarity. Slightly increase the speech rate for a more natural flow without sacrificing intelligibility.*
2) *Strengthen emotional expression to make the speech more captivating. Clean up background noise for optimal audio clarity. Adjust the speech rate to a more natural tempo while preserving intelligibility.*

1. Improvement Dimensions
dynamic range;intelligibility
2. Detailed Description
Enhance intelligibility by reducing voice drops and missing segments. Improve dynamic range consistency for smoother volume transitions.
3. Alternative Description
1) *Minimize voice dropouts and missing segments to enhance intelligibility. Stabilize dynamic range for more uniform volume transitions. Refine distortion management to reduce speech interruptions.*
2) *Address voice gaps and dropouts to improve intelligibility. Adjust dynamic range for steadier volume levels. Optimize distortion control to ensure smoother speech continuity.*
3) *Reduce speech dropouts and missing segments for better intelligibility. Fine-tune dynamic range to achieve smoother audio transitions. Enhance distortion mitigation to prevent speech breaks.*

Figure 14: Suggestion example for improving a low-quality speech sample.

1. Improvement Dimensions
emotional impact;frequency balance;intelligibility
2. Detailed Description
Enhance emotional expressiveness to engage listeners. Adjust frequency balance to improve voice clarity. Optimize intelligibility for better comprehension.
3. Alternative Description
1) *Increase emotional variation to enhance engagement. Adjust tonal balance to refine voice clarity. Boost intelligibility for better listener comprehension.*
2) *Incorporate more expressive vocal modulation. Optimize frequency distribution for improved clarity. Strengthen intelligibility to aid listener perception.*
3) *Add more dynamic vocal expression to engage the audience. Modify frequency settings to enhance voice clarity. Increase intelligibility for smoother listener comprehension.*

1. Improvement Dimensions
Expressiveness
2. Detailed Description
To enhance expressiveness: Incorporate subtle vocal inflections to add emotional depth adjust pacing slightly to emphasize key phrases for greater engagement.
3. Alternative Description
1) *To improve expressiveness: Introduce gentle vocal modulations to enrich emotional tone. Fine-tune pacing to highlight important phrases for better engagement. Vary pitch subtly to prevent a monotonous effect.*
2) *To enhance expressiveness: Employ subtle pitch variations to inject emotional depth. Adjust rhythm slightly to underscore key moments. Experiment with tonal shifts to prevent a flat presentation.*
3) *To refine expressiveness: Integrate delicate vocal inflections for emotional texture. Optimize pacing to draw attention to pivotal phrases. Vary tone slightly to avoid a monotonous effect.*

1. Improvement Dimensions
None
2. Detailed Description
No specific improvements are needed as the speech performs optimally across all evaluated dimensions.
3. Alternative Description
1) *No immediate enhancements are required, as the speech excels in every assessed aspect.*

Figure 15: Suggestion example for refining a high-quality speech sample.

Sub-dimensions	Annotation Criteria	
Overall Quality	**1 point (Extremely Poor)**	Basic standard for lowest quality.
	2 points (Poor)	Below average quality with obvious defects.
	3 points (Average)	Moderate quality meeting basic listening needs.
	4 points (Good)	High quality with minor imperfections.
	5 points (Excellent)	Top-tier quality with no perceivable defects.

Table 8: Annotation Protocol for Overall Quality.

Sub-dimensions	Annotation Criteria	
Intelligibility	**1 point (Extremely Poor)**	Pronunciation is vague. Most words or sentences are difficult to recognize or easily misheard. Listeners must concentrate very hard or replay multiple times to understand.
	2 points (Poor)	General idea is barely understandable with many ambiguities. Speech becomes difficult to hear even in slightly noisy environments.
	3 points (Average)	Most words and sentences are understandable. Only a few words or complex sentences require extra concentration.
	4 points (Good)	Speech is clear and easy to understand even in general background noise. Only occasional minor vagueness appears.
	5 points (Excellent)	Speech is extremely clear and almost all content is understood immediately. Pronunciation is accurate and expression is detailed.
Distortion	**1 point (Extremely Poor)**	Obvious and continuous noise or heavy synthetic artifacts, which seriously affect understanding.
	2 points (Poor)	Audible background noise or buzzing is present, and frequent noise, splicing traces, or jitter can be heard.
	3 points (Average)	Minor noise or occasional synthetic traces, no serious listening interference.
	4 points (Good)	Very slight noise or synthetic feeling, detectable only in rare moments.
	5 points (Excellent)	The background is clean or synthetic traces are almost undetectable. The overall sound is natural and realistic.
Speech Rate	**Slow**	Obviously lower than average speed.
	Slightly Slow	Slightly lower than common speed.
	Appropriate	Matches normal speed, easy for most people to understand.
	Slightly Fast	Slightly higher than conventional speed.
	Fast	Significantly higher than average spoken speed.
Dynamic Range	**1 point (Extremely Poor)**	Sharp volume fluctuations, seriously affecting understanding.
	2 points (Poor)	Frequent unreasonable volume changes.
	3 points (Average)	Acceptable overall volume, but sudden changes in some paragraphs.
	4 points (Good)	Stable volume most of the time, occasional minor fluctuations.
	5 points (Excellent)	Stable and balanced volume.

Table 9: Annotation Protocol for Production Quality.

B Model

B.1 Unified Prompt Format for SQ-LLM

To support a unified evaluation across four distinct tasks, quality assessment, comparison, suggestion, and deepfake detection, we design standardized prompt templates to guide SQ-LLM's behavior. These templates ensure consistent instruction formats while accommodating the unique requirements of each task. The following presents the representative prompts.

Assessment Prompt

```
[Audio: <AUDIO_PLACEHOLDER>]
Please evaluate the overall quality of
    this speech.
```

Comparison Prompt

```
Sample A: [Audio: <AUDIO_A_PLACEHOLDER>]
Sample B: [Audio: <AUDIO_B_PLACEHOLDER>]
Compare the speech quality of Sample A
    and Sample B. Please provide a
    comprehensive assessment.
```

Suggestion Prompt

```
[Audio: <AUDIO_PLACEHOLDER>]
Please suggest specific aspects for
    improvement to enhance the overall
    quality of this speech.
```

Deepfake Detection Prompt

```
[Audio: <AUDIO_PLACEHOLDER>]
Determine if this speech is real or a
    deepfake.
```

Sub-dimensions	Annotation Criteria	
Emotional Impact	**1 point (Extremely Poor)**	Strong machine-like feeling, no emotional fluctuation.
	2 points (Poor)	Slight emotion, but still plain.
	3 points (Average)	Adds emotional color to some paragraphs, but not sufficient.
	4 points (Good)	Natural and smooth emotions, obvious emotional resonance.
	5 points (Excellent)	Perfectly matches text atmosphere, deeply touches listeners with strong appeal.
Artistic Expression	**1 point (Extremely Poor)**	Speech sounds mechanical and disconnected from the context. There is no sense of rhythm or layered meaning.
	2 points (Poor)	The speech shows only basic text expression and lacks contextual processing.
	3 points (Average)	Limited rhythm/stress variation, beginning to express the text's mood.
	4 points (Good)	Integrates rhythm, context, and meaning well, with artistic appeal.
	5 points (Excellent)	Rhythm, context, and nuance perfectly align, creating strong artistic tension.
Subjective Experience	**1 point (Extremely Poor)**	Uncomfortable or boring, cannot be endured for a long time.
	2 points (Poor)	Barely audible, lacks attractiveness and comfort.
	3 points (Average)	Acceptable, neither boring nor pleasant.
	4 points (Good)	Comfortable and natural, listeners are willing to continue.
	5 points (Excellent)	Extremely pleasant, leaves a deep and happy impression.

Table 10: Annotation Protocol for Content Enjoyment.

Sub-dimensions	Annotation Criteria	
Distortion Type	**Background Noise**	Continuous or intermittent environmental noise.
	Jitter	Tremor or slight pitch flutter.
	Drop / Missing	Dropped frames or truncation, clipped words, brief silences or jumps.
	Timbre & Quality	Unnatural tone.
	Artifacts	Non-natural sounds.
Emotion Type	**Happiness**	Higher pitch, stronger energy, slightly faster tempo, bright tone.
	Sadness	Lower pitch, softer loudness, slower rate, subdued affect.
	Anger	Strong energy, laryngeal tension, increased sharpness or bursts.
	Fear	More tremor, rising terminals, unstable delivery.
	Disgust	Nasal timbre, curt phrasing, flattened intonation.
	Surprise	Sudden increase in pitch and loudness, with clear and abrupt emphasis.
	Neutral	Pitch, energy, and rhythm remain stable, showing minimal emotional cues.
Speaker Gender	**Male**	Perceived masculine timbre.
	Female	Perceived feminine timbre.
	Unknown	Indeterminate.

Table 11: Annotation Protocol for three types of categorical metadata.

B.2 Chain-of-Thought Example for Speech Quality Assessment

To better illustrate the reasoning procedure , we provide an example of Chain-of-Thought for the speech quality assessment task. In this case, the model is instructed to evaluate a single speech clip by assigning scores to eight perceptual dimensions, accompanied by brief justifications. The <think> block captures the internal reasoning process with structured judgments, while the <answer> block summarizes the evaluation in fluent natural language.

<think>

For each dimension, give a score and a short explanation.

Overall Quality: 1/5

Intelligibility: 1/5

Distortion: 1/5 (timbre and quality;artifacts)

Speech Rate: suitable

Dynamic Range: 2/5

Emotional Impact: 1/5 (Neutral)

Artistic Expression: 1/5

Subjective Experience: 2/5 (male, middle-aged)

</think>

<answer>

The speech suffers from extremely poor overall quality. Objectively, intelligibility is severely compromised, with timbre and quality distortions lasting 0-3 seconds, making the vocal track unclear.

Sub-dimensions	Annotation Criteria
Distortion Severity	Explanation: The perceived intensity or impact level of the distortion. Example: *"Slight." / "Noticeable." / "Severe."*
Perceptual Description	Explanation: Short free-form notes describing the perceptual impression. Example: *"The voice has poor timbre."*
Speaker Age	Explanation: Estimated or perceived age range of the speaker. Example: *"Thirties."*
Speaking Tone	Explanation: Brief description of tone style. Example: *"The voice sounds mellow but not clear enough."*

Table 12: Annotation Protocol for five open-ended fields.

The speech rate is suitable, but the dynamic range lacks smoothness. Subjectively, the neutral emotional impact and weak artistic expression fail to engage. The middle-aged male voice sounds hollow and thick, offering an unremarkable listening experience. In summary, the speech is marred by severe distortion, weak expressiveness, and a dull delivery.

</answer>

B.3 Chain-of-Thought Example for Speech Quality Comparison

The following is an example of Chain-of-Thought reasoning used in the speech quality comparison task. The `<think>` section captures the intermediate thought process, where the model compares two speech clips dimension by dimension, offering judgments (e.g., "A and B are similar" or "B is better than A"). This structured comparison allows the model to form a reasoned basis for its conclusion. The `<answer>` section then synthesizes these observations into a natural language summary that reflects the dimension-wise analysis.

<think>

Compare the two audio samples across different quality dimensions.

Overall Quality: A and B are similar
Intelligibility: A and B are similar
Distortion: A and B are similar
Speech Rate: A and B are similar
Dynamic Range: A and B are similar
Emotional Impact: B is better than A
Artistic Expression: B is better than A
Subjective Experience: A and B are similar

</think>

<answer>

Overall, both speech samples (A and B) demonstrate comparable quality. Objectively, they are equally intelligible, with similar speech rates, dynamic range, and minimal distortion, though B has slightly better timbre. From a subjective perspective, B excels in emotional impact and artistic expression, delivering a more engaging performance, while the listening experience remains similar for both. In summary, B stands out slightly due to its expressive strengths, though both are technically balanced.

</answer>

B.4 Reward Criteria

The four reward criteria are applied to three task formats (SingleEval, Suggest, CompareEval). Each evaluator outputs an integer in $[0, 10]$.

[Helpfulness]

Assessment **0–2** Barely helpful or misleading; advice doesn't fit the context or could make results worse. **2–4** Mostly generic advice (e.g., "improve clarity/intonation") with little operational detail; not readily applicable. **4–6** Partly useful: identifies issues (distortion, limited emotional engagement) but lacks precise locations/parameters/practice details. **6–8** Actionable plan with minor gaps; mostly tailored; at least one check or fallback is suggested. **8–10** Directly usable and targeted: gives specific actions (e.g., reduce minor distortion at 0–0.5s, 1.3–2.1s, 3.1–3.8s via denoise/retake; adjust pacing; widen pitch/energy for expressiveness).

Comparison **0–2** No decision, or misleading guidance that would pick the worse sample. **2–4** Vague preference ("A sounds nicer") with minimal justification; weak for decision-making. **4–6** Gives a decision but reasons are partly generic or mixed; reader must infer the main takeaway. **6–8** Decision is usable with specific reasons, but lacks either prioritization, trade-offs, or implications. **8–10** Clear decision (A/B or justified tie) with ranked reasons that map to the brief's goals; highlights where and

Speech Quality Assessment Prompt

Please give a detailed evaluation of the following speech sample:
<audio>
[Evaluation Guidelines Start]
- Overall Quality: the general impression and performance.
- Intelligibility: clarity of pronunciation and ease of understanding.
- Distortion: presence and severity of noise, artifacts, or synthetic glitches, and their types.
- Speech Rate: whether the pace is slow, appropriate, or fast, and how it affects comprehension.
- Dynamic Range: stability and balance of volume throughout the sample.
- Emotional Impact: the degree to which emotions are expressed and conveyed naturally.
- Artistic Expression: rhythm, emphasis, and prosody in reflecting meaning and context.
- Subjective Experience: the overall listening experience, including comfort, naturalness, and listener appeal.
[Evaluation Guidelines End]

[Output Requirements]
- Provide a natural, flowing evaluation rather than a checklist or bullet points.
- Integrate the above aspects into a coherent description.
- Highlight both technical precision and subjective listening experience.
[Output Requirements End]

Figure 16: Speech Quality Assessment Prompt.

why one sample is better (e.g., "A cleaner in 0–3.2 s; more balanced dynamics in verse lines"), plus what that implies (intelligibility, listening comfort). If close, states trade-offs and when to prefer each.

Suggestion **0–2** Misleading or counterproductive suggestions; could worsen quality. **2–4** Mostly platitudes ("be more expressive"); little practical value without extra work. **4–6** Partly helpful: directions are generic or incomplete; user must figure out key details to execute. **6–8** Solid and usable, but misses one element (priority, verification, or justification). **8–10** Directly improves the piece: prioritized, actionable steps with clear how-to.

[Relevance]

Assessment **0–2** Largely off-topic or at odds with instructions/context. **2–4** Weak focus: substantial digressions or misses a key element from the context (e.g., ignores the provided distortion intervals). **4–6** Mostly relevant but includes nontrivial extraneous content or overlooks a minor stated point (e.g., underemphasizes dynamic range being stable). **6–8** Generally focused with brief, still-useful tangents; constraints are respected with only trivial deviations. **8–10** Stays tightly on clarity/intelligibility, pacing, distortion with timestamps, dynamic range stability, tone/timbre, expressiveness/emotion; no irrelevant digressions; follows the reference.

Comparison **0–2** Off-topic or contradicts the comparison instruction (e.g., evaluates only A). **2–4** Weak alignment: talks about single-sample qualities without comparing, or ignores key axes. **4–6** Some relevant content, but includes nontrivial digressions or underplays a required axis. **6–8** Mostly on-topic with brief tangents that still inform the comparison. **8–10** Tightly focused on A vs B across the specified axes (clarity/intelligibility, pacing, distortion with timestamps, dynamic range, tone/expressiveness); no unrelated commentary; follows output constraints (e.g., compare rather than re-summarize background).

Suggestion **0–2** Largely off-topic or contradicts instructions. **2–4** Weak alignment; misses key issues mentioned/implied in context. **4–6** Some relevance, but includes nontrivial digressions or overlooks a stated focus. **6–8** Mostly on-topic, with brief tangents that still help. **8–10** Tightly aligned to the observed issues (e.g., rate, emotional nuance, expressiveness) and any provided constraints; no off-topic or format violations (stays in "suggest" mode, not scoring/explaining).

[Accuracy]

Assessment **0–2** Predominantly incorrect or fabricated; contradicts core facts; unsafe/misleading. **2–4** Key error or contradiction; parts may remain usable. **4–6** Some incorrect details or unsupported claims, but the main idea is salvageable. **6–8** Minor imprecision that doesn't change the outcome (e.g., "slight" vs. "minor"); overall alignment remains correct. **8–10** Fully consistent with the ref-

Speech Quality Comparison Prompt

Please compare the following two speech samples, referred to as Sample A and Sample B, and provide a detailed evaluation.
Sample A: ***<audio>*** *Sample B:* ***<audio>***
[Evaluation Guidelines Start]
- Overall Quality: the general impression and performance.
- Intelligibility: clarity of pronunciation and ease of understanding.
- Distortion: presence and severity of noise, artifacts, or synthetic glitches, and their types.
- Speech Rate: whether the pace is slow, appropriate, or fast, and how it affects comprehension.
- Dynamic Range: stability and balance of volume throughout the sample.
- Emotional Impact: the degree to which emotions are expressed and conveyed naturally.
- Artistic Expression: rhythm, emphasis, and prosody in reflecting meaning and context.
- Subjective Experience: the overall listening experience, including comfort, naturalness, and listener appeal.
[Evaluation Guidelines End]

[Output Requirements]
- Provide a natural, flowing comparative assessment rather than a checklist.
- Highlight differences between Sample A and Sample B, noting strengths and weaknesses.
- Conclude with overall judgment of which sample is superior and why.
[Output Requirements End]

Figure 17: Speech Quality Comparison Prompt.

erence: correctly notes every details; no fabricated metrics or contradictions.

Comparison **0–2** Largely inaccurate/fabricated; contradicts core facts or uses faulty audio concepts. **2–4** Major misread of evidence leading to doubtful verdict. **4–6** Some incorrect or unsubstantiated claims, but the main verdict is still plausible. **6–8** Minor imprecision that doesn't change the outcome; overall alignment intact. **8–10** Consistent with context: correctly attributes issues. No fabricated timestamps/metrics; no internal contradictions; terminology used correctly.

Suggestion **0–2** Fabricated issues, unsafe/incorrect advice, or contradictions with context. **2–4** Major mismatch or questionable technique likely to underperform. **4–6** Some mismatches yet partially useful. **6–8** Minor imprecision but generally appropriate; outcomes remain valid. **8–10** Suggestions match the evidence (no invented problems), choose appropriate, safe remedies, and avoid technical/myth errors.

[Level of Detail]

Assessment **0–2** Little to no actionable detail; mostly restates the prompt or uses vague descriptors. **2–4** High-level outline; major gaps. **4–6** Main points present but lacks important parameters; some ambiguity remains. **6–8** Strong detail with minor omissions (eg. missing one timestamp or a small validation step); still largely reproducible. **8–10** Reproducible detail: cites exact timestamps for issues; separates objective vs. subjective aspects.

Comparison **0–2** Vague restatements; no actionable or verifiable detail. **2–4** High-level bullets with little localization or evidence; hard to verify. **4–6** Names main areas but have some ambiguity. **6–8** Specific and useful but missing one element. **8–10** Concrete evidence: cites timestamps/segments for differences; separates objective from subjective.

Suggestion **0–2** Vague or filler; no operational detail. **2–4** High-level bullets only; little to reproduce in practice. **4–6** Main areas named, but lacks crucial parameters/examples; ambiguity remains. **6–8** Detailed but missing one element. **8–10** Specific targets, concrete how-to steps and actionables.

C Experimental Setup

C.1 Data splitting details

In the Deepfake Speech Detection task, the partitioning method of the dataset is crucial for effective model training and evaluation. The ASVspoof challenge was the first to establish the common practice of dataset construction for this task, which has since been widely adopted in subsequent research. Following this established practice, we partition the dataset for the Deepfake Speech Detection task accordingly. The dataset comprises

Speech Quality Improvement Prompt

Please provide constructive improvement suggestions for the following speech sample:
<audio>
[Improvement Guidelines Start]
- Overall Quality: ways to enhance the general impression and performance.
- Intelligibility: how to make pronunciation clearer and understanding easier.
- Distortion: how to reduce noise, artifacts, or glitches and improve naturalness.
- Speech Rate: adjustments to pacing that could improve comprehension.
- Dynamic Range: refinements to volume stability and balance.
- Emotional Impact: methods to convey emotions more naturally and effectively.
- Artistic Expression: how rhythm, emphasis, and prosody could better reflect meaning and context.
- Subjective Experience: strategies to improve overall listening comfort, naturalness, and appeal.
[Improvement Guidelines End]

[Output Requirements]
- Provide a natural, flowing set of suggestions rather than a checklist.
- Integrate the above aspects into a coherent description.
- Emphasize both technical precision and practical advice.
[Output Requirements End]

Figure 18: Speech Quality Improvement Prompt.

Deepfake Speech Detection Prompt

Please determine whether the following speech sample is real or synthetic:
<audio>
[Decision Guidelines Start]
- Real: the speech is naturally produced by a human.
- Fake: the speech is generated or synthesized by a machine.
[Decision Guidelines End]

[Output Requirements]
- Output only one word: "real" or "fake".
[Output Requirements End]

Figure 19: Deepfake Speech Detection Prompt.

a total of 32,207 samples, divided into training, validation, and test sets in proportions of 20.5%, 17.8%, and 61.7%, respectively. To better simulate real-world applications, the proportion of real samples gradually decreases across the subsets, with rates of 33.1%, 22.7%, and 20.3%, ensuring the adequacy and diversity of the training set and aligning with the challenging Zero-shot evaluation scenario, which is a key focus of this task. The specific partitioning strategies are as follows: for certain fake sample sources, a 5:5:0 train-validation split is used to assess the model's ability to fit known spoofing distributions, while a portion of fake sample sources is fully assigned to the test set with a 0:0:10 ratio, evaluating the model's zero-shot generalization capability in the context of previously unseen spoofing scenarios. Furthermore, due to significant variations in language and data scale across sample sources, an imbalance in training data arises. To mitigate this, we adopt four different partitioning ratios (2:2:6, 4:2:4, 6:2:2, and 1:1:8) for certain sample sources to ensure sufficient training samples while maintaining balanced distributions across the subsets. Finally, we verify that the resulting partitions are mutually exclusive across tasks, ensuring that no data overlap or leakage occurs.

C.2 LLM-based Metrics

To evaluate the outputs produced by the three baseline models in the ms-swift framework, we employ the DeepSeek to score model inferences on the Speech Quality Assessment and Speech Quality Comparison prompt tasks, and quantify synthesized speech along multiple dimensions.

For the SQA task, we first assess the overall accuracy and quality of the generated paragraph-level descriptions. Concretely, human-annotated speech evaluation texts are used as ground-truth and, together with each baseline model's generated output, are submitted to the DeepSeek API. We require DeepSeek to perform a comparative evaluation along the dimensions of Helpfulness, Relevance, Accuracy, and Level of Detail, returning a 10-point score and a textual justification for each baseline output. The prompt used for the paragraph-level assessment is given in Figure 20.

Next, we extract fine-grained dimensions from

Speech Quality Assessment API Prompt

You are an evaluator assessing the quality of an AI-generated response based on a provided human-labeled context and a human-written ground-truth response.
Context: *{context}*

Question: *{question}*

Answer_1 (Ground Truth Answer): *{answer_1}*

Answer_2 (Model-generated Answer): *{answer_2}*

Task:
Compare Answer_2 with Answer_1 using the provided Context and Question. Focus on the following aspects:
- Helpfulness: Does the answer provide useful and relevant information?
- Relevance: Does it stay on topic and align with the context?
- Accuracy: Does it reflect the facts stated in the context?
- Level of Detail: Is the response thorough and precise?

Then, do the following:
1. Provide a brief explanation comparing Answer_2 with Answer_1.
2. Give a numerical score (0 to 10) for Answer_2 based on its quality relative to Answer_1.
3. Do not score Answer_1. Use it only as the gold reference.

Output Format:
- Explanation: <Your reasoning here>
- Score: <A number from 0 to 10>

Figure 20: Speech Quality Assessment API Prompt.

Speech Quality Assessment Score API Prompt

Task:
Evaluate the following synthesized speech based on the eight dimensions below. For each dimension, provide a score from 1 to 5 (1 = worst, 5 = best), and classify "Speech Rate" as one of the following: slow, slightly slow, suitable, slightly fast, fast.

Output Format:
- Overall Quality: [score]
- Intelligibility: [score] Distortion: [score]
- Speech Rate: [classification]
- Dynamic Range: [score]
- Emotional Impact: [score]
- Artistic Expression: [score]
- Subjective Experience: [score]

Description: {description}

Figure 21: Speech Quality Assessment Score API Prompt.

the paragraph-level descriptions and perform an accuracy assessment per dimension. As shown in Figure 21, each baseline model's natural-language SQA output is sent to DeepSeek to be quantitatively mapped onto eight quality dimensions: Overall Quality, Intelligibility, Distortion, Speech Rate, Dynamic Range, Emotional Impact, Artistic Expression, and Subjective Experience. Speech Rate is treated as a categorical label with five levels: Slow, Slightly Slow, Appropriate, Slightly Fast, and Fast, while the other seven dimensions are mapped to a five-point numeric scale. The resulting quantitative scores are then compared against human ground-truth annotations to compute accuracy statistics.

For the SQC task, we follow an analogous procedure: DeepSeek is used to evaluate model outputs both at the paragraph level and across the same eight fine-grained dimensions, producing comparable numeric scores and textual explanations. The SQC prompt used to elicit these comparative judgments is shown in Figure 22, and the scores thus obtained are evaluated against ground-truth annotations to enable direct comparison across models.

C.3 Deepfake Detection Metrics

The ACC metric applies to all LLM outputs, regardless of whether the token at the expected position is semantically valid. A prediction is counted as correct only when the generated token matches the ground-truth response among the key responses ("Fake" and "Real"). In contrast, the EER and minDCF are evaluated only on semantically valid outputs. Since LLMs do not provide explicit confi-

Speech Quality Comparison API Prompt

You are an evaluator assessing the quality of an AI-generated response based on a provided human-labeled context and a human-written ground-truth response.
Context: *{context}*

Question: *{question}*

Answer_1 (Ground Truth Answer): *{answer_1}*

Answer_2 (Model-generated Answer): *{answer_2}*

Task:
1. Extract the concluding judgments from both Answer_1 and Answer_2 for the following aspects:
- Overall Quality: Which answer (A or B) is better, or are they the same?
- Intelligibility: Which answer has better intelligibility, or are they the same?
- Distortion: Which answer has less distortion, or are they the same?
- Speech Rate: Which answer has a better speech rate, or are they the same?
- Dynamic Range: Which answer shows better dynamic range, or are they the same?
- Emotional Impact: Which answer has better emotional impact, or are they the same?
- Artistic Expression: Which answer has better artistic expression, or are they the same?
- Subjective Experience: Which answer provides a better subjective experience, or are they the same?

2. For each aspect, compare the extracted conclusions:
- If the conclusions for an aspect are the same (e.g., both say Answer_1 is better), mark them as "consistent."
- If the conclusions for an aspect differ, or if sentence 2 contains no relevant content, mark them as "inconsistent."

3. Evaluate the overall quality of Answer_2 based on the following general aspects:
- Helpfulness: Does the answer provide useful and relevant information?
- Relevance: Does it stay on topic and align with the context?
- Accuracy: Does it reflect the facts stated in the context?
- Level of Detail: Is the response thorough and precise?

4. Provide a brief explanation of your reasoning, comparing Answer_2 with Answer_1. After that:
- Provide a simple numerical score (0 to 10) for Answer_2 based on its quality relative to Answer_1.
- For each aspect (Overall Quality, Intelligibility, Distortion, Speech Rate, Dynamic Range, Emotional Impact, Artistic Expression, and Subjective Experience), indicate whether the conclusions are "consistent" or "inconsistent."
- Display the extracted conclusion for each aspect from both Answer_1 and Answer_2.

Important Output Formatting Guidelines:
- All eight dimensions must be extracted.
- Use the exact phrasing for each field. For example, Conclusion Consistency for Overall Quality: Consistent must appear exactly as shown.
- There should be no extra words, punctuation, or deviations from this format.

Output Format:
Explanation: *<Your simple reasoning here>*
Score: *<A number from 0 to 10>*
Conclusion Consistency for XX: <Consistent/Inconsistent>
Extracted Conclusion from Answer_1 (XX): <The extracted conclusion sentence from Answer_1 for XX>
Extracted Conclusion from Answer_2 (XX): <The extracted conclusion sentence from Answer_2 for XX>

Figure 22: Speech Quality Comparison API Prompt.

dence scores, we approximate the posterior probabilities by applying a two-dimensional softmax over the logits of the key response tokens, following the method described in Gu et al. (2025). Note that for expert systems, the EER and minDCF are computed directly from the inference scores, and the ACC is subsequently obtained by binarizing these scores into categorical labels.

C.4 Baselines

We evaluate four speech-related tasks, Speech Quality Assessment, Speech Quality Comparison, Speech Quality Improvement Suggestion, and Deepfake Speech Detection by using three models from the ms-swift framework: Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2.5-Omni-7B (Xu et al., 2025), and MiDashengLM-7B (Dinkel et al., 2025).

Qwen2-Audio-7B-Instruct is a large-scale audio-language model that accepts various audio inputs and is optimized for instruction-following in both voice chat and audio analysis modes. Its audio encoder is based on the Whisper-large-v3 model.

Qwen2.5-Omni-7B is an end-to-end omni-multimodal model designed to perceive text, images, audio, and video. It uses a Thinker-Talker architecture to generate both text and streaming speech responses.

MiDashengLM-7B is an open-source model designed for efficient and comprehensive audio understanding by training on "general audio captions" that fuse speech, sound, and music information into a single textual representation.

These models have comparable parameter scales and all support mixed multimodal inputs consisting of multiple speech segments and text. During inference, we configure the models to generate textual outputs. The hyperparameters are set uniformly

MODEL	OVR	INT	DST	DYN	EMO	EXP	SUBJ	AVG PCC	SR ACC
Qwen2-Audio-7B	0.073	0.037	0.080	0.043	0.070	0.078	0.072	0.065	0.664
Qwen2.5-Omni-7B	0.070	0.078	0.070	0.019	0.057	0.050	0.106	0.064	**0.736**
MiDashengLM-7B	0.158	0.114	0.138	0.071	0.156	0.102	0.159	0.128	0.712
Qwen3-8B + Whisper	0.424	0.387	0.470	0.274	0.363	0.329	0.418	0.381	0.676
Qwen2.5 + Audiobox	0.457	0.423	0.585	0.209	0.376	0.348	0.420	0.403	0.644
Qwen3-4B + WavLM	0.422	0.387	0.460	0.207	0.314	0.334	0.378	0.357	0.665
FT Qwen2-Audio-7B	0.244	0.192	0.455	0.202	0.209	0.295	0.269	0.267	0.706
SQ-LLM (OURS)	**0.520**	**0.505**	**0.592**	**0.329**	**0.434**	**0.378**	**0.456**	**0.459**	0.731

Table 13: Detailed speech quality assessment across models: each dimension Pearson correlation coefficients (PCC) with human ratings and speech rate accuracy; abbreviations are OVR (overall quality), INT (intelligibility), DST (distortion), DYN (dynamics), EMO (emotional), EXP (expression), SUBJ (subjective). AVG PCC denotes the mean PCC over all dimensions, and SR ACC denotes speech rate accuracy.

with max_new_tokens = 2048 and max_batch_size = 64, while all other parameters remain at their default settings.

For the Deepfake Speech Detection task, three baseline systems, RawNet2 (Tak et al., 2021), AASIST (Jung et al., 2022), and AASIST2 (Zhang et al., 2024), are employed. RawNet2 and AASIST follow the official ASVspoof 5 Challenge implementations[4], while AASIST2 is reimplemented using the pretrained feature extractor *voidful/wav2vec2-xlsr-multilingual-56* from Hugging Face. All systems retain their original architectures, and the input features and training setups are standardized to ensure a fair comparison.

RawNet2 uses a learnable Sinc front-end (Ravanelli and Bengio, 2018) with residual stacks for end-to-end learning on raw waveforms, incorporating filter-wise feature map scaling to enhance discriminability. It has been widely adopted for deepfake speech detection and serves as an official baseline in the ASVspoof 2021 challenge. AASIST augments a raw-waveform encoder with a spectro-temporal heterogeneous graph attention mechanism, coupled with max graph operations and an extended readout scheme to model artifacts introduced by synthesis and conversion. AASIST2 integrates a wav2vec2 pretrained feature extractor (Babu et al., 2021) with Res2Net-based multi-scale blocks (Gao et al., 2019), and applies dynamic chunking with adaptive large-margin fine-tuning to improve robustness across durations, particularly for short utterances.

For reproducibility, all models are trained using the Adam optimizer (learning rate = 0.0001, batch size = 24) with a class-weighted cross-entropy loss (negative : positive = 7 : 3). Training is conducted for up to 50 epochs on a single NVIDIA RTX 4090 GPU, with early stopping based on validation EER and a patience of 5.

C.5 Prompt for Zero-Shot Model

For models like MiDashengLM-7B that are evaluated directly without fine-tuning, we carefully design the prompt templates shown below to maximize the model's performance while ensuring a fair comparison.

For the four speech quality assessment tasks, we developed dedicated prompt templates to guide the model in performing precise, context-aware evaluations.

Speech Quality Assessment focuses on analyzing a single speech sample in terms of clarity, naturalness, emotional expression, and overall listening experience, as illustrated by the prompt template in Figure 16.

Speech Quality Comparison compares two samples, highlighting differences in intelligibility, distortion, dynamic range, and artistic expression to judge which sample performs better. The corresponding prompt template is shown in Figure 17.

Speech Quality Improvement Suggestion provides actionable recommendations for enhancing a speech sample, addressing issues such as pacing, pronunciation, emotional conveyance, and technical quality, with the prompt presented in Figure 18.

Deepfake Speech Detection determines whether a speech sample is human-produced or synthetically generated, offering a binary real/fake judgment. The prompt diagram is depicted in Figure 19.

Each prompt outlines evaluation criteria and output requirements, and attaches the relevant audio file paths (two for the SQC task and one for the

[4] https://github.com/asvspoof-challenge/asvspoof5

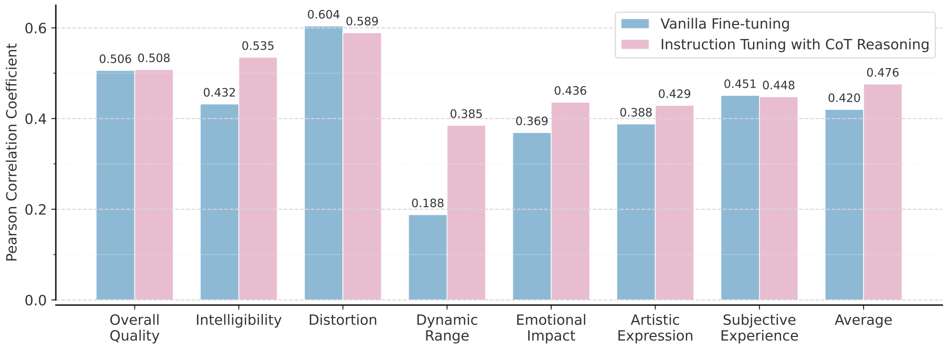


Figure 23: Pearson correlation coefficients of SQ-LLM predictions with human ratings across dimensions. The comparison is made between vanilla fine-tuning and instruction tuning with Chain-of-Thought reasoning.

Lang	Real	Open	Comm.	Overall
English	93.3	85.9	97.9	88.6
French	100.0	99.9	53.5	86.6
Japanese	99.9	100.0	95.9	99.1
Chinese	99.9	69.9	97.0	82.9

Table 14: Deepfake detection accuracy (%) by language, broken down by real speech and two fake-source types (open-source and commercial), together with the overall results.

others). Following standardization, we performed inference on these four tasks using three baseline models from the ms-swift framework.

D Additional Results

Speech Quality Assessment Table 13 presents dimension-wise Pearson correlations with human ratings and speech rate accuracy. In the direct-evaluation setting (Qwen2-Audio-7B, Qwen2.5-Omni-7B, MiDashengLM-7B), models generally struggle to capture speech quality; Qwen2.5-Omni-7B is notably strong on speech rate, achieving the best SR ACC without task-specific training, indicating that the pretrained model has a strong grasp of speech tempo. When we move to trained systems, custom-constructed pipelines, or fine-tuned audio LLMs raise several dimensions, yet the gains are uneven across categories. In contrast, SQ-LLM delivers uniform improvements in PCC on every dimension and yields the strongest average correlation with human raters. This pattern is consistent with the effect of CoT guidance combined with GRPO, which helps the model distinguish low-level degradations from higher-level prosodic and affective cues and reason about them explicitly.

Deepfake Detection Analysis Table 14 further breaks down deepfake detection accuracy by language and fake-source type. Overall performance is strong on Japanese, with consistently high accuracy across all categories. In contrast, the overall scores drop for Chinese and French, but the degradation is source-dependent: Chinese is mainly affected by open-source fakes , while French shows a sharp decline on commercial fakes. Notably, accuracy on real speech remains near-perfect across languages, suggesting low false-positive rates; the main weakness lies in generalizing to specific fake generation pipelines.

CoT Reasoning Analysis Figure 23 presents a comparative analysis of SQ-LLM models trained with and without Chain-of-Thought reasoning. The results show that while both models perform comparably on overall quality and distortion, CoT reasoning leads to noticeable improvements on more subjective and perceptual aspects. Specifically, dimensions like dynamic range, emotional impact, artistic expression, and dynamic range benefit the most from CoT supervision, indicating that step-by-step reasoning helps the model better capture subtle cues in expressiveness and listener experience.