SPEECHJUDGE: TOWARDS HUMAN-LEVEL JUDGMENT FOR SPEECH NATURALNESS

Xueyao Zhang[1*] **Chaoren Wang**[1*] **Huan Liao**[1] **Ziniu Li**[1] **Yuancheng Wang**[1]
Li Wang[1] **Dongya Jia**[2] **Yuanzhe Chen**[2] **Xiulin Li**[3] **Zhuo Chen**[2] **Zhizheng Wu**[1,4,5,6†]
[1]The Chinese University of Hong Kong, Shenzhen [2]ByteDance Seed [3]DataBaker Technology
[4]Shenzhen Loop Area Institute [5]City University of Macau [6]Amphion Technology Co., Ltd

ABSTRACT

Aligning large generative models with human feedback is a critical challenge. In speech synthesis, this is particularly pronounced due to the lack of a large-scale human preference dataset, which hinders the development of models that truly align with human perception. To address this, we introduce ***SpeechJudge***, a comprehensive suite comprising a dataset, a benchmark, and a reward model centered on *naturalness*—one of the most fundamental subjective metrics for speech synthesis. First, we present ***SpeechJudge-Data***, a large-scale human feedback corpus of 99K speech pairs. The dataset is constructed using a diverse set of advanced zero-shot text-to-speech (TTS) models across diverse speech styles and multiple languages, with human annotations for both intelligibility and naturalness preference. From this, we establish ***SpeechJudge-Eval***, a challenging benchmark for speech naturalness judgment. Our evaluation reveals that existing metrics and AudioLLMs struggle with this task; the best-performing model, Gemini-2.5-Flash, achieves less than 70% agreement with human judgment, highlighting a significant gap for improvement. To bridge this gap, we develop ***SpeechJudge-GRM***, a generative reward model (GRM) based on Qwen2.5-Omni-7B. It is trained on SpeechJudge-Data via a two-stage post-training process: Supervised Fine-Tuning (SFT) with Chain-of-Thought rationales followed by Reinforcement Learning (RL) with GRPO on challenging cases. On the SpeechJudge-Eval benchmark, the proposed SpeechJudge-GRM demonstrates superior performance, achieving 77.2% accuracy (and 79.4% after inference-time scaling @10) compared to a classic Bradley-Terry reward model (72.7%). Furthermore, SpeechJudge-GRM can be also employed as a reward function during the post-training of speech generation models to facilitate their alignment with human preferences.

1 INTRODUCTION

The collection and integration of human feedback corpora for model alignment has become a critical stage in the development of modern large-scale generative models, proving indispensable in domains such as text (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), image (Xu et al., 2023; Kirstain et al., 2023), and video generation (Xu et al., 2024; Liu et al., 2025b).

In the field of speech synthesis, ***naturalness*** has long been a cornerstone subjective metric for quality assessment (Ju et al., 2024; Anastassiou et al., 2024; Du et al., 2024a; Xu et al., 2025; KimiTeam et al., 2025), representing one of the most general-purpose indicators of performance (Taylor, 2009; Tan, 2023). Prior research has explored automated speech assessment through MOS predictors (Saeki et al., 2022; Huang et al., 2024) and constructed the human feedback corpora for specific attributes like the low-level acoustic quality (Wang et al., 2025c). However, a large-scale human feedback corpus centered on the holistic quality of naturalness—and a corresponding reward model trained to capture these preferences—remains a notably underexplored area. To fill this void, this paper focuses on the dimension of speech naturalness and present a three-part contribution:

[*]Equal Contribution.
[†]Corresponding author.

1. A Large-scale Human Feedback Dataset: *SpeechJudge-Data*. We recruit human annotators to provide feedback on synthesized speeches, with a focus on assessing two fundamental speech aspects: *intelligibility* and *naturalness*. For data synthesis, we employ a diverse set of advanced, open-source zero-shot TTS models with varying architectures (such as CosyVoice2 (Du et al., 2024b), Ints (Zhang et al., 2025b), F5-TTS (Chen et al., 2025b), and MaskGCT (Wang et al., 2025e)) to produce the compared speech pairs. We prepare speech references in both regular and expressive styles, construct multilingual target texts, and cover both monolingual and cross-lingual synthesis scenarios to ensure data diversity (Section 3.1). We instruct human annotators to perform two tasks based on a speech pair (Figure 1): (a) pointwise annotation of text accuracy to assess intelligibility, and (b) pairwise preference annotation to judge relative speech naturalness. This extensive effort, involving 69 labelers over two months, results in 99K annotated pairs, with each pair receiving an average of 2.49 annotations from different labelers. We believe the SpeechJudge-Data can serve as a valuable corpus for alignment research in speech synthesis (e.g., DPO alignment (Rafailov et al., 2023) or reward modeling (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022) in Section 5).

2. An Evaluation Benchmark for Speech Naturalness Judgment: *SpeechJudge-Eval*. We design a dedicated evaluation benchmark for the task of speech naturalness judgment. The task is structured as follows: given a target text and two corresponding speech samples, a model needs to judge which one is more natural. To construct the evaluation set, we select a subset from the SpeechJudge-Data where human annotators demonstrated high inter-annotator agreement, ensuring a high-quality ground truth. We assess the naturalness judgment capabilities of a wide range of metrics and models, including Word Error Rate (WER) (Radford et al., 2023; Gao et al., 2023), Fréchet Audio Distance (FAD) (Kilgour et al., 2019), MOS predictors (Saeki et al., 2022; Reddy et al., 2022; Tjandra et al., 2025), Deepfake Detectors (Jung et al., 2022; Wang et al., 2025b), and AudioLLMs (Xu et al., 2025; KimiTeam et al., 2025; Xiaomi, 2025; Comanici et al., 2025; Hurst et al., 2024). Our evaluations reveal that even the most capable model—specifically, Gemini-2.5-Flash (Comanici et al., 2025) in our experiments—achieved less than 70% agreement with human preferences. This finding highlights a significant performance gap and underscores the substantial room for research and improvement in automated speech naturalness judgment.

3. A Generative Reward Model for Speech Naturalness: *SpeechJudge-GRM*. To develop a reward model that more effectively captures human preferences, we develop SpeechJudge-GRM, a generative reward model (GRM) (Zhang et al., 2025a; Liu et al., 2025c) trained on the SpeechJudge-Data. Specifically, we base our model on Qwen2.5-Omni-7B (Xu et al., 2025) and design a two-stage post-training process. During the first stage, we perform Supervised Fine-Tuning (SFT) as the "cold start" to improve the model's instruction-following and rationale-based reasoning capabilities. To achieve this, we leverage Gemini-2.5-Flash (Comanici et al., 2025) to generate Chain-of-Thought (CoT) data for speech naturalness judgment task. In the second stage, we focus on more challenging cases of SpeechJudge-Data, which we define as instances where Gemini-2.5-Flash fails to make the correct judgment. Treating the human-annotated labels as the verifiable reward (DeepSeek-AI et al., 2025; Liu et al., 2025c), we apply the GRPO-based Reinforcement Learning (RL) stage (Shao et al., 2024). Our experiments demonstrate that when trained on the same data, SpeechJudge-GRM significantly outperformed the classic Bradley-Terry reward model (BTRM) (Bradley & Terry, 1952; Rafailov et al., 2023), achieving a higher accuracy in predicting human preferences (77.2% for SpeechJudge-GRM vs. 72.7% for SpeechJudge-BTRM, Table 3). Besides, SpeechJudge-GRM also supports inference-time scaling and offers explainability through its CoT outputs. Furthermore, SpeechJudge-GRM can also be employed as an objective naturalness metric for sample selection (Figure 5) or as a reward function in RL algorithms to enhance the quality of existing speech generation models (Figure 6).

We release all resources from this study at https://github.com/AmphionTeam/SpeechJudge. Audio samples are available at https://speechjudge.github.io/.

2 Related Work

Human Alignment for Speech Generation Aligning generative models with human feedback has proven crucial, a process also known as RLHF in LLMs (Ouyang et al., 2022; Bai et al., 2022). In the vision domain, many similar human preference datasets exist, such as Pick-a-Pic (Kirstain et al., 2023), ImageReward (Xu et al., 2023), and VideoReward (Liu et al., 2025b). The speech

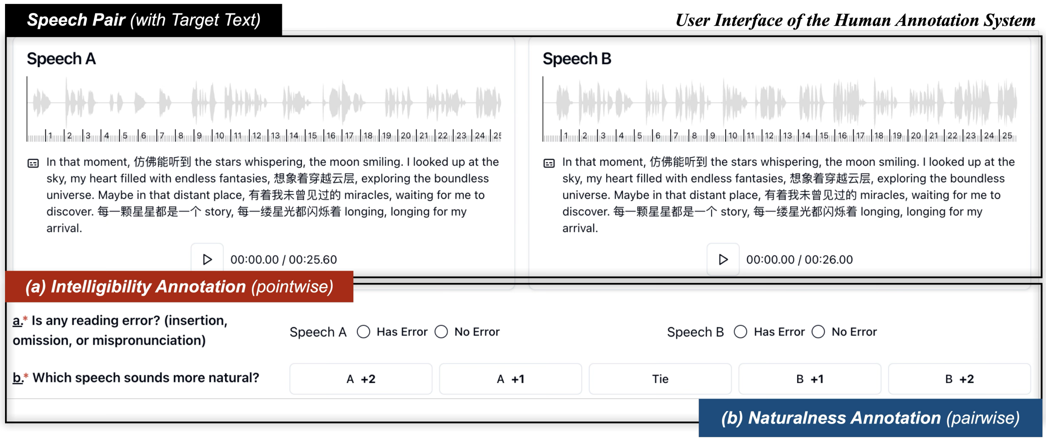


Figure 1: SpeechJudge-Data consists of speech pairs (with corresponding text) synthesized by multiple zero-shot TTS models. For each pair, human annotators need to perform **(a)** a pointwise annotation of text accuracy to assess intelligibility, and **(b)** a pairwise preference annotation to judge the relative speech naturalness.

synthesis field, pioneering efforts to construct human corpora involved MOS datasets (Saeki et al., 2022; Huang et al., 2024). However, these datasets often did not use advanced TTS models for data generation, provided only the pointwise labels rather than the direct pairwise human preference, and were limited in scale. More recently, efforts have focused on building human feedback corpora centered on specific speech attributes, such as low-level acoustic quality (Wang et al., 2025c), intelligibility (Zhang et al., 2025b), or the instruction-following capabilities of spoken dialogue systems (Ji et al., 2025; Ge et al., 2025). Despite this progress, a large-scale human feedback corpus built specifically around *naturalness*—one of the most general-purpose and fundamental metrics for speech synthesis (Taylor, 2009; Tan, 2023)—has remained a critical missing piece.

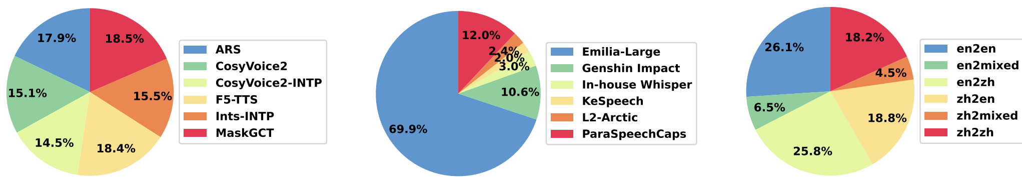
AudioLLM as a Judge Using LLMs as automated quality evaluators is a prominent topic in the textual LLM field, popularized by the "LLM-as-a-judge" paradigm (Zheng et al., 2023). This idea has recently been extended to the audio domain. A concurrent work, AudioJudge (Manakul et al., 2025), evaluates the capabilities and limitations of using AudioLLMs for speech quality assessment and paralinguistic understanding via prompt engineering. Furthermore, many studies have focused on fine-tuning AudioLLMs to better expose their understanding capabilities for specific tasks, such as discriminating the human-likeness of audio (Wang et al., 2025d), modeling low-level acoustic qualities (Chen et al., 2025a; Wang et al., 2025c), unifying multiple speech quality evaluation tasks into a single AudioLLM (Wang et al., 2025a), and enhancing the assessment of instruction-following in spoken dialogue systems (Ji et al., 2025; Ge et al., 2025). However, how to improve the ability of AudioLLMs to understand and judge speech naturalness, and how to use their quality-assessment capabilities as a reward to improve the post-training of speech generation models themselves, remain significantly underexplored.

3 SPEECHJUDGE-DATA

Our work is grounded in **SpeechJudge-Data**, a large-scale human feedback corpus for assessing the *intelligibility* and *naturalness* of synthesized speech. Formally, we aim to construct a dataset $\mathcal{D} = \{(t, a_1, a_2)\}$, where each triplet comprises a pair of synthesized speech samples (a_1, a_2) and the corresponding target text t. We instruct annotators to provide pointwise intelligibility and pairwise naturalness preference annotations based on $\mathcal{D}$ (Figure 1).

3.1 DATASET CONSTRUCTION

We employ a diverse set of recent advanced zero-shot TTS models to prepare the dataset $\mathcal{D}$. Formally, for each sample (t, a_1, a_2), we denote the synthesized speech a_i as being produced by the model $\mathcal{M}_{tts}$, i.e., $a_i \sim \mathcal{M}_{tts}(a_{ref}, t)$, where a_{ref} is the reference speech.



(a) TTS Models. (b) Speech References. (c) Language Settings.

Figure 2: Distribution of SpeechJudge-Data.

Model Selection For $\mathcal{M}_{tts}$, we select the following six models of three architectures to enrich the distribution of the synthetic data (Figure 2a): (1) **AR-based**: ARS (Wang et al., 2025e), CosyVoice2 (Du et al., 2024b), CosyVoice2-INTP (Zhang et al., 2025b), and Ints-INTP (Zhang et al., 2025b). The latter two are released by Zhang et al. (2025b) as intelligibility-enhanced models. (2) **FM-based**: F5-TTS. (3) **MGM-based**: MaskGCT (Wang et al., 2025e).

Prompt Construction To build diverse prompts (a_{ref}, t) for TTS, for a_{ref}, we adopt both **regular** and **expressive** speech samples. The regular samples are randomly selected from the Emilia-Large dataset (He et al., 2025). The expressive samples are sourced from corpora rich in paralinguistics, including the emotional corpora: ParaSpeechCaps (Deb et al., 2024), the accented corpora: L2-Arctic (Zhao et al., 2018) and KeSpeech (Tang et al., 2021), the whisper samples from an in-house corpus, and the character voices from video games Genshin Impact (simon3000, 2023). We display the detailed distribution of speech references in Figure 2b.

The target text t paired with each a_{ref} is constructed as follows: For regular a_{ref} samples, we randomly sample transcriptions from the Emilia-Large dataset (He et al., 2025). These are then refined using DeepSeek-V3 (DeepSeek-AI et al., 2024) to correct typos and normalize punctuations. For expressive a_{ref} samples, we instruct DeepSeek-V3 to generate several scripts in different writing styles, tailored to the topic of a_{ref} (see Appendix B.1 for more details). The languages of the target texts included Chinese (*zh*), English (*en*), and Chinese-English code-switching (*mixed*). For the combinations (a_{ref}, t), we include both monolingual settings (*en2en* and *zh2zh*) and cross-lingual settings (*zh2en*, *en2zh*, *zh2mixed*, and *en2mixed*), where *zh2en* denotes Chinese a_{ref} with English t, and similarly for others. The distribution of the language settings of (a_{ref}, t) is shown in Figure 2c.

Speech Pair Construction To ensure the diversity of the (a_1, a_2) pairs being compared, we follow Zhang et al. (2025b) and adopt both intra-model (i.e.,a_1 and a_2 being generated by the same model) and inter-model pairs (i.e., a_1 and a_2 being generated by the different models). The distribution of the speech pair is shown in Figure 7.

3.2 Human Annotation

Given a sample (t, a_1, a_2), human annotators are instructed to perform both pointwise intelligibility and pairwise naturalness annotations (Figure 1). For intelligibility, annotators perform a binary classification to determine whether the speech (a_1 and a_2) accurately reads the text t without any content insertion, omission, or mispronunciation. For naturalness, they perform a five-scale Comparative Mean Opinion Score (CMOS) annotation to determine which of the two audio clips (a_1 or a_2) sounds more natural and human-like.

We recruited professional annotators from a specialized data annotation firm in China and provided them with training for speech naturalness judgement. All annotators assigned to Chinese data were native speakers. For the English and code-switching datasets, annotators were required to have a proficiency level equivalent to at least CET-6. All personnel underwent standardized training based on a detailed annotation manual. Initially, we conducted a pilot study among researchers to refine the guidelines for clarity and unambiguity. To ensure annotation quality, each sample (t, a_1, a_2) was independently annotated by two individuals. A third annotator was introduced if any disagreements. The detailed annotation guidelines are provided in Appendix C.

Statistics We recruit 69 annotators and conduct annotations over two months. The resulting constructed dataset D, which we denote as SpeechJudge-Data (raw), contains 99K (t, a_1, a_2) samples, with each sample receiving an average of 2.49 annotations from different labelers. The market value of this annotation scale is estimated at over 500K RMB (about 70K USD). Based on the raw dataset, we also construct several subsets for analysis and reward model training. We provide detailed descriptions of each subset and its applications in the following sections and in Appendix B.2.

Human Agreement Analysis We analyze the human annotations for naturalness in this section; discussions regarding intelligibility are provided in Appendix C.3. For naturalness annotations, we evaluate the inter-annotator agreement across our constructed dataset. To simplify the analysis, given the sample (t, a_1, a_2), we transform the five-scale naturalness scale (CMOS) into a ternary classification system: either a_1 is better, a_2 is better, or their quality is a Tie. Based on this simplified classification, we categorize the annotation results into four distinct levels of agreement[1]: (1) **Full Agreement (FA)**: A consensus is reached among all annotators, with all ratings pointing to the same outcome (e.g., "2A", "3A", "2B", "3B"). We use "2A" to indicate that two annotators both rated a_1 as better, while "3B" denotes three annotators all rating a_2 as better. (2) **Weak Agreement (WA)**: This level captures cases where two annotators agree on a specific polarity, while the third annotator marks a Tie (e.g., "2A+1T", "2B+1T"). We also include the "2T+1A" and "2T+1B" cases in this level. (3) **Weak Disagreement (WD)**: This occurs when two annotators' ratings share the same polarity, but the third's rating is the opposite (e.g., "2A+B", "2B+A"). (4) **Full Disagreement (FD)**: This represents a complete lack of consensus, where all three annotators provide different classifications, denoted as "1A+1B+1T".

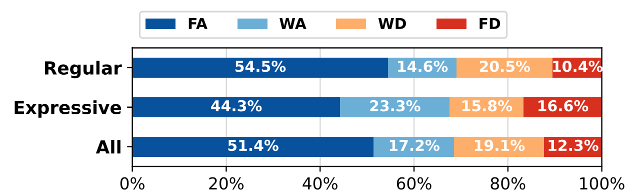


Figure 3: Distribution of SpeechJudge-Data on different levels of human agreement.

In Figure 3, we demonstrate the distribution of these human agreement levels for the SpeechJudge-Data and its two subsets, *regular* and *expressive* (which are defined by their speech references). The figure shows that about 70% of the entire dataset falls into the Full Agreement (51.5%) or Weak Agreement (17.2%) levels. Furthermore, we observe that the *expressive* subset has a lower agreement level than the *regular* subset, which suggests that human evaluation of expressive speech generation is inherently a more challenging problem. Besides this sample-level agreement analysis, we also analyze the reliability of individual annotators, and we will discuss this in the Appendix C.1.

4 SpeechJudge-Eval

To evaluate speech naturalness, existing studies typically organize their own listening tests, which often have inconsistent settings across different papers (Ju et al., 2024; Anastassiou et al., 2024; Du et al., 2024a; Xu et al., 2025; KimiTeam et al., 2025). Alternatively, previous researchers use proxy MOS predictors, such as UTMOS (Saeki et al., 2022), as an objective metric. However, it remains an underexplored problem whether these metrics can accurately judge the naturalness of more advanced speech generation models (Wang et al., 2025e; Du et al., 2024b; Chen et al., 2025b; Zhang et al., 2025b) and align with human preferences. Motivated by this, we construct a benchmark, **SpeechJudge-Eval**, specifically for the speech naturalness judgment task.

4.1 Task Description

Task Formulation We formulate the naturalness judgment task as a pairwise comparison, specifically a *win-or-lose* binary classification task: Given a target text t and a corresponding audio pair (a_1, a_2), a model needs to determine which audio has better naturalness. This results in a binary choice: either a_1 is better or a_2 is better. We use the human answer as the ground truth, and use

[1]**Note**: Each sample of SpeechJudge-Data is independently annotated by a minimum of two and a maximum of three annotators (Appendix C).

Table 1: Protocols of different models for naturalness judgment.

Type	Protocol
Objective Metrics	WER ↓, Naturalness ↑ SIM ↑, Naturalness ↑ FAD ↓, Naturalness ↑
MOS Predictors	MOS ↑, Naturalness ↑
Deepfake Detectors	Fake ↓, Naturalness ↑
AudioLLMs	Score ↑, Naturalness ↑

Prompts of AudioLLMs

- We are comparing the naturalness of two models' outputs. The models need to speak the target text accurately and naturally.
- Target text: $\{t\}$, Output A: $\{a_1\}$, Output B: $\{a_2\}$. Analyze the two outputs above, and score them with number from 1 to 10. Note:
  - Please evaluate the naturalness of both audio outputs based on the following criteria: ***Prosody and Intonation***, ***Pacing and Rhythm***, ***Articulation and Clarity***, and ***Overall Naturalness***.
  - After conducting a detailed analysis of each criterion, using the following output template to highlight your conclusion: Output A: X, Output B: X.

[*] We instruct AudioLLMs using two modes of prompt: *plain* and *CoT*. The text in blue is only employed during the *CoT* mode.

Table 2: Accuracy of speech naturalness judgment across different models on SpeechJudge-Eval.

Model	Regular	Expressive	Total
Objective Metrics			
WER	59.3	57.0	57.9
SIM	47.5	42.5	44.5
FAD	50.3	47.5	48.6
MOS Predictor			
DNSMOS	61.0	55.8	57.9
UTMOS	54.0	53.5	53.7
Content Enjoyment (CE)	69.3	55.2	60.8
Content Usefulness (CU)	61.3	54.7	57.3
Production Complexity (PC)	39.3	48.7	44.9
Production Quality (PQ)	61.3	54.3	57.1
Deepfake Detectors			
AASIST	40.5	50.8	46.7
ADV	35.3	40.3	38.3
AudioLLMs (Open-source)			
Phi-4-Multimodal	54.8	58.5	57.0
Qwen2.5-Omni-7B	62.0	59.7	60.6
Kimi-Audio-7B-Instruct	65.5	68.0	67.0
Gemma-3n-E4B-it	49.0	47.7	48.2
Voxtral-Mini-3B-2507	60.0	53.3	56.0
MiDashengLM	58.8	63.5	61.6
MiMo-Audio-7B-Instruct	61.3	49.3	54.1
AudioLLMs (Closed-source)			
Gemini-2.5-Flash	73.5	66.2	69.1
Gemini-2.5-Pro	73.0	62.2	66.5
GPT-4o mini Audio	56.3	46.7	50.5
GPT-4o Audio	71.5	64.7	67.4

[*] We use the protocols of Table 1 to establish judgment rules for different models. The results of AudioLLMs here are obtained using the *plain* prompt of Table 1.

Accuracy to measure the judgment performance of a model $\mathcal{M}$ on the evaluation set $\mathcal{D}$:

$$\textbf{Accuracy} = \frac{1}{|\mathcal{D}|}\sum_{d=0}^{|\mathcal{D}|}\mathbb{I}(y_{\mathcal{M}} = y_{\mathcal{H}}), \tag{1}$$

where $|\mathcal{D}|$ is the total number of samples in the evaluation set, $y_{\mathcal{M}}$ and $y_{\mathcal{H}}$ represent the answers of the model $\mathcal{M}$ and human for the sample d, respectively. $\mathbb{I}$ is the indicator function.

Evaluation Data We sample a subset of the SpeechJudge-Data to create the evaluation set for SpeechJudge-Eval. Specifically, we first select a subset that contains only *preference data* (i.e., we filter out samples with the "Tie" annotation), and then choose only those with full-agreement-level (FA) samples to ensure a high-quality ground truth. We perform sampling from both the *regular* and *expressive* subsets of SpeechJudge-Data and proportionally cover the three target text languages (*zh*, *en*, and *mixed*) within each subset. The final SpeechJudge-Eval dataset consists of 1,000 samples. The construction details of SpeechJudge-Eval and its distribution can be found in Appendix B.2.

4.2 Benchmark for Different Models

We test the naturalness judgment capability of various models based on SpeechJudge-Eval. We consider four different categories of models, whose evaluation protocols are shown in Table 1:

1. **Objective metrics**, such as WER (Radford et al., 2023; Gao et al., 2023), SIM (Chen et al., 2022), and FAD (Kilgour et al., 2019) in audio generation tasks. We assume that a better value of these metrics (e.g., lower for WER and FAD; higher for SIM) indicates better naturalness.
2. **MOS Predictors**, including DNSMOS (Reddy et al., 2022), UTMOS (Saeki et al., 2022), and predictors from audiobox-aesthetics (CE, CU, PC, and PQ) (Tjandra et al., 2025). We assume that a higher MOS score corresponds to better naturalness.

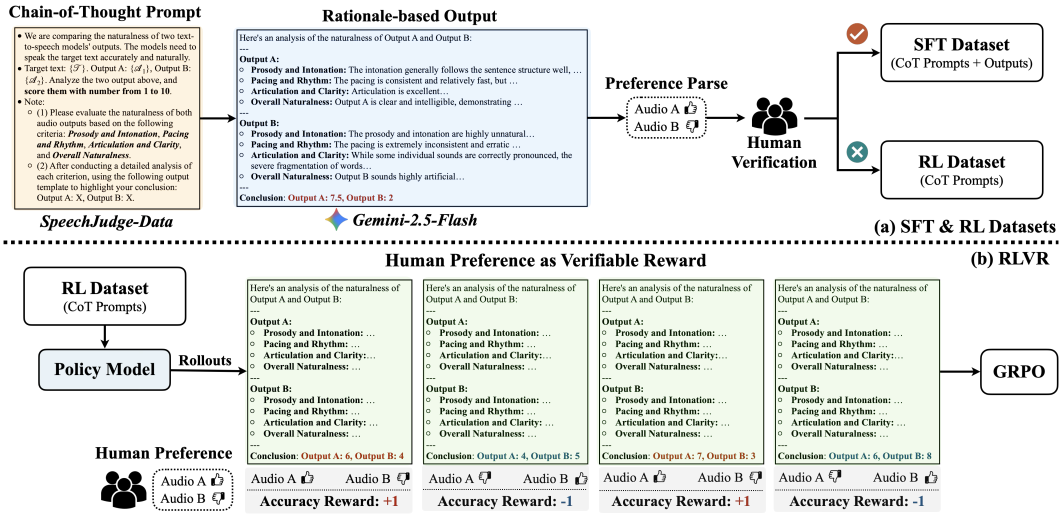


Figure 4: **SpeechJudge-GRM**: (a) We employ Gemini-2.5-Flash as a teacher model to generate CoT rationales for SpeechJudge-Data. We use the samples where Gemini-2.5-Flash's preference aligns with human as the SFT dataset, while the remaining samples are reserved for the RL stage. (b) We treat the human preference as a verifiable reward to train the GRM with GRPO.

3. **Deepfake detectors**, which are typically pre-trained on a binary classification task to predict whether an audio is fake or not (Jung et al., 2022; Wang et al., 2025b). We assume that an audio with a lower fake probability should have better naturalness.
4. **AudioLLMs**, which are employed to test their speech naturalness understanding capabilities in a *zero-shot* manner[2]. We include the open-source Phi-4-Multimodal (Abouelenin et al., 2025), Qwen2.5-Omni (Xu et al., 2025), Kimi-Audio (KimiTeam et al., 2025), Gemma-3n (Kamath et al., 2025), Voxtral (Liu et al., 2025a), MiDashengLM (Dinkel et al., 2025), Mimo-Audio (Xiaomi, 2025), and the closed-source Gemini-2.5 (Comanici et al., 2025) and GPT-4o (Hurst et al., 2024). We use the *plain* prompt of Table 1 to instruct the model to pairwise score the naturalness of two audios. We use their grading to determine the naturalness preference.

The performance of different models on SpeechJudge-Eval is presented in Table 2. A key observation is that speech naturalness judgment is a highly challenging task. The best-performing model, Gemini-2.5-Flash, still only achieves less than 70% agreement with human preferences. When comparing different models, we find that: (1) common objective metrics and MOS predictors show only a weak correlation with human preferences, often achieving less than 60% accuracy and sometimes performing at the level of a random guess (around 50%). (2) While deepfake detectors are highly effective at distinguishing between machine-generated and human-recorded speech (Wang et al., 2025b; Jung et al., 2022), their ability to do so is not well-aligned with the naturalness objective when comparing two generated samples. (3) AudioLLMs demonstrate significant potential for this task. While some models, such as Gemma-3n and GPT-4o mini Audio, perform at a chance level, a number of others achieve an accuracy exceeding 60%. This promising performance motivates us to further leverage these AudioLLMs for the design of a reward model for speech naturalness.

5 SPEECHJUDGE-GRM

Based on the proposed SpeechJudge-Data, we further explore how to train a reward model capable of accurately capturing human preferences. Specifically, we propose **SpeechJudge-GRM**, where we leverage the inherent audio understanding capabilities of AudioLLMs (specifically, Qwen2.5-Omni-7B (Xu et al., 2025)) to elicit their speech naturalness judgment capability. Compared to the classic BTRM (Bradley & Terry, 1952), the key strengths of GRM are its ability to enable Chain-of-Thought (CoT) reasoning and its support for test-time computation via majority voting, which ultimately leads to improved preference judgment performance (Zhang et al., 2025a).

[2]We assume that our adopted AudioLLMs have not been directly trained on the speech naturalness judgment task. Their performance on this benchmark is therefore considered a *zero-shot* capability.

Table 3: Accuracy of speech naturalness judgment of SpeechJudge-GRM.

Model	Regular	Expressive	Total
Qwen2.5-Omni-7B	62.0	59.7	60.6
Gemini-2.5-Flash	73.5	66.2	69.1
SpeechJudge-BTRM	77.5	69.5	72.7
SpeechJudge-GRM (SFT)	77.8	73.7	75.3
w/ Voting@10	77.4	77.6	77.6
SpeechJudge-GRM (SFT+RL)	79.0	76.0	77.2
w/ Voting@10	80.5	78.7	79.4

* Our evaluation is conducted on SpeechJudge-Eval (like Table 2). **w/ Voting@10**: For each prompt, the GRM generates 10 outputs, and we use the majority voting from these 10 outputs as the final result.

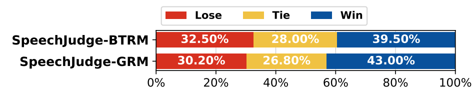


Figure 5: Subjective evaluation of using SpeechJudge-GRM for high-naturalness sample selection. Human subjects compare a best-of-100 output of Qwen2.5-Omni-7B (Talker), chosen by either SpeechJudge-BTRM or SpeechJudge-GRM, against a randomly output.

5.1 Methodology

We develop SpeechJudge-GRM based on Qwen2.5-Omni-7B (Thinker) (Xu et al., 2025). Inspired by the powerful capabilities of RL with the verifiable reward (RLVR) (Shao et al., 2024; DeepSeek-AI et al., 2025), our natural initial approach is to treat the human preference $y_{\mathcal{H}}$ for the pair (a_1, a_2) as a verifiable reward, and launch a RLVR training based on Qwen2.5-Omni. However, in practice, we find that the instruction-following reasoning capabilities of Qwen2.5-Omni are very weak (more detailed discussions can be found in Appendix E). Therefore, we adopt a two-stage post-training process ("SFT + RL") to develop SpeechJudge-GRM (Figure 4). We describe the details as follows.

SFT Stage We consider SFT as a "cold start" stage to improve the Qwen2.5-Omni's instruction-following, reasoning, and speech naturalness understanding capabilities. We select Gemini-2.5-Flash (Comanici et al., 2025)—one of the best-performing closed-source models on SpeechJudge-Eval (Table 2)—to serve as a teacher model, and instruct it to generate the CoT data. Specifically, for each sample $d = (t, a_1, a_2, y_{\mathcal{H}})$ from SpeechJudge-Data, we use the CoT prompt from Table 1 (denoted as $\mathbf{I}_{CoT}$) to instruct Gemini-2.5-Flash to generate a rationale-based output (denoted as $\mathbf{O}_{teacher}$). We then extract the preference judgment ($y_{\mathcal{M}}$) from this output. For samples where Gemini-2.5-Flash's preference is consistent with the human (i.e., $y_{\mathcal{M}} = y_{\mathcal{H}}$), we concatenate the CoT prompt and the model's output, $[\mathbf{I}_{CoT}, \mathbf{O}_{teacher}]$, to create a data point for our SFT dataset. Conversely, we consider the sample d a challenging case and reserve the prompt $\mathbf{I}_{CoT}$ for the second-stage RL dataset. During the SFT stage, for each training sample $[\mathbf{I}_{CoT}, \mathbf{O}_{teacher}]$, we perform the next token prediction only on the segment $\mathbf{O}_{teacher}$.

RL Stage We treat the annotated human preference as a verifiable reward, and, building on the SFT model, we further trained it using the GRPO algorithm (Shao et al., 2024). Specifically, for each sample $d = (t, a_1, a_2, y_{\mathcal{H}})$ in the RL dataset, we adopt the CoT prompt to instruct the policy model to conduct multiple rollouts during each iteration. For the i-th rollout, we parse the model's preference for (a_1, a_2), denoted as $y^i_{\mathcal{M}}$. Following (Liu et al., 2025c), we use an accuracy-based rule to calculate the reward: the reward is 1 if $y^i_{\mathcal{M}} = y_{\mathcal{H}}$, and -1 otherwise. In other words, during the RL stage, we only constrain the model's final naturalness judgment to align with human preferences, allowing the model to autonomously optimize its reasoning and rationale generation capabilities.

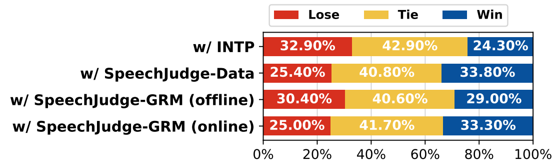
We denote the training dataset of SpeechJudge-GRM as SpeechJudge-Data (train). Its construction process is as follows (see Appendix B.2 for more details). Based on the raw SpeechJudge-Data, we first filter out all samples at the Full Disagreement (FD) level. For the other samples—at the FA, WA, and WD levels—we apply a majority voting principle among annotators to determine the final label for each. We then further exclude samples with a "Tie" label, using only the remaining preference data to form the SpeechJudge-Data (train). We use LoRA (Hu et al., 2022) to fine-tune the GRM during both the SFT and RL stages. Other experimental setup details are provided in Appendix F.

5.2 Effectiveness of SpeechJudge-GRM on Naturalness Judgement

To verify the effectiveness of SpeechJudge-GRM for naturalness judgment, we evaluate it on the SpeechJudge-Eval benchmark. We develop SpeechJudge-BTRM as a baseline, which utilizes the BTRM paradigm (Bradley & Terry, 1952; Rafailov et al., 2023) by adding a linear layer on Qwen2.5-Omni-7B (Thinker) to produce a single scalar reward prediction. SpeechJudge-BTRM also uses LoRA fine-tuning and uses the same training data as SpeechJudge-GRM.

Model	T-ACC	N-CMOS
Qwen2.5-0.5B-TTS	84.0%	0.00
w/ INTP	87.0%	0.18 $_{\pm 0.07}$
w/ SpeechJudge-Data	91.0%	0.16 $_{\pm 0.08}$
w/ SpeechJudge-GRM (offline)	91.0%	0.21 $_{\pm 0.12}$
w/ SpeechJudge-GRM (online)	90.0%	0.25 $_{\pm 0.09}$

(a) Text Accuracy (T-ACC) and Naturalness CMOS (N-CMOS).



(b) Win/Lose/Tie of speaker similarity after post-training.

Figure 6: Post-training of Qwen2.5-0.5B-TTS based on SpeechJudge. We display the objective results (WER and SIM) in Table 14 of Appendix H.2.

From the results of Table 3, we can observe that: (1) The SpeechJudge-BTRM achieves a 72.7% agreement with human preferences on SpeechJudge-Eval, a level of performance comparable to the initial development of BTRMs in the textual LLM RLHF field (Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022). (2) After conducting SFT training with the CoT data, the accuracy of SpeechJudge-GRM (SFT) reaches 75.3%. Besides, further RLVR training improves the final model SpeechJudge-GRM (SFT+RL) to an accuracy of 77.2%. (3) Due to the generative nature of the GRM, we can further enhance the accuracy of SpeechJudge-GRM using inference-time scaling. For example, by using majority voting across 10 outputs instead of just one, the accuracy is improved by approximately 2 percentage points (75.3% $\rightarrow$ 77.6%; 77.2% $\rightarrow$ 79.4%). These results collectively verify the effectiveness of our proposed SpeechJudge-GRM for judging speech naturalness.

5.3 High-Quality Sample Selection based on SpeechJudge-GRM

We investigate the effect of SpeechJudge-based reward models for high-quality sample selection. We use the hard cases from SeedTTS-Eval (Anastassiou et al., 2024) and the code-switching cases from Amphion-TTS-Eval (Zhang et al., 2024) as target texts. For each text, we instruct the Qwen2.5-Omni-7B (Talker) (Yang et al., 2024) to generate 100 speeches. We then ask human subjects to compare the best-of-100 output—as selected by either SpeechJudge-BTRM or SpeechJudge-GRM—against a randomly sampled output. The evaluation measures the win/lose/tie ratios based on speech naturalness. From Figure 5, we observe that the best-of-100 samples selected by both SpeechJudge-BTRM and SpeechJudge-GRM are more likely to outperform a randomly selected sample from the same set. This finding demonstrates the advantage of using the SpeechJudge-Data corpus for training human-aligned reward model. Furthermore, SpeechJudge-GRM exhibits better performance than SpeechJudge-BTRM, which highlights the superiority of the proposed GRM.

5.4 Post-Training of Zero-Shot TTS based on SpeechJudge-GRM

We investigate the effect of using SpeechJudge-GRM as a reward function for post-training of TTS model. Specifically, we develop a new zero-shot TTS model, **Qwen2.5-0.5B-TTS**, to serve as the base model, which was not involved in the construction of the SpeechJudge-Data. This model is based on Qwen2.5-0.5B (Yang et al., 2024), adopts the classic two-stage "AR+Diffusion" architecture (Anastassiou et al., 2024; Du et al., 2024a), uses the speech tokenizer from DualCodec (Li et al., 2025a), and is pre-trained on the Emilia dataset (He et al., 2025).

Based on this pre-trained model, we design four comparative methods: (1) **w/ INTP**: We use the intelligibility preference dataset, INTP (Zhang et al., 2025b), to perform offline DPO alignment (Rafailov et al., 2023). (2) **w/ SpeechJudge-Data**: We use the SpeechJudge-Data (train) to perform offline DPO alignment. (3) **w/ SpeechJudge-GRM (offline)**: We use SpeechJudge-GRM as an offline preference data annotator. We take all speech pairs from the INTP dataset and re-annotate their preference labels using SpeechJudge-GRM, then perform offline DPO alignment on the resulting data. (4) **w/ SpeechJudge-GRM (online)**: We use SpeechJudge-GRM as a reward function for the online DPO algorithm (Guo et al., 2024). The training data consists of only the prompts from INTP (i.e., the target texts and speech references for zero-shot TTS).

We use SeedTTS-Eval (Anastassiou et al., 2024) and Amphion-TTS-Eval (Zhang et al., 2025b, 2024, 2025c) as evaluation sets. We present the objective results (WER and SIM) in Table 14 and the subjective results in Figure 6. We observe that both intelligibility and naturalness are enhanced for all the four methods after post-training. Additionally, the post-training method based on SpeechJudge-

GRM achieves a greater improvement in naturalness (Figure 6a). Besides, the SpeechJudge-based methods could match or lead to a slight improvement in speaker similarity (Figure 6b).

6 Limitations and Future Work

While SpeechJudge-Data and SpeechJudge-GRM represent a step toward human-aligned speech naturalness judges, several limitations remain and open up directions for future work.

Scope of data and annotators. Our corpus is constructed entirely from synthetic TTS outputs in Chinese, English, and Chinese–English code-switching, and our annotators are professional raters in China (native Mandarin speakers with high but still L2 English proficiency). As shown in Appendix C.2, inter-annotator agreement is noticeably higher on Chinese than on English and mixed subsets, indicating that the current dataset primarily reflects the preferences of Chinese and Chinese–English bilingual listeners, and is tailored to TTS-style read speech rather than spontaneous conversation. Extending SpeechJudge-Data to more languages, speaking styles, and listener populations (including native speakers of other languages and more diverse cultural backgrounds) is an important direction for building more universal naturalness judges.

Residual failure cases. Our error analysis in Appendix G.3 shows that SpeechJudge-GRM's remaining mistakes on SpeechJudge-Eval concentrate on some specific trade-offs, such as clean but robotic vs. slightly noisy but lively speech, prosody vs. articulation, and extreme expressive styles like very high-F0 emotional speech or whispers. In these regimes the model can over-weight cleanness, under-weight style-appropriate prosody, or become effectively indifferent when preference gaps are extremely small. Future work could incorporate explicit modeling of recording conditions (e.g., background noise), style-aware priors, and targeted augmentation of expressive and cross-lingual examples to better capture these nuanced preferences.

CoT quality and teacher bias. The CoT capability of SpeechJudge-GRM is bootstrapped from a proprietary teacher (Gemini-2.5-Flash) via an SFT stage. Although our analyses in Appendix G.2 suggest that GRM's CoT is largely self-consistent and moderately faithful, the reasoning style still inherits biases from the teacher, and we do not perform large-scale human verification of the intermediate explanations. An interesting direction is to involve humans more directly in assessing and curating CoT rationales—similar in spirit to the concurrent SQ-LLM work (Wang et al., 2025a), which incorporates human involvement in CoT annotation—for example, by collecting human-written or human-edited analyses, or learning from explicit feedback on explanation quality, and to explore alternative, fully open-source teachers for bootstrapping.

From coarse-grained to fine-grained naturalness. In this work, naturalness is annotated and modeled at the *utterance* level: for each sentence, annotators choose which of two speeches is more natural, and SpeechJudge-GRM outputs a single decision per pair. However, in real-world speech, naturalness is often highly non-uniform within an utterance—some segments sound very natural while others contain local artifacts, disfluencies, or prosodic issues. Our current formulation does not explicitly localize such fine-grained phenomena. A promising future direction is to collect segment-level or time-aligned human feedback and to train reward models that can produce not only utterance-level judgments but also fine-grained scores or rationales over time.

7 Conclusion

In this work, we tackle the challenge of aligning speech synthesis with human perception of naturalness by introducing SpeechJudge: a suite consisting of a large-scale human preference dataset (SpeechJudge-Data), a challenging benchmark (SpeechJudge-Eval), and a generative reward model (SpeechJudge-GRM). Our benchmark shows that even strong AudioLLMs struggle at naturalness judgment, reaching under 70% agreement with human preferences. In contrast, the proposed SpeechJudge-GRM achieves 77.2% accuracy on SpeechJudge-Eval (up to 79.4% with inference-time scaling @10), outperforming a classic Bradley–Terry reward model (72.7%). We further demonstrate that SpeechJudge-GRM serves as an effective reward function for post-training TTS models, leading to improved perceived naturalness in downstream evaluations. By releasing our data, benchmark, and models, we hope to enable further research on human-aligned speech generation and more reliable evaluation of speech naturalness.

Ethics Statement

This research adheres to the ICLR Code of Ethics. Our dataset was constructed with feedback from paid professional annotators under fair labor conditions, and the data itself consists of synthesized speech from properly licensed corpora, safeguarding the privacy of all individuals. All human annotations were collected from professional annotators recruited by a third-party data annotation company under written informed consent, on synthetic TTS audio only, with no collection of personally identifying information. We acknowledge that our models may reflect linguistic biases present in the English and Chinese source data and recognize that generative speech technology has dual-use potential. We do not condone any malicious use of our work, such as the creation of misleading deepfakes.

Reproducibility Statement

To ensure full reproducibility, we will publicly release all key resources from this study. This includes the SpeechJudge-Data corpus, the SpeechJudge-Eval benchmark, the trained model checkpoints for SpeechJudge-GRM, and the source code for both reward model training and downstream experiments. The main paper and its appendices provide detailed descriptions of our methodology, data construction protocols, and experimental setups to allow for the complete verification of our findings. Audio samples are available on our project website.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China under Grant 62376237, 2023 Shenzhen stability Science Program, Internal Project Fund from Shenzhen Research Institute of Big Data (Grant No. T00120230002), and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No. 2023ZT10X044). We thank Yushun Zhang and the anonymous reviewers for their insightful comments and suggestions. We appreciate the efforts of all the subjects during the subjective evaluation.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Jun-Kun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint*, abs/2503.01743, 2025.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint*, abs/2406.02430, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson

Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint*, abs/2204.05862, 2022.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *ACL (1)*, pp. 6255–6271. Association for Computational Linguistics, 2025b.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint*, abs/2507.06261, 2025.

Chayan Deb, Adway Mitra, Pratik Kumar, Soumya Jahar, and Biplab Das. ParaSpeechCaps: A Parallel Speech-Caption Dataset for Speech-based Image Retrieval, 2024.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng

Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *arXiv preprint*, abs/2412.19437, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, abs/2501.12948, 2025.

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

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

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint*, abs/2412.10117, 2024b.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *INTERSPEECH*, pp. 2063–2067. ISCA, 2022.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*, pp. 1593–1597. ISCA, 2023.

Yuan Ge, Junxiang Zhang, Xiaoqian Liu, Bei Li, Xiangnan Ma, Chenglong Wang, Kaiyang Ye, Yangfan Du, Linfeng Zhang, Yuxin Huang, et al. Sagelm: A multi-aspect and explainable large language model for speech judgement. *arXiv preprint arXiv:2508.20916*, 2025.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Ramé, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online AI feedback. *arXiv preprint*, abs/2402.04792, 2024.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *arXiv preprint*, 2501.15907, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

Wen-Chin Huang, Szu-Wei Fu, Erica Cooper, Ryandhimas E. Zezario, Tomoki Toda, Hsin-Min Wang, Junichi Yamagishi, and Yu Tsao. The voicemos challenge 2024: Beyond speech quality prediction. In *SLT*, pp. 803–810. IEEE, 2024.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *arXiv preprint*, abs/2410.21276, 2024.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, Siqi Zheng, Jin Xu, Junyang Lin, and Zhou Zhao. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint*, abs/2505.09558, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning*, 2024.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas W. D. Evans. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP*, pp. 6367–6371. IEEE, 2022.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku

Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle K. Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry (Dima) Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *arXiv preprint*, abs/2503.19786, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pp. 2350–2354. ISCA, 2019.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *arXiv preprint*, abs/2504.18425, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. Dualcodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. In *INTERSPEECH*. ISCA, 2025a.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *ICLR*. OpenReview.net, 2025b.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de Las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, and Victor Paltz. Voxtral. *arXiv preprint*, abs/2507.13264, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. *arXiv preprint*, abs/2501.13918, 2025b.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint*, abs/2504.02495, 2025c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019.

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

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, pp. 886–890. IEEE, 2022.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *INTERSPEECH*, pp. 4521–4525. ISCA, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint*, abs/2402.03300, 2024.

simon3000. Genshin Impact Voice Lines. `https://huggingface.co/datasets/simon3000/genshin-voice`, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. In *NeurIPS*, 2020.

Xu Tan. *Neural Text-to-Speech Synthesis*. Springer, 2023.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *NeurIPS*, 2021.

Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint*, abs/2502.05139, 2025.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. Speechllm-as-judges: Towards general and interpretable speech quality evaluation. *arXiv preprint*, abs/2510.14664, 2025a.

Li Wang, Junyi Ao, Linyong Gan, Yuancheng Wang, Xueyao Zhang, and Zhizheng Wu. Audio deepfake verification. *arXiv preprint arXiv:2509.08476*, 2025b.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *ACL (1)*, pp. 23588–23609. Association for Computational Linguistics, 2025c.

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

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *ICLR*. OpenReview.net, 2025e.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL `https://github.com/XiaomiMiMo/MiMo-Audio`.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint*, abs/2412.21059, 2024.

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. Qwen2.5-omni technical report. *arXiv preprint*, abs/2503.20215, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint*, abs/2412.15115, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *arXiv preprint*, abs/2503.14476, 2025.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *ICLR*. OpenReview.net, 2025a.

Xueyao Zhang, Liumeng Xue, Yicheng Gu, Yuancheng Wang, Jiaqi Li, Haorui He, Chaoren Wang, Ting Song, Xi Chen, Zihao Fang, Haopeng Chen, Junan Zhang, Tze Ying Tang, Lexiao Zou, Mingxuan Wang, Jun Han, Kai Chen, Haizhou Li, and Zhizheng Wu. Amphion: An open-source audio, music and speech generation toolkit. In *SLT*. IEEE, 2024.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *ACL*. Association for Computational Linguistics, 2025b.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*, 2025c.

Guanlong Zhao, Siyuan Feng, Xinyuan Li, Shiyin Kang, Helen Zhao, and Shinji Watanabe. L2-ARCTIC: A Non-native English Speech Corpus. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 312–318, 2018. doi: 10.1109/SLT.2018.8639538.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

A THE USE OF LARGE LANGUAGE MODELS

We utilized Large Language Models (LLMs) for two assistive tasks: (1) to correct grammar and polish the language of this manuscript, and (2) to generate and normalize the target texts for SpeechJudge-Data during the dataset construction. All core scientific contributions, including ideation, methodology, and analysis, are entirely those of the authors.

B DETAILS OF SPEECHJUDGE-DATA

B.1 DETAILS OF PROMPT CONSTRUCTION

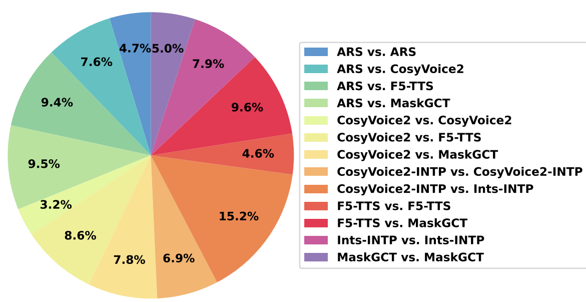


Figure 7: Distribution of the speech pairs of SpeechJudge-Data. The chart illustrates the percentage of both intra-model (e.g., ARS vs. ARS) and inter-model (e.g., ARS vs. CosyVoice2) pairs.

For the target texts paired with the regular speech references, we use DeepSeek-V3 (DeepSeek-AI et al., 2024) to fix typos and normalize punctuations, the prompt used is listed below.

System Prompt:
I obtained a text from an audio file based on some ASR models. Please help me clean it up (e.g., correct typos, add proper punctuation marks, and make the sentences semantically coherent). Note: (1) You can modify, add, or replace words that better fit the context to ensure semantic coherence. (2) Please only return the cleaned-up result without any explanation.

User Prompt (Example):
a panda eats shoes and leaves

System Output (Example):
A panda eats shoots and leaves.

For the target texts paired with the expressive speech references, we use DeepSeek-V3 to generate several scripts in different writing styles based on the speech reference's text, the prompt used is listed below.

System Prompt:
请按句子原始语言进行续写（中文：续写中文，英文：续写英文）。将给出的句子以10种截然不同的文风续写（中文：如古典文学风格、科幻风格、武侠风格等；英文：Formal, Informal,

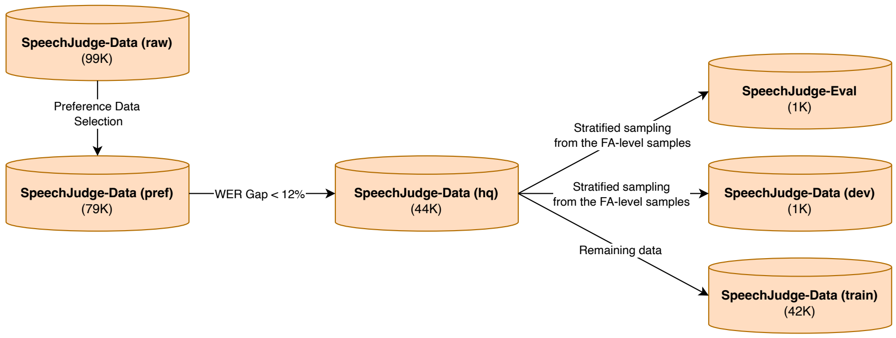


Figure 8: From the raw SpeechJudge-Data to its different subsets.

Literary, Journalistic, Advertising, etc.），直接续写即可，续写时避免重复原句内容。每段续写文本长度应控制在正常语速朗读约10-20秒左右。
请严格按照以下JSON格式返回结果，不添加任何其他解释或备注：
{
”continue”: [
{”type”: ”文风类型”, ”content”: ”续写内容”},
{”type”: ”文风类型”, ”content”: ”续写内容”},
...（共10条）
],
”random”: ”从以上续写中随机选出一条续写内容”,
”translation”: ”将random内容经过中英互译处理后的文本”,
”codeSwitching”: ”将random内容进行适度的中英code-switching，即在句中选取部分关键词汇转换为英文或中文后的文本”
}

User Prompt (Example):
A panda eats shoots and leaves.

System Output (Example):
{
”continue”: [
{”type”: ”Formal”, ”content”: ”This behavior is characteristic of the species Ailuropoda melanoleuca, which sustains itself primarily on a herbivorous diet consisting of various bamboo species. The consumption of both young shoots and mature leaves is essential for meeting its nutritional requirements.”},
{”type”: ”Informal”, ”content”: ”And then, after chowing down on all that bamboo, it just kinda chills out, maybe takes a long nap or something. They’re pretty laid-back animals, you know? Not much else going on when you’re full of leaves.”},
...
{”type”: ”Fantasy”, ”content”: ”Legend says that with every shoot consumed, a little more of the forest’s ancient magic is absorbed into its being. The leaves it leaves behind are not mere waste, but enchanted tokens that can heal or curse, depending on the heart of the finder.”},
...
],
”random”: ”Legend says that with every shoot consumed, a little more of the forest’s ancient magic is absorbed into its being. The leaves it leaves behind are not mere waste, but enchanted tokens that can heal or curse, depending on the heart of the finder.”,
”translation”: ”传说中，它每吃下一根竹笋，就会吸收一点森林的古老魔力。它身后留下的叶子并非单纯的废弃物，而是附有魔法的信物，可治愈亦可诅咒，全凭发现者的心地。”,
”codeSwitching”: ”Legend says that with every shoot consumed, a little more of the forest’s ancient magic 被吸收into its being. The leaves it leaves behind are not mere waste, but enchanted tokens that can heal 或curse, depending on the finder 的heart.”
}

B.2 SUBSETS OF SPEECHJUDGE-DATA

We construct several subsets based on SpeechJudge-Data (Figure 8). We begin with the **SpeechJudge-Data (raw)** corpus, containing 99K pairs, where each pair is annotated by multiple labelers as a five-scale naturalness CMOS. We aggregate these annotations via a majority vote for each pair, and subsequently discard all "Tie" pairs, yielding the 79K-pair human *preference* data, denoted as **SpeechJudge-Data (pref)**.

During our preliminary analysis based on SpeechJudge-Data (pref), we observe that a significant disparity in intelligibility between two speech samples can overshadow the subtler quality of naturalness, biasing human preference toward the more comprehensible sample. To mitigate this confounding factor and create a more high-quality dataset focused specifically on naturalness, we further refine the data. Specifically, we retain only pairs where the absolute WER gap of those is below 12%. This process results in the 44K-pair high-quality **SpeechJudge-Data (hq)** subset, ensuring that its preference labels are more reflective of genuine differences in naturalness.

From SpeechJudge-Data (hq), we construct our benchmark, **SpeechJudge-Eval**, by applying stratified sampling to FA-level pairs, resulting in 1,000 pairs; its composition is detailed in Table 4. Similarly, we use the same strategy to construct a validation set of the same size, **SpeechJudge-Data (dev)**. The remaining 42K pairs, **SpeechJudge-Data (train)**, constitute the training set for our reward models.

Table 4: Distribution of the SpeechJudge-Eval benchmark.

Subset	Source of Speech References	Languages of Target Texts	# Pairs
Regular	Emilia-Large	*en*	200
		zh	200
Expressive	ParaSpeechCaps, L2-Arctic, KeSpeech, Genshin, etc.	*en*	200
		zh	200
		mixed	200

C HUMAN ANNOTATION DETAILS

The complete annotation guidelines are attached below:

TASK INTRODCUTION

In each task, you are required to complete two evaluations:

1. **Pronunciation Error Detection**
   For each audio clip, we provide the **target text** that is intended to be read aloud. You need to determine if there are any pronunciation errors in the audio, such as omissions (missing words), insertions (extra words), or substitutions (wrong words).
2. **Naturalness Comparison**
   Compare two audio clips and determine which one sounds more natural and more like a real human speaking.

Attention:

- Please use headphones for the evaluation to better capture audio details and improve judgment accuracy.

ANNOTATION CRITERIA

A. PRONUNCIATION ERROR DETECTION

Pronunciation errors include the following three categories:

- **Omission:** Certain words from the target text are missed.
- **Insertion:** Extra words not in the target text are added, e.g., repeating words.
- **Substitution:** Certain words are misread, e.g., reading names, numbers, polyphonic characters, or other words incorrectly, or making word order errors.

Attention: These pronunciation errors can occur at any point in the audio. For example:

- At the beginning of the audio, words are spoken that are not in the target text.
- At the end of the audio, some content from the target text is omitted.
- In the middle of the audio, omissions, insertions, or substitutions occur.
- ...

B. NATURALNESS COMPARISON

Natural speech should sound like a real person talking. Specific criteria include:

- The audio is clear, free from robotic/electronic tones or obvious noise (e.g., unnatural laughter, shouting, or other irrelevant background voices).
- The intonation is natural and expressive, not flat or mechanical.
- The speaking rhythm is reasonable, with moderate speed and appropriate pauses. (Note: Inappropriate pauses affect the naturalness of the audio but are not classified as "pronunciation errors").
- Word stress is placed correctly, conforms to normal linguistic sense, and does not sound abrupt.

The rating scale is as follows:

- **A +2:** Audio A is significantly more natural than B (large difference).
- **A +1:** Audio A is slightly more natural than B (slight difference).
- **Tie:** The naturalness of the two audio clips is similar and difficult to judge.
- **B +1:** Audio B is slightly more natural than A (slight difference).
- **B +2:** Audio B is significantly more natural than A (large difference).

Attention:

- If there are minor errors in individual words, please do not let them affect your overall judgment of naturalness.
- However, if a large number of content errors are found that severely interfere with your listening experience, this will affect the audio's naturalness.

C.1 INDIVIDUAL ANNOTATOR RELIABILITY

To assess the reliability of individual annotators, we computed agreement rate for each participant. This rate measures the extent to which an annotator's judgments align with those of their peers on the same sample (t, a_1, a_2).

For a given sample annotated by a group of M annotators, the agreement score for annotator i is calculated as the fraction of the other $M - 1$ annotators who assigned the exact same label. An annotator's final reliability score is the average of these scores across all samples they evaluated. We excluded participants who annotated fewer than 10 samples from this analysis.

Formally, for an annotator i who labeled N_i samples, the agreement rate r_{ij} for sample j is defined as:

$$r_{ij} \;=\; \frac{1}{M-1} \sum_{x \neq i} \mathbb{I}[y_{xj} = y_{ij}]$$

The overall agreement rate for annotator i, denoted as R_i, is then:

$$R_i \;=\; \frac{1}{N_i} \sum_{j=1}^{N_i} r_{ij}$$

where $y_{ij} \in \{A, B, T\}$ is the label assigned by annotator i to sample j. The label T (i.e., Tie) is treated as a distinct category, and it's agreement is counted only on exact matches.

Figure 9 illustrates the distribution of these agreement rates for our 69 annotators for SpeechJudge-Data (raw). The distribution is generally unimodal with a peak in the 60–70% range[3], which indicates a consistent and reliable level of performance across the annotation pool.

[3]We have noted that one annotator's agreement with the others is less than 30%, so we ultimately removed his data from SpeechJudge-Data.

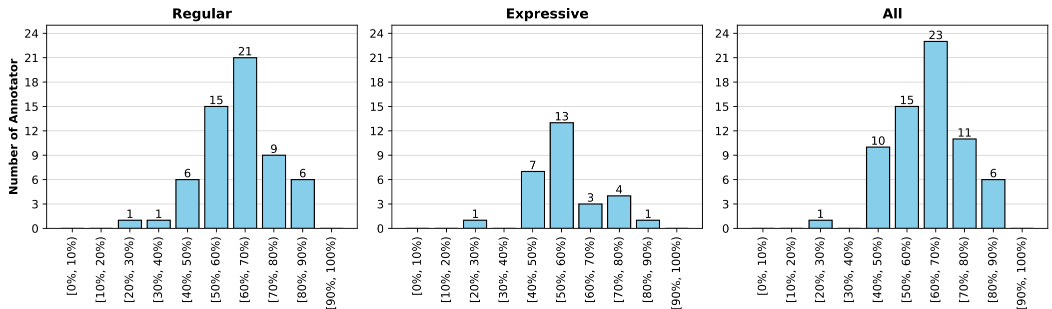


Figure 9: Distribution of individual annotator agreement rates. Most annotators fall between 50% and 80%, typically peaking in the 60–70% range.

C.2 Inter-Annotator Agreement

Complementary to the per-annotator reliability analysis in Appendix C.1, we now quantify the overall level of agreement among annotators at the dataset level. Following common practice in RLHF-style preference datasets, we compute inter-annotator agreement as the probability that two annotators chosen at random assign the same preference label to the same pair (Stiennon et al., 2020; Ouyang et al., 2022; Xu et al., 2023). Table 5 summarizes the results.

Ternary vs. binary preferences. SpeechJudge-Data (raw) contains ternary labels (i.e., A better / B better / Tie), and the corresponding inter-annotator agreement is $50.7\% \pm 0.1\%$ over the whole corpus. After removing "Tie" cases and restricting to clear binary preferences (SpeechJudge-Data (pref), labels A better / B better only), the agreement rises to $69.0\% \pm 0.2\%$. This level is comparable to well-established RLHF datasets in text and vision: ImageReward reports an agreement of $65.3\% \pm 1.6\%$ (Xu et al., 2023), and InstructGPT reports $72.6\% \pm 1.5\%$ on binary A/B comparisons (Ouyang et al., 2022). These numbers indicate that, despite the inherent subjectivity of the task, our binary preference data are on par with existing human-feedback corpora.

Effect of style. Within each language, expressive prompts are slightly harder than regular ones: for example, on SpeechJudge-Data (pref), regular *zh* reaches $74.4\% \pm 0.2\%$ agreement while expressive *zh* reaches $73.0\% \pm 0.6\%$. A similar, mildly lower agreement is observed for expressive *en* and *mixed* subsets. Overall, however, regular vs. expressive speech remains in a similar agreement range, suggesting that our guidelines allow annotators to make consistent naturalness judgments.

Effect of language and code-switching. A more pronounced pattern emerges across languages. In both the raw ternary data and the binary preference subset, agreement on Chinese (*zh*) is noticeably higher than on English (*en*) and especially on code-switched (*mixed*) samples (e.g., 74.4% vs. 61.7% and 62.5% in SpeechJudge-Data (pref)). We attribute this gap to the annotator population: our annotators are predominantly native Mandarin speakers; while English and code-switched items are assigned only to annotators who pass a high English proficiency bar, they are still L2 or bilingual listeners. Similar phenomena—lower agreement for L2 or cross-lingual annotations compared to L1 ones—have also been discussed in recent text RLHF studies (Bai et al., 2022).

In the current work, we therefore interpret SpeechJudge-Data as primarily reflecting the preferences of *Chinese and Chinese–English bilingual* listeners. In future work, we plan to augment the corpus by recruiting native English speakers for the English subset and stronger bilingual/native speakers for code-switched data, so as to improve agreement on these subsets and broaden the cultural and linguistic coverage of the dataset.

C.3 Intelligibility Annotation Analysis

We provide a detailed analysis of the relationship between the mostly common used objective intelligibility metric, Word Error Rate (WER), and the subjective human judgments of intelligibility. Our goal is to determine the extent to which WER can serve as a reliable proxy for human perception.

We use all the speech samples from SpeechJudge-Data (raw) for this analysis. We visualize the relationship between WER and the subjective text accuracy in Figure 10. For the regular speeches

Table 5: Inter-annotator agreement (mean ± std, in %) on SpeechJudge-Data and prior human preference datasets in the filed of RLHF.

Dataset	Preference	Regular		Expressive			Total
		en	*zh*	*en*	*zh*	*mixed*	
SpeechJudge-Data (raw)	Ternary	54.9 $_{\pm 0.3\%}$	55.5 $_{\pm 0.2\%}$	50.2 $_{\pm 0.7\%}$	49.5 $_{\pm 0.4\%}$	36.3 $_{\pm 0.3\%}$	50.7 $_{\pm 0.1\%}$
SpeechJudge-Data (pref)	Binary	61.7 $_{\pm 0.3\%}$	74.4 $_{\pm 0.2\%}$	59.9 $_{\pm 0.8\%}$	73.0 $_{\pm 0.6\%}$	62.5 $_{\pm 0.5\%}$	69.0 $_{\pm 0.2\%}$
ImageReward (Xu et al., 2023)	Binary	–	–	–	–	–	65.3 $_{\pm 5.6\%}$
InstructGPT (Ouyang et al., 2022)	Binary	–	–	–	–	–	72.6 $_{\pm 1.5\%}$

* SpeechJudge-Data (raw) uses ternary labels (A better / B better / Tie); SpeechJudge-Data (pref) removes ties and uses binary labels (A better / B better). Results of SpeechJudge-Data are further broken down by style and language.

(the orange curve), we observe a consistent negative correlation: as the WER increases, its perceived text accuracy steadily declines. For the expressive speeches (the green curve), the similar trend holds for expressive speech when WER is under about 12%. When WER is over the threshold, however, the correlation between WER and the subjective text accuracy weakens significantly. We think this divergence is sourced from that the greater stylistic variations in expressive speech pose a substantial challenge to the robustness of ASR systems compared to the regular samples.

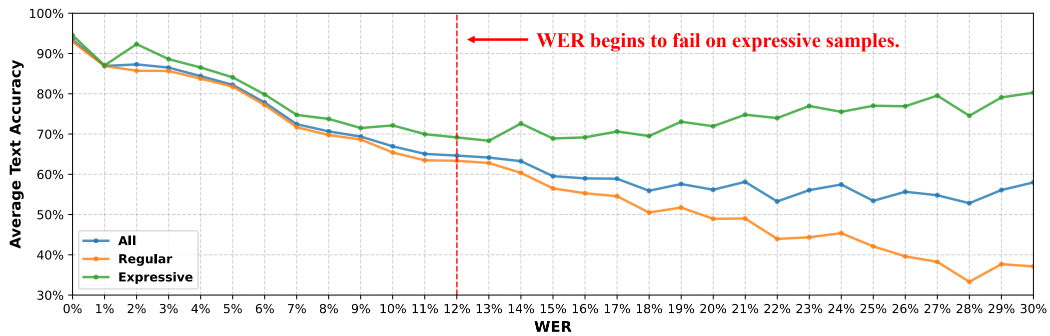


Figure 10: The relationship between the human-annotated text accuracy and WER.

D Details of Evaluation on the SpeechJudge-Eval Benchmark

During the evaluation on the SpeechJudge-Eval Benchmark of Table 2, we adopt the following protocol for each model:

- **WER** (Radford et al., 2023; Gao et al., 2023): We employ `Whisper-large-v3`[4] (Radford et al., 2023) for English texts, and `Paraformer-zh`[5] (Gao et al., 2022, 2023) for Chinese and code-switching texts.
- **SIM** (Chen et al., 2022): We compute the cosine similarity between the WavLM TDNN[6] (Chen et al., 2022) speaker embeddings of generated samples and the prompt samples.
- **FAD** (Kilgour et al., 2019): We use the officially released checkpoint, `VGGish`, to obtain the FADs of audios.
- **DNSMOS** (Reddy et al., 2022): We use the officially released script[7] to calculate the DNSMOS of audios.
- **CE, CU, PC, and PQ** (Tjandra et al., 2025): We use the officially released toolkit[8] to get the predicted MOS for audios.

[4]https://huggingface.co/openai/whisper-large-v3
[5]https://huggingface.co/funasr/paraformer-zh
[6]https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification
[7]https://github.com/microsoft/DNS-Challenge/blob/master/DNSMOS/dnsmos_local.py
[8]https://github.com/facebookresearch/audiobox-aesthetics

- **AASIST** (Jung et al., 2022): It is a common baseline model for audio deepfake detection, employs a heterogeneity-aware approach to integrate spectral and temporal sub-graphs. We use a large-scale in-house corpus to train the model.
- **ADV** (Wang et al., 2025b): It is a state-of-the-art (SOTA) deepfake detection model built upon the pre-trained `w2v-bert-2.0`[9]. It utilizes a multi-task training approach involving deepfake source tracing to extract robust audio deepfake features. We use the same corpus of AASIST to train the model.
- **AudioLLMs**: We use the plain prompt of Table 1 to instruct AudioLLMs to pairwise score the naturalness of two audios. For the closed-source models, we use the official API released by Google[10] for Gemini and OpenAI[11] for GPT. We use the model variants `gemini-2.5-flash`, `gemini-2.5-pro`, `gpt-4o-mini-audio-preview-2024-12-17`, and `gpt-4o-audio-preview-2025-06-03` for Gemini-2.5-Flash, Gemini-2.5-Pro, GPT-4o mini Audio, and GPT-4o Audio.

E MORE EVALUATION RESULTS OF EXISTING AUDIOLLMS

Using AudioLLM as a judge models, prompt engineering strategies are usually believed crucial for improving the performance (Zheng et al., 2023; Manakul et al., 2025). Some common prompt engineering strategies include using the CoT prompts to activate the model's thinking and reasoning abilities (Zhang et al., 2025a; Liu et al., 2025c; Manakul et al., 2025), or employing few-shot evaluation formats (Zheng et al., 2023; Xiaomi, 2025).

E.1 CHAIN-OF-THOUGHT PROMPTING FOR AUDIOLLM JUDGES

In this study, we investigate whether using the CoT from Table 1 helps AudioLLMs better judge speech naturalness. We display the results in Table 6. Interestingly, we find that some closed-source AudioLLMs, such as Gemini-2.5-Flash, improve their performance on SpeechJudge-Eval through this thinking and reasoning process. However, this strategy often does not work for existing open-source AudioLLMs. For example, the results in Table 6 show that Qwen2.5-Omni-7B and Kimi-Audio-7B-Instruct, which is already the leading open-source models on SpeechJudge-Eval (Table 2), actually sees a decline in performance when using the CoT prompt.

Table 6: Performance of AudioLLMs on SpeechJudge-Eval when using the CoT prompt of Table 1.

Model	Regular	Expressive	Total
Qwen2.5-Omni-7B	62.0	59.7	60.6
w/ CoT prompt	54.4	50.6	52.9
Kimi-Audio-7B-Instruct	65.5	68.0	67.0
w/ CoT prompt	67.4	66.1	66.5
Gemini-2.5-Flash	73.5	66.2	69.1
w/ CoT prompt	75.0	67.5	70.5

Based on our preliminary qualitative analysis, we believe the reason why the open-source AudioLLMs do not work well with the CoT prompt is that their foundational capabilities are relatively weak. These weaknesses include instruction-following (such as format-following), multiple-audio understanding, long-context processing, and reasoning abilities. This is also why, when we developed SpeechJudge-GRM, we did not directly apply RLVR on top of Qwen2.5-Omni-7B. Instead, we used an initial SFT stage as a cold start.

[9]https://huggingface.co/facebook/w2v-bert-2.0
[10]https://ai.google.dev/gemini-api/docs/models
[11]https://platform.openai.com/docs/models

E.2 Few-Shot Prompting for AudioLLM Judges

Motivated by the common belief that in-context examples can improve LLM judging ability (Zheng et al., 2023; Manakul et al., 2025), we also investigate *few-shot prompting* for Qwen2.5-Omni-7B on SpeechJudge-Eval.

Table 7: Performance of Qwen2.5-Omni-7B on SpeechJudge-Eval with different k-shot prompts.

Model	Regular	Expressive	Total
Qwen2.5-Omni-7B (0-shot)	62.0	59.7	60.6
w/ 2-shot	50.9	46.6	48.2
w/ 4-shot	46.1	52.0	49.6
w/ 6-shot	50.3	53.0	51.9
w/ 8-shot	51.0	54.8	53.3
w/ 16-shot	48.6	53.0	51.3

We start from the *plain* zero-shot prompt in Table 1, which asks the model to decide which of two audios is more natural. For the k-shot setting, we prepend k preference exemplars to this prompt. Each exemplar consists of: (i) a target text, (ii) an audio pair associated with this text, and (iii) the corresponding human naturalness label (which of the two audios is preferred). The model is then queried on a new SpeechJudge-Eval pair with the same instruction and output format. We evaluate Qwen2.5-Omni-7B with $k \in \{2, 4, 6, 8, 16\}$; results are reported in Table 7.

Contrary to the usual expectation that few-shot prompting should help, we observe that none of the k-shot configurations improves over the zero-shot baseline. On the contrary, the overall accuracy drops from 60.6% in the 0-shot setting to 48–53% with few-shot prompts. The degradation is particularly pronounced on regular speech, while expressive speech shows small, inconsistent fluctuations.

These findings are consistent with our observations in Appendix E.1 on chain-of-thought prompting: current open-source AudioLLMs such as Qwen2.5-Omni-7B still have limited instruction-following, multi-audio understanding, and long-context handling capabilities. Although few-shot prompts provide more information in principle, the model struggles to reliably associate multiple text-audio pairs with their labels in the context. As a result, increasing k mainly adds complexity to the input without yielding better judgments. This further motivates our choice to move beyond pure prompt engineering, and instead train dedicated reward models (BTRM / GRM) on human preference data for robust naturalness evaluation.

F Training Details of SpeechJudge-GRM

SFT Stage We use Gemini-2.5-Flash (Comanici et al., 2025) to generate the CoT data for SpeechJudge-Data (train). For the total 42K samples, Gemini-2.5-Flash's judgments agree with human feedback on 25K samples, while they disagree on 17K samples. During the SFT stage, we fine-tune Qwen2.5-Omni-7B (Thinker) (Xu et al., 2025) on the 25K CoT data using LoRA (Hu et al., 2022) with a rank of 128. We use Adam (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) as the optimizer and set the learning rate to 5e-5. The maximum number of tokens per batch is 4000. We select the best checkpoint on SpeechJudge-Data (dev) as the SFT model, SpeechJudge-GRM (SFT).

RL Stage We use the 17K samples (as described above) to conduct DAPO (Yu et al., 2025), which is an enhanced variant of GRPO (Shao et al., 2024). We utilize the `ms-swift`[12] toolkit to launch the training process. We initialize the policy model with the SFT model and use LoRA training with a rank of 64. The number of rollouts for each prompt is set to 8, and the batch size is 32. The learning rate is 5e-6. We select the best checkpoint on SpeechJudge-Data (dev) as the final SpeechJudge-GRM model, i.e., SpeechJudge-GRM (SFT+RL).

[12]https://github.com/modelscope/ms-swift

Table 8: Accuracy of different models on SpeechJudge-Eval (regular) across language settings.

Model	Regular				
	en2en	*zh2en*	*zh2zh*	*en2zh*	**Avg**
Qwen2.5-Omni-7B	48.1	58.5	75.7	66.0	61.0
Gemini-2.5-Flash	59.4	62.8	81.6	87.6	72.8
SpeechJudge-BTRM	66.0	71.3	86.4	86.6	77.5
SpeechJudge-GRM (SFT)	67.0	74.5	84.5	85.6	77.8
w/ Voting@10	65.1	75.5	83.5	85.6	77.3
SpeechJudge-GRM (SFT+RL)	69.8	77.7	85.4	83.5	79.0
w/ Voting@10	75.5	79.8	80.6	86.6	80.5

Table 9: Accuracy of different models on SpeechJudge-Eval (expressive) across language settings.

Model	Expressive						
	en2en	*zh2en*	*zh2zh*	*en2zh*	*en2mixed*	*zh2mixed*	**Avg**
Qwen2.5-Omni-7B	43.6	51.7	61.1	70.9	64.5	69.6	59.7
Gemini-2.5-Flash	53.6	68.3	73.3	76.4	64.5	67.1	66.2
SpeechJudge-BTRM	62.9	56.7	72.2	85.5	68.6	67.1	69.5
SpeechJudge-GRM (SFT)	61.4	66.7	89.1	77.8	74.4	73.4	73.7
w/ Voting@10	69.3	75.0	88.9	90.9	71.1	74.7	77.8
SpeechJudge-GRM (SFT+RL)	71.4	65.0	81.1	86.4	70.2	81.0	76.0
w/ Voting@10	75.0	66.7	82.2	89.1	72.7	84.8	78.7

G More Evaluation Results for SpeechJudge-GRM

G.1 Performance under Different Data Distributions

In Table 2 of the main paper, we reported accuracies on SpeechJudge-Eval aggregated over all languages and styles. Here we provide a more fine-grained view of how different judges behave under different data distributions, and we additionally evaluate on a completely out-of-distribution (OOD) test set involving real human speech versus commercial TTS clones.

Languages on the regular subset. Table 8 breaks down performance on the *regular* subset of SpeechJudge-Eval by language setting: *en2en*, *zh2en*, *zh2zh*, and *en2zh*. Across all models we observe that pairs involving Chinese (*zh2zh* and *en2zh*) are consistently easier than purely English pairs (*en2en*). For example, SpeechJudge-GRM (SFT+RL) reaches 85.4/83.5% accuracy on *zh2zh*/*en2zh*, but only 69.8% on *en2en*, and the same trend holds for BTRM and Gemini-2.5-Flash. We believe several factors contribute to this gap. On the data side, as shown in Appendix C.2, Chinese (*zh*) subsets exhibit higher inter-annotator agreement than English and mixed subsets, so supervision for English-like conditions is more varied and harder to fit. On the modeling side, current TTS systems may also produce relatively high-quality English outputs compared to Chinese, making the naturalness differences between two English samples more subtle and thus more difficult to judge reliably. Even under these challenges, SpeechJudge-GRM (SFT+RL) with Voting@10 still achieves the strongest average accuracy (80.5%) among all open judges.

Languages on the expressive subset. Table 9 shows the same breakdown for the *expressive* subset. As expected, expressive speech is generally harder: all models are a few points lower than on their regular counterparts. The language pattern persists: Chinese-involving settings (*zh2zh*, *zh2mixed*) tend to be easier than *en2en*. For instance, SpeechJudge-GRM (SFT+RL) with Voting@10 attains 82.2% on *zh2zh* and 84.8% on *zh2mixed*, compared to 75.0% on *en2en*. This is consistent with the inter-annotator statistics in Appendix C.2, where expressive and code-switched English subsets show lower human agreement, and also reflects the increased linguistic and prosodic complexity of expressive and code-switched speech. Overall, expressive data are slightly more challenging than regular data, but the relative ranking of judges is stable and GRM maintains a clear advantage over BTRM and the AudioLLM baselines across all language settings.

Table 10: Accuracy of different models on SpeechJudge-Eval across prompt styles.

Model	Regular	Emotional	Accented	Whisper	Game
Qwen2.5-Omni-7B	61.0	56.7	64.4	57.1	61.3
Gemini-2.5-Flash	72.8	63.7	66.7	74.6	66.0
SpeechJudge-BTRM	77.5	69.8	71.3	76.2	66.8
SpeechJudge-GRM (SFT)	77.8	75.3	80.5	71.4	70.2
w/ Voting@10	77.3	76.7	80.5	74.6	78.7
SpeechJudge-GRM (SFT+RL)	79.0	74.4	81.6	79.4	74.5
w/ Voting@10	80.5	78.1	85.1	82.5	75.7

Table 11: Accuracy of different models on the out-of-distribution **human vs. SeedTTS clone** evaluation set. Each test pair consists of a human recording and a SeedTTS voice clone for the same sentence; the model must decide which of the two is more natural. In all pairs, the human recording is treated as the ground-truth more natural sample.

Model	Character1	Character2	Avg
Deepfake Detectors			
AASIST	97.2	100	98.6
ADV	99.6	100	99.8
AudioLLMs			
Qwen2.5-Omni-7B	48.0	44.8	46.4
Kimi-Audio-7B-Instruct	85.2	85.6	85.4
Gemini-2.5-Flash	52.8	48.8	50.8
Naturalness Reward Model			
SpeechJudge-BTRM	55.6	45.2	50.4
SpeechJudge-GRM (SFT)	37.6	44.0	40.8
w/ Voting@10	36.0	41.4	38.7
SpeechJudge-GRM (SFT+RL)	57.6	67.2	62.4
w/ Voting@10	59.8	67.5	63.7

Different prompt styles. Table 10 compares performance across prompt styles: regular, emotional, accented, whisper, and game-character speech. For SpeechJudge-BTRM we see that regular prompts are the easiest (77.5%), while emotional and game styles are notably harder (69.8% and 66.8%). In contrast, SpeechJudge-GRM benefits more from the expressive settings: with SFT+RL and Voting@10, GRM achieves 80.5% on regular, but rises to 85.1% on accented and 82.5% on whisper prompts, and narrows the gap on emotional and game prompts (from about 10 points for BTRM to only a few points). We believe this reflects the advantage of the generative reward model and its CoT-based training: by explicitly reasoning about artifacts, prosody, and style, GRM is better able to handle challenging conditions such as accented speech and whisper, rather than overfitting to the most frequent regular style.

OOD evaluation on real speech vs. commercial TTS clones. Finally, Table 11 presents a new, fully out-of-distribution evaluation designed to stress-test generalization to *real* speech and unseen TTS architectures. We select two native English voice actors, each recording 250 utterances (500 in total), and use a commercial SeedTTS voice-cloning API[13] to synthesize clones for each utterance. SeedTTS is a state-of-the-art proprietary system whose output quality is typically higher than that of the open-source TTS models used to construct SpeechJudge-Data, so this benchmark effectively probes the gap between very strong modern TTS and human recordings. For every sentence, we form a pair (human recording vs. SeedTTS clone) and treat the human recording as the ground-truth more natural sample. Neither the human recordings nor the SeedTTS outputs are present in SpeechJudge-Data, making this a challenging OOD test. The results show three interesting trends:

- First, deepfake detectors (AASIST and ADV) achieve almost perfect accuracy (about 99%) on this task, even though they perform at chance level on SpeechJudge-Eval (Table 2), confirming that they mainly learn to discriminate real vs. synthetic rather than judge naturalness between two synthetic samples.

[13]https://www.volcengine.com/product/voicecloning

- Second, among AudioLLMs, Kimi-Audio-7B-Instruct performs strongly (85.4% on average), possibly because its training includes tasks or signals related to authenticity detection. In contrast, Gemini-2.5-Flash attains only 50.8% accuracy, roughly at chance level and similar to Qwen2.5-Omni-7B (46.4%). This indicates that Gemini's strong performance on SpeechJudge-Eval (69.1% in Table 2) does not automatically transfer to the human–vs–clone setting: it appears to be a good judge for *synthetic-vs-synthetic* naturalness comparisons, but it is not explicitly biased toward humans when facing a high-quality commercial clone.
- Third, for naturalness reward models, SpeechJudge-BTRM is close to random guessing (50.4%), suggesting that the classical Bradley-Terry training may overfit more heavily to the specific synthetic generators in SpeechJudge-Data. Interestingly, SpeechJudge-GRM (SFT) alone performs even worse than BTRM, which we hypothesize is due to SFT encouraging the model to over-memorize our prepared CoT patterns and thus hurting OOD generalization (Li et al., 2025b; Chu et al., 2025). Once we add the RLVR stage, however, SpeechJudge-GRM (SFT+RL) improves substantially to 62.4% (and 63.7% with Voting@10), outperforming BTRM as well as generic AudioLLMs such as Qwen2.5-Omni-7B and Gemini-2.5-Flash on this benchmark. Given that the SeedTTS clones are already very close to human quality, this performance indicates that SpeechJudge-GRM has the potential to provide useful feedback not only for open-source TTS models, but also for strong proprietary systems that were never seen during training. This suggests that the generative reward modeling paradigm is more robust to distribution shift than classical BTRM and off-the-shelf AudioLLM judges, and that RL on human preferences is crucial for recovering and enhancing generalization beyond the synthetic training distribution.

G.2 Quality Analysis of Chain-of-Thought Reasoning

Beyond scalar accuracy, we further analyze the *quality* of the Chain-of-Thought (CoT) rationales produced by different judges. Specifically, we compare Gemini-2.5-Flash (teacher), SpeechJudge-GRM (SFT), and SpeechJudge-GRM (SFT+RL) along three aspects: (i) logical consistency between reasoning and conclusion, (ii) faithfulness and hallucination rate as judged by human experts, and (iii) differences in reasoning style.

Consistency between reasoning and conclusion. We first ask whether a model's CoT reasoning is logically compatible with its final decision. Using DeepSeek-V3 (DeepSeek-AI et al., 2024) as a meta-judge, we prompt it with the instruction like: given the CoT analysis of A and B (over prosody, pacing, articulation, and overall naturalness) and the final scores assigned to A and B, decide whether the conclusion is consistent with the reasoning. DeepSeek-V3 returns a binary label (consistent / not consistent) and a brief justification. The specific instruction is as follows:

System Prompt:
I am having a model decide whether A or B is better using a Chain-of-Thought (CoT) process. Now, I need you to help me determine whether the model's CoT output—its reasoning process and its final conclusion (which is presented as "A: X points, B: Y points")—is consistent.
Please note:

- You only need to check whether, in the CoT reasoning, if the model analyzes that A is worse than B, then in the final scores A should also be lower than B. As long as this condition is met, you can regard it as consistent. Conversely, if the model reasons that A is better than B but assigns A a lower score than B in the final output, then it is not consistent.
- You only need to return a JSON string with two keys.
  - The first key is "result" with a value of 0 or 1—0 means not consistent, 1 means consistent.
  - The second key is "reason", where you briefly explain your reasoning in English.

User Prompt (Example):

Output A:

- **Prosody and Intonation:** The intonation is somewhat flat and lacks natural variability, making it sound a bit robotic. There are some unnatural stress patterns, and the overall prosody doesn't convey the meaning with human-like nuance.
- **Pacing and Rhythm:** The pacing is a bit slow and deliberate, and there's a slight unnatural pause before "They will frequently...". The rhythm feels somewhat monotonous.

Table 12: Consistency between Chain-of-Thought (CoT) reasoning and final conclusion for different models on SpeechJudge-Eval, as judged by DeepSeek-V3. A prediction is counted as consistent when the preference implied by the CoT analysis matches the model's final decision.

Model	Consistency
Gemini-2.5-Flash	97.9%
SpeechJudge-GRM (SFT)	97.6%
SpeechJudge-GRM (SFT + RL)	98.2%

Table 13: Human-rated quality of Chain-of-Thought (CoT) explanations for different models on SpeechJudge-Eval. Scores (1–3, ↑) are assigned by expert raters on three dimensions—Prosody & Intonation, Pacing & Rhythm, and Articulation & Clarity—matching the dimensions specified in our CoT prompt (Table 1). We report the mean and its 95% confidence interval for each dimension.

Model	Prosody & Intonation	Pacing & Rhythm	Articulation & Clarity	Avg
Gemini-2.5-Flash	1.90 $_{\pm 0.38}$	2.00 $_{\pm 0.44}$	2.10 $_{\pm 0.43}$	2.00 $_{\pm 0.34}$
SpeechJudge-GRM (SFT)	2.00 $_{\pm 0.39}$	2.15 $_{\pm 0.29}$	1.95 $_{\pm 0.42}$	2.03 $_{\pm 0.33}$
SpeechJudge-GRM (SFT+RL)	2.10 $_{\pm 0.20}$	2.00 $_{\pm 0.21}$	2.40 $_{\pm 0.32}$	2.17 $_{\pm 0.12}$

- **Articulation and Clarity:** While most words are clear, there is a very noticeable and significant mispronunciation of "will" as "woof" ("they woof frequently make a show"). This is a major error that severely impacts naturalness.
- **Overall Naturalness:** The "woof" mispronunciation is a critical flaw that immediately breaks the illusion of natural speech. Combined with the somewhat flat prosody, the output sounds distinctly synthetic and flawed.

Output B:

- **Prosody and Intonation:** The prosody is much more natural and varied, with appropriate pitch contours and stress patterns that align well with the meaning of the text. It sounds like a human speaker.
- **Pacing and Rhythm:** The pacing is excellent, flowing smoothly and naturally without any awkward pauses or rushed segments. The rhythm is consistent and pleasant to listen to.
- **Articulation and Clarity:** All words are articulated clearly and precisely. There are no noticeable mispronunciations, sibilance issues, or other artifacts.
- **Overall Naturalness:** This output sounds highly natural and human-like. The combination of good prosody, smooth pacing, and clear articulation results in a very convincing and pleasant listening experience.

Conclusion: Output A: 4, Output B: 8.5

System Output (Example):
{"result": 1, "reason": "In the reasoning, Output A is judged clearly worse than Output B (robotic prosody, major mispronunciation) while Output B is described as highly natural and human-like. The final scores reflect this ordering with A (4) $<$ B (8.5), so the conclusion is consistent with the analysis."}

Table 12 summarizes the results on CoT outputs for SpeechJudge-Eval across the three models. All three models exhibit very high internal consistency: Gemini-2.5-Flash, SpeechJudge-GRM (SFT), and SpeechJudge-GRM (SFT+RL) achieve 97.9%, 97.6%, and 98.2% consistency, respectively. This indicates that, at least at the coarse level captured by this automatic check, the CoT analyses are not arbitrary narratives but align well with the final naturalness preference.

Human evaluation of CoT faithfulness. Logical consistency does not guarantee that the reasoning is correct or grounded in the audio. To measure faithfulness and hallucination, we conduct a human evaluation with experienced speech researchers (the background of these subjects are detailed in Appendix H.1). For each model and each sampled SpeechJudge-Eval pair, the experts examine the CoT and assign a 1–3 score on three dimensions—(1) Prosody and Intonation, (2) Pacing and Rhythm, and (3) Articulation and Clarity—which match the dimensions specified in our CoT prompt (Table 1) and used by the models in their explanations. A score of 3 means "highly sensible (e.g., the CoT cites concrete audio details and the analysis is correct)", 2 means "partially sensible (e.g., the overall good/bad tendency is right but details are coarse or partially off)", and 1 means "not sensible / mostly hallucinatory".

The results are reported in Table 13. On average, all three models obtain scores around 2.0, indicating that their CoT rationales are generally meaningful rather than dominated by hallucination. Importantly, SpeechJudge-GRM does not lose CoT quality compared to the Gemini teacher even though it is initialized from Gemini-generated rationales: SpeechJudge-GRM (SFT) slightly improves the average score to 2.03 ± 0.34, and after RL, SpeechJudge-GRM (SFT+RL) further increases it to 2.17 ± 0.12. The largest gain appears in the "Articulation and Clarity" dimension (2.40 vs. 2.10 for Gemini), suggesting that RLVR encourages the model to focus more accurately on pronunciation errors and intelligibility-related artifacts when explaining its decisions. Overall, the human study suggests that the RL stage not only improves preference alignment but also mildly enhances the faithfulness of the CoT reasoning.

Differences in reasoning style. Finally, we examine whether the three models share the same reasoning style or develop distinct emphases. We randomly sample 20 SpeechJudge-Eval cases on which all three models predict the *correct* preference label, and collect their CoT outputs. We then submit these triplets of CoTs (anonymized as "Model 1/2/3") to three strong text LLMs—GPT 5.1[14], Gemini 3 Pro[15], and DeepSeek-V3[16], asking them to compare the similarities and systematic differences among the models. The specific instruction is as follows:

System Prompt:
I would like you to analyze the reasoning processes in the Chain-of-Thought (CoT) outputs of the following three models (referred to as Model 1, Model 2, and Model 3, respectively). What are the similarities and significant pattern differences among them? I will provide you with 20 cases, and under each case, the outputs of Model 1, Model 2, and Model 3 will be given.

User Prompt (Example):

******************** Case: 1/20 ********************

[Text] The rainbow serves as a metaphor for life's diversity - just as varying droplet sizes create different spectra, our unique experiences shape the breadth of our existence. The wider the band, the richer the experience.
[Model 1] ...
[Model 2] ...
[Model 3] ...

******************** Case: 2/20 ********************

[Text] They have committed to sending us the proposals to address our concerns by the end of January.
[Model 1] ...
[Model 2] ...
[Model 3] ...
...

The three analyzers broadly agree on the following qualitative picture: all models follow a similar structural template (prosody $\rightarrow$ pacing $\rightarrow$ articulation $\rightarrow$ overall naturalness, followed by a 1–10 score), but they differ in what they emphasize:

- **Gemini-2.5-Flash:** Described as the most sophisticated at linking prosody to semantic meaning and discourse structure, often explaining why an intonation pattern distorts the intended message.
- **SpeechJudge-GRM (SFT):** Viewed as emotionally and narratively focused, with stable formatting and slightly more generous tone; it emphasizes human-likeness and expressiveness but is somewhat less detailed on low-level signal artifacts.
- **SpeechJudge-GRM (SFT+RL):** Characterized as more critical and technically oriented: it pays more attention to mispronunciations, noise, and clarity, sometimes with blunter wording. Some analyzers note that its formatting is slightly less uniform than SpeechJudge-GRM (SFT), but its focus on error severity and technical correctness is stronger.

[14]https://chatgpt.com/
[15]https://aistudio.google.com/
[16]https://chat.deepseek.com/

Taken together, these analyses suggest that SpeechJudge-GRM does not simply imitate the teacher's explanation style. Instead, SFT initializes a shared analytical framework, while the RLVR stage shifts the model toward rationales that are more tightly coupled to human preferences and to concrete acoustic evidence (especially articulation and clarity), without sacrificing internal consistency. We will release the full set of CoT examples and meta-analyses in our open-source release to facilitate further research on CoT quality and reasoning in audio reward models.

G.3 Error Analysis of SpeechJudge-GRM

Although SpeechJudge-GRM performs well on SpeechJudge-Eval, it still disagrees with the human on a subset of it. We manually inspected these errors and found several recurring patterns.

Over-weighting cleanness vs. liveliness. In many errors the human-preferred sample contains mild background noise but clearly more human-like prosody and articulation, while the alternative is cleaner yet more robotic or over-smoothed. Annotators consistently favor the livelier sample, whereas GRM sometimes chooses the cleaner one, indicating that it can over-emphasize acoustic cleanness when the trade-off is subtle.

Prosody-articulation trade-offs. Another frequent pattern is a trade-off between expressive prosody and technical correctness. Humans often prefer speech with natural rhythm and intonation despite minor pronunciation issues, while GRM occasionally favors the perfectly articulated but flatter reading. These cases reveal that the relative weighting between prosody and articulation is still imperfectly captured.

Extreme expressive styles. Errors also concentrate in highly expressive styles. For emotional speech with very high F0 or strong emphasis, humans interpret the exaggerated prosody as appropriate for the style, but GRM sometimes penalizes it as "unnatural" and prefers a neutral reading. For whispers, the lack of voicing makes prosody judgments difficult; GRM occasionally fails to distinguish the more fluent whisper when both samples sound degraded.

Very small preference gaps. A small number of mistakes arise when both clips are high-quality and differ only in subtle cues (micro-pauses, breathing, slight emphasis shifts). In these cases GRM's predictions are effectively close to random, which is unsurprising given the weak supervision signal.

In summary, SpeechJudge-GRM's errors concentrate on nuanced trade-offs (clean vs. lively, prosody vs. articulation) and on challenging expressive styles, suggesting future work on modeling recording conditions, style-aware priors, and more targeted training examples.

H High-Quality Sample Selection and Post-Training based on SpeechJudge-GRM

H.1 Details of Subjective Evaluation

During the construction of SpeechJudge-Data, we hired human labelers from a data crowdsourcing company. To verify the effectiveness of our training for them and to ensure the high quality of both the dataset and the resulting SpeechJudge-GRM, the human subjects for the final sample selection and TTS post-training experiments (Section 5.3 and 5.4) were all experienced speech generation researchers. All these researchers had extensive audio backgrounds, with a minimum of two years of experience in speech synthesis.

We randomly selected the subjective evaluation samples from both SeedTTS-Eval (Anastassiou et al., 2024)[17] and Amphion-TTS-Eval (Zhang et al., 2025b,c)[18]. The evaluation set for each system in Figure 6 consists of 70 samples, while the set for each system in Figure 5 contains 100 samples. Each audio sample in these evaluations received at least three independent ratings. These subjective evaluation results show that the annotation quality of SpeechJudge-Data largely aligns with the judgments of professional researchers.

[17] https://github.com/BytedanceSpeech/seed-tts-eval

[18] https://huggingface.co/datasets/amphion/Amphion-TTS-Eval

Table 14: Post-training of Qwen2.5-0.5B-TTS based on SpeechJudge.

Model	Regular		Articulatory		Code-switching		Cross-lingual		Expressive		Avg	
	WER	**SIM**	**WER**	**SIM**	**WER**	**SIM**	**WER**	**SIM**	**WER**	**SIM**	**WER**	**SIM**
Qwen2.5-0.5B-TTS	2.63	0.698	10.53	0.679	23.87	0.666	10.51	0.593	11.10	0.706	11.73	0.668
w/ INTP	2.06	0.697	8.62	0.694	18.37	0.663	7.12	0.588	9.80	0.708	9.19	0.670
w/ SpeechJudge-Data	2.12	0.698	8.92	0.678	19.01	0.657	7.72	0.583	9.97	0.707	9.55	0.664
w/ SpeechJudge-GRM (offline)	2.31	0.698	7.83	0.681	15.36	0.662	7.84	0.593	9.72	0.709	8.51	0.668
w/ SpeechJudge-GRM (online)	2.35	0.696	8.45	0.674	15.87	0.653	7.82	0.580	9.79	0.702	8.85	0.661

H.2 Objective Results

We present the objective results (WER and SIM) of the Qwen2.5-0.5B-TTS post-training in Table 14. The results show that all four post-training methods significantly improve the WER. This trend is similar to the subjective intelligibility results shown in Figure 6a.

Regarding the SIM metric, both w/ INTP and w/ SpeechJudge-GRM (offline) either match or slightly outperform the baseline model, while the other two methods show a slight decline. However, the objective SIM results appear to be in slight conflict with the subjective speaker similarity results in Figure 6b. For instance, in the subjective evaluation, w/ INTP actually shows a decrease in speaker similarity (Win: 24.30%, Lose: 32.90%).

Through follow-up interviews with the subjects who participated in our subjective evaluation, we gathered additional qualitative insights. Participants consistently reported that the synthesized samples, both before and after post-training, demonstrated excellent speaker similarity, closely matching the reference speaker's timbre and style. In most cases, participants found it challenging to distinguish any significant differences in similarity, leading them to prefer selecting "Tie". For example, in Figure 6b, all four methods have the highest "Tie" proportion, each exceeding 40%. This demonstrates that post-training methods centered on naturalness (SpeechJudge-based) or intelligibility (INTP-based) are not yet fully aligned with speaker similarity, which requires further research into speaker similarity alignment.