# Comprehensive Benchmarking of Long-Form Speech Generation in Diverse Scenarios

Anonymous ACL submission

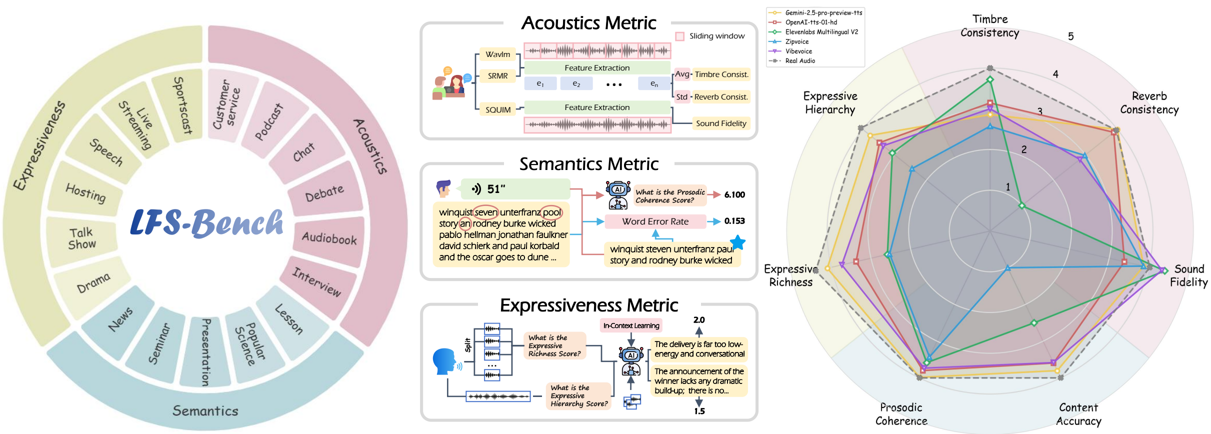



Figure 1: **Overview of LFS-Bench.** We propose LFS-Bench, a comprehensive benchmark designed to evaluate the performance of long-form speech generation models. ***Left:*** We construct test sets across 17 downstream speech scenarios, grounded in three core challenges of long-form generation: *Acoustics*, *Semantics*, and *Expressiveness*. ***Center:*** Along these three challenge axes, we propose seven disentangled metrics to comprehensively assess model performance and validate them through human alignment studies. ***Right:*** Extensive experiments show that existing models still have substantial room for improvement in reverb consistency, prosodic coherence, and expressiveness.

## Abstract

Recent advances in speech generation have enabled high-fidelity synthesis, yet systematic evaluation of models under long-context conditions remains largely underexplored. A comprehensive evaluation benchmark for long-form speech is indispensable for two reasons: 1) existing test scenarios are often confined to limited domains, creating a significant gap with the diverse downstream applications; 2) existing metrics overlook critical long-text factors such as consistency and coherence, failing to generalize reliably. To this end, we propose LFS-Bench, a comprehensive benchmark that decomposes “long-form speech quality” into specific, disentangled dimensions. LFS-Bench has three key properties. **1) Rich speech scenarios**: Focusing on long-form speech generation and dialog generation, LFS-Bench covers acoustics, semantics, and expressiveness challenges, and consists of 1,101 samples spanning 17 common speech scenarios; **2) Comprehensive evaluation dimensions**: Along the acoustics, semantics, and expressiveness axes, LFS-Bench defines an automated evaluation protocol with seven metrics to provide a comprehensive, accurate, and standardized assessment; **3) Valuable Insights**: Through extensive experiments, we reveal that current models still struggle in highly expressive scenarios and exhibit a notable gap in consistency and hierarchy compared to real recordings. The project page can be found at `https://lfs-bench.github.io`.

## 1 Introduction

Recent advances in generative modeling have revolutionized content creation across modalities (OpenAI, 2024; Esser et al., 2024; Guo et al., 2025). While Large Language Models (LLMs) have demonstrated impressive capabilities in long-context generation and understanding (Chen et al., 2023; Xiao et al., 2023; Bai et al., 2024), the speech community is similarly shifting focus from sentence-level to paragraph-level synthesis (Le et al., 2023; Shen et al., 2024). Compared to traditional concatenation strategies, end-to-end long-form TTS paradigms promise superior acoustic and semantic consistency, leveraging broader contextual cues (Peng et al., 2025; Park et al., 2024).

Despite these advancements, the systematic evaluation of long-form speech remains a significant challenge. While downstream applications involve complex multi-speaker interactions and rich semantic contexts, existing test scenarios are often confined to limited domains or single-speaker settings (Koizumi et al., 2023; Zhang et al., 2022). This discrepancy prevents a thorough assessment of how models handle the rich challenges inherent in long-form generation, leaving their capabilities in complex scenarios largely underexplored.

Furthermore, establishing an effective evaluation protocol that is both scalable and accurate is equally difficult. Existing sentence-level metrics like Word Error Rate (WER)(Ali and Renals, 2018) have become saturated (Chen et al., 2024b) and correlate poorly with human perception in long-text contexts (Minixhofer et al., 2025). Although human listening tests are the gold standard, they are non-scalable and costly. Recently, MLLM-based evaluators have emerged (Chen et al., 2024a; Manku et al., 2025), yet they typically provide coarse-grained comparative judgments rather than quantitative metrics, often overlooking the property of consistency (Li et al., 2024). Consequently, the field lacks an automated protocol aligned with the fine-grained nuances of long-form generation.

To this end, we propose LFS-Bench, a comprehensive benchmark for long-form TTS models with three core properties: 1) **rich** scenarios, 2) **comprehensive** evaluation, and 3) **valuable** insights.

First, LFS-Bench is defined over two fundamental long-form TTS paradigms: long-form speech generation and dialog generation. Starting from three core dimensions of long-form speech, namely *acoustics*, *semantics*, and *expressiveness*, LFS-Bench constructs 1,101 test samples spanning 17 downstream scenarios, providing broad coverage of long-form TTS applications.

Second, our framework establishes an automatic evaluation protocol that employs a hierarchical approach to decomposing "long-form speech quality". Transcending the traditional focus on Fidelity and Accuracy, we introduce novel dimensions tailored for long-form characteristics, specifically Acoustic Consistency, Prosodic Coherence, and Expressive Hierarchy. These metrics effectively address the limitations of existing protocols by quantifying temporal stability and expressive dynamics. Moreover, we conduct user studies to validate the reliability of these automated metrics, ensuring they serve as a scalable proxy for human perception.

Table 1: Comparison of speech generation benchmarks and test datasets. **Pipe.** indicates availability of an automatic evaluation pipeline, and ✗ marks that only part of the metrics are objectively computable. * denotes non-public data, with results estimated from the paper.

| Benchmark | Clips | Scenario | Spk-Num | Avg-Word | Pipe | Dim. |
|---|---|---|---|---|---|---|
| SeedTTS-Eval | 6612 | 1 | 1 | 19.57 | ✗ | 3 |
| EmergentTTS-Eval | 1645 | 6 | 1 | 33.93 | ✓ | 5 |
| TTSDS2 | 60 | 4 | 1 | 24.24 | ✓ | 4 |
| Choice of Voices | 1 | 1 | 1 | 988 | ✗ | 5 |
| MinutesSpeech-test | 1221 | 1 | 1 | 134 | ✗ | 6 |
| LibriSpeech-long | 960 | 1 | 1 | 534.5 | ✗ | 6 |
| NeuralTTS-eval | 250 | 1 | 1 | 260* | ✗ | 9 |
| MultiDialog | 831 | 3 | 2 | 319.8 | ✗ | 4 |
| **LFS-Bench** | 1101 | 17 | 1-4 | 228.6 | ✓ | 7 |

Finally, through extensive experiments on LFS-Bench, we derive critical insights detailed in Section 5. Our empirical results reveal that while current models rival human recordings in fidelity and accuracy, they exhibit substantial gaps in reverb consistency, prosodic coherence, and expressive hierarchy. Notably, performance deteriorates in highly expressive scenarios, underscoring the persisting challenges in modeling long-term dependencies and dynamic stylistic variations.

We're open-sourcing LFS-Bench, including test samples, and evaluation scripts with prompts. We'll also include more models in LFS-Bench to drive forward the field of long-form speech generation.

## 2 Related Work

**Long-form TTS** Generating long-form speech and dialogues presents significant challenges in maintaining prosodic coherence, modeling long sequences, and managing speaker transitions. To ensure prosodic consistency, recent studies have explored joint style modeling and cross-sentence memory mechanisms (Guo et al., 2024a; Li et al., 2025). Concurrently, to enhance long-sequence modeling efficiency, researchers have introduced compact representations via multi-resolution quantization (Nishimura et al., 2024) or low frame-rate tokenization (Peng et al., 2025), as well as state space models to alleviate memory bottlenecks (Park et al., 2024). Regarding speaker transitions, while early works combined autoregressive (AR) and non-autoregressive (NAR) components (Borsos et al., 2023), recent advancements have further developed both paradigms: NAR approaches increasingly employ flow-matching techniques, whereas AR models leverage speaker tokens to handle long-context dialogues (Ju et al., 2025; Xie et al., 2025a). Despite these technical strides, existing metrics remain insufficient for evaluating prosodic coherence, emotional richness, and transition quality. To bridge this gap, LFS-Bench introduces a unified evaluation framework with targeted test cases and human-aligned metrics designed to quantify these critical properties.

**Evaluation for Speech Generation Models** Current TTS evaluation mainly relies on four objective metric families: signal-based metrics (Taal et al., 2010), MOS prediction networks (Saeki et al., 2022), distributional metrics (Minixhofer et al., 2024), and accuracy metrics (Ali and Renals, 2018). These metrics are nearly saturated for recent state-of-the-art systems (Ju et al., 2024). Follow-up benchmarks (Huang et al., 2025; Anastassiou et al., 2024) increase difficulty via harder texts or controllability, but remain sentence level and are not directly suitable for long-form speech (Clark et al., 2019). Long text test sets like MinutesSpeech- (Nishimura et al., 2024) and LibriSpeech-Long (Park et al., 2024) partially address this gap, yet cover only a narrow range of scenarios, as shown in Table 1. Benchmarks for dialog models also face similar issues (Ao et al., 2024). Moreover, existing protocols rely heavily on subjective evaluations (Cambre et al., 2020; Zhang et al., 2023), which do not scale and lack standardized procedures. In contrast, LFS-Bench jointly covers long-form speech and dialog generation, spans 17 scenarios, and provides comprehensive automatic metrics aligned with humans, thereby addressing key limitations of current evaluation practices.

# 3 LFS-Bench

## 3.1 Overview

Long-form speech generation requires multi-dimensional evaluation to ensure immersion and realism. For instance, in an online education scenario, a generated lecture must not only preserve timbre and acoustic environment (acoustics) but also deliver accurate content with natural pacing (semantics), while exhibiting dynamic variations to sustain engagement (expressiveness). Motivated by these requirements, we propose LFS-Bench, a hierarchical benchmark comprising 1,101 samples across 17 downstream applications. And as detailed in Section 3.4, our evaluation protocol is organized around three primary dimensions:

**Acoustics Challenge** focuses on sound quality, environmental fidelity, and speaker identity. Hence, we carefully curate samples from six relevant scenarios: customer service, podcast, chat, debate, audiobook, and interview, and evaluate acoustic performance based on *timbre consistency*, *reverb consistency*, and *sound fidelity*.

**Semantics Challenge** targets correctness and fluency to probe the upper limits of semantic modeling. We derive complex test cases from five information-dense scenarios (lesson, popular science, presentation, seminar, and news), evaluating them by *content accuracy* and *prosodic coherence*.

**Expressiveness Challenge** addresses the issues of flat emotion and low engagement in long-form speech. We incorporate highly expressive scenarios such as drama, talk show, hosting, speech, live streaming, and sportscast. Performance is assessed through *expressive richness* (sentence-level emotional impact) and *expressive hierarchy* (paragraph-level expressive dynamics).

## 3.2 Data Collection

To provide a high-quality benchmark, we curate the test samples from three sources: online text corpora, online audio media, and LLM generation.

**Online Text Corpora** For scenarios such as audiobooks, drama, and news, where abundant transcripts are available online, we directly construct test sets from the web. After crawling the raw data, we clean irrelevant content such as illegal characters, and normalize the text into a clear and readable format. We then employ human annotators to proofread the transcripts and add speaker labels, yielding the final curated test samples.

**Online Audio Media** This source constitutes the main component of LFS-Bench. For web audio data, after crawling, we first denoise the raw audio (Wang and Tian, 2025), and then use DNS-MOS (Reddy et al., 2021) scores to filter out low-quality cases. After that, speaker diarization is conducted (Zheng et al., 2023) to obtain audio segments for each speaker. Finally, we use SenseVoice (An et al., 2024) to transcribe audio clips. Upon completion of the script processing, we perform manual verification to correct errors from the previous steps and curate the final test samples.

**LLM generation** We use GPT-5 (OpenAI, 2025) to augment our test set and increase the diversity of data sources. Specifically, we first design prompts that include scenario, topic, and task information. Then we use them to guide the LLM to generate high-quality test cases. All generated samples are then checked and verified by human annotators.

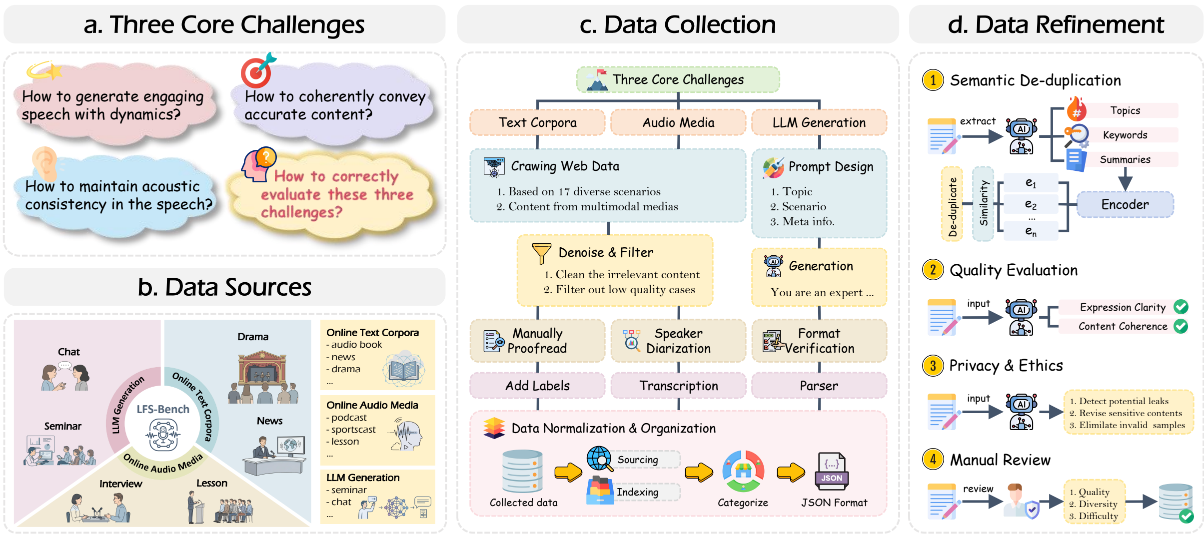



Figure 2: **Overview of dataset construction and refinement.** The process consists of four stages: 1) Formulating LFS-Bench based on three core challenges; 2) Selecting 17 downstream speech scenarios aligned with these challenges; 3) Designing a hybrid data collection pipeline; 4) Performing data refinement on the constructed dataset.

### 3.3 Data Refinement

To ensure the quality of curated samples, we implement a rigorous refinement pipeline. The process begins with semantic de-duplication, where we employ GPT-5 to extract topics, keywords, and summaries for each sample. These fields are concatenated and encoded using SentenceBERT (Reimers and Gurevych, 2019) to identify and remove highly similar instances based on cosine similarity. Subsequently, we filter for content quality by leveraging GPT-5 to evaluate expression clarity and content coherence, discarding any samples that fall below predefined thresholds. To address privacy and ethical concerns, we utilize DeepSeek V3.2 (Liu et al., 2024a) with a chain-of-thought (Wei et al., 2022) procedure to detect potential leaks, revise sensitive content, and eliminate samples posing social or ethical risks. Finally, we conduct a manual review to purge remaining low-quality samples and replenish the dataset, ultimately yielding 1,101 samples that cover three core challenges and span 17 downstream scenarios, as shown in the left side of Fig. 1.

### 3.4 Evaluation Metrics

We disentangle the challenges into seven objective metrics to comprehensively assess the performance of TTS models. More details in Appendix C.

**Timbre Consistency**. Compared with prior work that evaluates zero-shot capability using speaker similarity, we directly measure within-utterance timbre consistency to assess a model's ability to maintain or switch speaker identity. For single-speaker long-form speech $w$, we apply a sliding window over the waveform and extract a speaker embedding for each window, yielding a sequence $\{\mathbf{e}_i\}_{i=1}^{n}$, where $n$ is the number of windows. We then compute the cosine similarity for every pair of distinct embeddings and take the average of the resulting similarity sequence $\{\mathrm{sim}_{i,j}\}_{i,j=1,i\neq j}^{n}$ as the measure of timbre consistency. For dialog, we first use forced alignment (McAuliffe et al., 2017) to obtain segments of each speaker. The final metric is obtained by averaging the consistency scores of individual speakers.

**Reverb Consistency**. We assess whether synthesized audio maintains a stable acoustic environment by measuring the consistency of reverberation over time. For a generated utterance $w$, we apply a sliding window over the waveform and compute the speech-to-reverberation modulation energy ratio (SRMR) for each window, obtaining a sequence of reverberation scores $\{r_i\}_{i=1}^{n}$. We then compute the standard deviation of this sequence, which serves as our reverb consistency metric; lower variance indicates a more consistent reverberation pattern across the utterance.

**Sound Fidelity**. We evaluate the perceptual quality and clarity of the generated speech using the Perceptual Evaluation of Speech Quality (PESQ) metric. Given that standard PESQ requires a reference signal unavailable in our setting, we employ `SQUIM-PESQ` to perform non-intrusive, reference-free evaluation for the synthesized audio.

**Content Accuracy**. Faithful content rendering is a cornerstone of robust TTS systems. To investigate the impact of long-sequence modeling on content fidelity, we employ an ASR-based evaluation, calculating the Word Error Rate (Character Error Rate for Chinese) between the transcripts of the synthesized audio and the ground truth text.

**Prosodic Coherence**. While content accuracy ensures lexical correctness, prosodic coherence evaluates the naturalness of delivery. This metric focuses on pauses, speaking rate, and the consistency of overall prosody to capture the naturalness of generated speech. LFS-Bench leverages SpeechJudge (Zhang et al., 2025b), a scoring model fine-tuned from Qwen2.5-Omni-7B (Xu et al., 2025a). We refine the input prompt to strengthen the model's sensitivity to prosodic consistency in long-form contexts, utilizing the resulting scalar score (1–5) as our metric for coherence.

**Expressive Richness**. In long-form synthesis, expressiveness becomes crucial, as monotonous delivery fails to sustain user engagement or support immersive experiences. To address this need, LFS-Bench evaluates expressive richness along three dimensions: emotional resonance, character portrayal, and storytelling. Following EmergentTTS-Eval (Manku et al., 2025), we employ LALMs as evaluators using a comprehensive prompt to score audio on a 1–5 scale. To ensure fine-grained assessment, we segment inputs into 10-second intervals and calculate the average score across all segments.

**Expressive Hierarchy**. Beyond sentence-level expressiveness, paragraph-level expressive hierarchy is a defining characteristic of long-form speech. We employ LALMs to evaluate this attribute on a scale of 1 to 5, designing prompts that specifically target emotional variation, vocal dynamics, and scene appropriateness. Crucially, we evaluate the full utterance rather than via segmentation to preserve the integrity of the narrative flow.

### 3.5 Human Perception Alignment Test

To further validate the effectiveness of our evaluation protocol, we conduct a subjective assessment in which human raters score a randomly selected subset of the test data. Additional implementation details and results are provided in the Appendix D.

**Prosody Evaluation**. We randomly sample 50 pairs of audio clips, each synthesized from identical text by different models, and conduct a subjective preference test with 10 human evaluators. For each pair $(A, B)$, raters assess the comparative prosodic coherence on a 5-point scale ranging from -2 to 2. The human preference score is defined as:

$$\mathcal{S}_{\text{pref}}(A, B) = \frac{1}{N} \sum_{i=1}^{N} s_i, \tag{1}$$

where $s_i$ denotes the score assigned by the $i$-th rater, and $N$ represents the total number of raters. We compute the Spearman Rank Correlation Coefficient (SRCC) between human preference scores and the differential of our metric. The SRCC of 0.82 shows that our metric effectively captures the perceived prosodic coherence of long-form speech.

**Expressiveness Evaluation**. We randomly sample 200 audio clips across all models and tasks, recruiting 10 human evaluators to score each sample, strictly adhering to the same expressiveness prompts used for the LALM evaluation. In parallel, we benchmark three MOS prediction networks and six LALMs by computing the correlation between their predicted scores and the human Mean Opinion Scores (MOS). Finally, we select Gemini3-Pro as our primary evaluator, due to its highest alignment with human judgment, yielding SRCC scores of 0.71 for expressive richness and 0.62 for expressive hierarchy. We also validate the stability of Gemini 3 Pro through independent repeated trials. More results are detailed in Appendix D.4.

## 4 Experiments

### 4.1 Settings

**Model Evaluated** For single-speaker long-form speech, we evaluate ten open-source models: ZipVoice (Zhu et al., 2025b), Spark-TTS (Wang et al., 2025), CosyVoice2-0.5B (Du et al., 2024), CosyVoice3-0.5B (Du et al., 2025), GLM-TTS (Cui et al., 2025), MegaTTS3 (Jiang et al., 2025), IndexTTS2 (Zhou et al., 2025b), FishSpeech-1.5 (Liao et al., 2024), F5TTS (Chen et al., 2024c), and VibeVoice (Peng et al., 2025). And we evaluate six closed-source flagship systems: Gemini-2.5-pro-preview-tts, OpenAI-tts-1-hd, ElevenLabs Multilingual V2, Minimax-speech-02-hd (Zhang et al., 2025a), InWorld-TTS-1-max (Atamanenko et al., 2025), and Seed-TTS2 (Anastassiou et al., 2024). In the dialogue generation setting, we select six open-source models capable of long-form synthesis—ZipVoice-Dialog (Zhu et al., 2025a), MoonCast (Ju et al., 2025), MOSS-TTSD (Zhao et al., 2025), FireRedTTS2 (Xie et al., 2025b), VibeVoice, and SoulX-Podcast (Xie et al., 2025a)—and compare

Table 2: **Evaluation results of long-form TTS models across multi-dimensional metrics.** Metrics cover Acoustics (Timbre/Reverb Consistency, Fidelity), Semantics (Content Accuracy, Prosodic Coherence), and Expressiveness (Richness, Hierarchy). CER and WER apply to Chinese and English, respectively. Closed-source models and open-source models is separately marked, with the best results in **bold** and the second best underlined.

| Model | Acoustics | | | Semantics | | Expressiveness | |
|---|---|---|---|---|---|---|---|
| | Timbre(↑) | Reverb(↓) | Sound Fidelity(↑) | CER/WER(↓) | Prosody(↑) | Richness(↑) | Hierarchy(↑) |
| *Open-Source Models* | | | | | | | |
| CosyVoice-2 | 0.92±0.018 | 2.35±0.78 | 3.80±0.27 | **0.032** / 0.168 | 3.23±1.01 | 3.02±0.68 | 2.76±0.88 |
| CosyVoice-3 | **0.94±0.008** | 2.26±0.59 | 3.83±0.10 | 0.034 / 0.141 | 3.31±0.71 | 2.80±0.70 | 2.45±0.75 |
| FishSpeech | 0.93±0.014 | 1.79±0.65 | **4.10±0.09** | 0.043 / 0.113 | 3.80±0.86 | 2.66±0.78 | 2.90±0.74 |
| F5TTS | 0.90±0.022 | 1.82±0.77 | 3.39±0.33 | 0.072 / 0.113 | 3.41±0.99 | 3.07±0.63 | 2.77±0.84 |
| GLM-TTS | 0.94±0.010 | **1.62±0.61** | 3.95±0.13 | 0.035 / 0.118 | 3.64±0.87 | 2.68±0.71 | 2.54±0.88 |
| IndexTTS-2 | **0.94±0.008** | 1.72±0.53 | 2.77±0.41 | 0.033 / 0.135 | 3.64±0.52 | 3.59±0.72 | 2.96±0.81 |
| MegaTTS-3 | 0.93±0.008 | 1.81±0.45 | 3.55±0.19 | 0.035 / **0.108** | 3.61±0.84 | 2.81±0.55 | 2.53±0.63 |
| SparkTTS | 0.93±0.033 | 1.79±1.70 | 3.59±0.40 | 0.329 / 0.240 | 2.58±1.24 | 3.47±0.58 | 2.38±0.83 |
| VibeVoice | 0.93±0.024 | 2.15±0.88 | 3.82±0.42 | 0.047 / 0.111 | **3.90±0.79** | **3.71±0.58** | **3.34±0.88** |
| ZipVoice | 0.90±0.011 | 2.06±1.08 | 3.51±0.19 | 0.072 / 0.396 | 3.19±1.11 | 2.44±0.85 | 2.11±1.05 |
| Average | 0.93 | 1.95 | 3.63 | 0.073 / 0.164 | 3.43 | 3.03 | 2.67 |
| *Closed-Source models* | | | | | | | |
| Elevenlabs Multilingual V2 | **0.96±0.008** | 3.05±0.59 | **4.02±0.11** | 0.100 / 0.115 | 3.50±0.73 | 2.33±0.74 | 2.68±0.81 |
| Gemini-2.5-pro-preview-tts | 0.91±0.018 | 1.44±0.50 | 3.16±0.36 | 0.058 / 0.169 | 3.91±0.72 | **4.14±0.65** | **3.51±0.84** |
| Inworld-TTS-1-max | 0.93±0.025 | 2.19±0.64 | 3.73±0.17 | 0.053 / **0.113** | 3.71±0.51 | 3.68±0.86 | 3.03±0.92 |
| Minimax-Speech-02-hd | 0.93±0.010 | **1.38±0.35** | 3.82±0.09 | **0.032** / 0.119 | **3.95±0.73** | 3.80±0.44 | 3.26±0.79 |
| OpenAI-tts-01-hd | 0.92±0.011 | 1.74±0.42 | 2.68±0.12 | 0.043 / 0.119 | 3.91±0.52 | 3.46±0.62 | 3.25±0.81 |
| SeedTTS-2 | 0.94±0.022 | 1.95±0.74 | 3.88±0.18 | 0.106 / 0.193 | 3.74±0.44 | 3.10±0.80 | 2.34±0.65 |
| Average | 0.93 | 1.96 | 3.55 | 0.065 / 0.138 | 3.79 | 3.42 | 3.01 |
| **Real Speech** | 0.96 | 1.91 | 3.62 | 0.070 / 0.074 | 4.04 | 4.35 | 3.94 |

them with four closed-source baselines: Gemini-2.5-pro-preview-tts, OpenAI-tts-1-hd, ElevenLabs Multilingual V2, and SeedTTS-Podcast.

**Evaluation Models** For the timbre consistency evaluation, we use WavLM TDCNN[1] to extract speaker embeddings, and perform forced alignment with Paraformer[2] on Chinese data and WhisperX (Bain et al., 2023) on English data. For WER computation, we adopt FunASR Nano[3] as the transcription model. For all expressiveness-related metrics, we use Gemini3-pro (Google DeepMind, 2025) with prompt enhancement as the evaluator.

### 4.2 Evaluation from Different Perspectives

**Per-Dimension Evaluation** We demonstrate LFS-Bench scores across all dimensions following the evaluation protocol outlined in Section 3.4, with results summarized in Tables 2 and 3. Additionally, we incorporate two reference baselines: *Real Speech* and *Real Dialogue*, which are derived from the source dataset in Section 3.2, serving as the topological upper bound for audio quality.

**Per-Scenario Evaluation** We evaluate the long-form speech and dialog generation models across three core categories spanning 17 different scenarios, and then calculate their performance via the evaluation protocol. Fig. 3 visualizes the evaluation results of each model in terms of three categories.

**Evaluations On Generated Length** We evaluate five representative models (MegaTTS3, F5TTS, Cosyvoice2, SparkTTS, and VibeVoice) across increasing input lengths among 100 samples in three core scenarios (Acoustics, Semantics, and Expressiveness). The results are shown in Fig 4.

## 5 Insights and Discussions

### 5.1 Observations

**Gap to Ground-Truth Audio** As shown in Tables 2 and 3, among the evaluated systems, ***VibeVoice*** and ***SoulX-Podcast*** emerge as the strongest open-source models, while ***Minimax-Speech-02-hd*** and ***Gemini-2.5-pro-preview-tts*** lead their proprietary counterparts. We also observe that, although SOTA open-source models already match or even surpass the best proprietary systems on several evaluation dimensions, Proprietary models still exhibit consistently stronger overall performance than open-source models for long-form speech generation. However, benchmarking against real recordings reveals persistent and systematic gaps. For long-form synthesized speech, even the best-performing models remain below human speech in overall expressiveness: the closed-

[1] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

[2] https://modelscope.cn/models/iic/speech_timestamp_prediction-v1-16k-offline

[3] https://huggingface.co/FunAudioLLM/Fun-ASR-Nano-2512

Table 3: **Results of dialogue generation models across LFS-Bench's metrics**. The performance of closed-source models and open-source models is separately marked, with the best results in **bold** and the second best underlined.

| Model | Acoustics | | | Semantics | | Expressiveness | |
|---|---|---|---|---|---|---|---|
| | Timbre(↑) | Reverb(↓) | Sound Fidelity(↑) | CER/WER(↓) | Prosody(↑) | Richness(↑) | Hierarchy(↑) |
| *Open-Source Models* | | | | | | | |
| FireRedTTS-2 | 0.93±0.017 | 3.48±1.06 | 2.62±0.69 | 0.075 / 0.131 | 3.24±1.04 | 2.72±0.75 | 2.81±0.97 |
| MoonCast | 0.90±0.022 | **3.06±1.84** | 2.62±0.37 | 0.313 / 0.125 | 3.16±1.18 | 2.68±0.68 | 2.70±0.99 |
| MOSS-TTSD | 0.91±0.028 | 3.55±1.16 | 2.89±0.55 | 0.148 / 0.239 | 2.79±1.14 | 3.21±0.79 | 2.99±1.06 |
| SoulX-Podcast | **0.93±0.016** | 3.51±0.80 | **3.96±0.09** | **0.061 / 0.090** | **4.01±0.78** | 3.44±0.69 | **3.71±0.81** |
| VibeVoice | 0.91±0.028 | 3.59±0.85 | 3.35±0.72 | 0.106 / 0.125 | 3.57±1.05 | **3.76±0.63** | 3.37±0.83 |
| ZipVoice-Dialog | 0.91±0.021 | 3.53±0.85 | 2.66±0.24 | 0.069 / 0.114 | 3.67±0.89 | 2.62±0.60 | 2.80±0.88 |
| Average | 0.92 | 3.45 | 3.02 | 0.129 / 0.137 | 3.41 | 3.07 | 3.06 |
| *Closed-Source models* | | | | | | | |
| Elevenlabs Multilingual V2 | 0.93±0.016 | 4.43±1.01 | 3.48±0.44 | 0.127 / 0.109 | 3.67±0.78 | 2.84±0.79 | 3.46±0.87 |
| Gemini-2.5-pro-preview-tts | 0.92±0.017 | 3.17±0.68 | 3.01±0.24 | 0.086 / **0.092** | **4.06±0.39** | **4.06±0.48** | **4.02±0.68** |
| OpenAI-tts-1-hd | **0.93±0.013** | 2.98±0.63 | 2.28±0.17 | 0.104 / 0.103 | 3.69±0.62 | 3.29±0.75 | 3.70±0.88 |
| SeedTTS-Podcast | 0.91±0.017 | **2.85±0.78** | **3.89±0.17** | **0.063** / 0.108 | 3.93±0.46 | 3.84±0.72 | 3.84±0.88 |
| Average | 0.92 | 3.36 | 3.17 | 0.095 / 0.103 | 3.83 | 3.51 | 3.76 |
| **Real Dialogue** | 0.95 | 2.73 | 2.94 | 0.050 / 0.137 | 3.95 | 4.42 | 4.17 |

source average lags behind real speech by nearly one MOS point in richness and over half a point in hierarchy. A similar pattern holds in dialog scenarios, where closed-source systems obtain higher expressiveness, but still fall short of the natural expressivity implied by real dialogue. In acoustic metrics, synthesized speech approaches real recordings in Fidelity, but long-form outputs show a deficit in Timbre Consistency. For dialog generation, the marked gap in Reverb Consistency (3.36 vs. 2.73) underscores a core challenge: sustaining global acoustic consistency across multiple speakers. In terms of Semantics, current models achieve Content Accuracy comparable to real speech, demonstrating strong capability in pronunciation. Nevertheless, deficiencies in prosodic coherence persist, limiting the naturalness of the synthesized audio.

**Impact of Scenarios.** As illustrated in Figure 3, downstream scenarios significantly impact generation performance. *Acoustic challenge scenarios* present distinct difficulties, particularly in maintaining acoustic field consistency. This struggle likely stems from frequent speaker transitions that disrupt reverberation unity, also causing minor fidelity degradation. Notably, however, timbre consistency remains stable, demonstrating the robustness of current models in this dimension. For *semantic-dominated scenarios*, linguistic complexity in semantic-dominated scenarios does not compromise content accuracy, thanks to robust text normalization. However, it poses substantial challenges to prosody modeling, indicating a need for improved comprehension of intricate syntactic structures. An intriguing finding emerges *in expressiveness settings*. Here, all models exhibit performance degradation across nearly all metrics, particularly in Expressive Richness. Theoretically, these scenarios should represent a higher upper bound for expressiveness. Consequently, this counter-intuitive performance suggests that models may lack effective training on expressive data. Furthermore, it highlights the substantial gap remaining in achieving immersive and expressive generation. More data support, experimental results, and detailed analysis can be found in Appendix G.2.

## 5.2 Discussions

**AR v.s. NAR** In long-form TTS, the choice between AR and NAR paradigms centers on the trade-off between expressiveness and robustness. NAR models, leveraging parallel generation mechanisms, demonstrate superior robustness and efficiency in long-text synthesis (Ren et al., 2020). However, they tend to produce over-smoothed rhythms, often failing to capture the vocal dynamics and emotional nuances required for extended narration. As observed in Table 2 and 3, F5TTS, despite being the top-performing NAR model, lags significantly behind most AR counterparts in expressive hierarchy. Similarly, ZipVoice-Dialog ranks among the lowest in expressiveness within the dialogue category. Conversely, AR models, typically built upon language model backbones, excel in prosody modeling but suffer from error propagation in long-form scenarios. While they achieve superior expressiveness, they exhibit a lower bound on Content Accuracy; for instance, both SparkTTS and MoonCast show suboptimal performance in this dimension. Furthermore, as illustrated in Figure 4, SparkTTS suffers from a substantial decline in con-

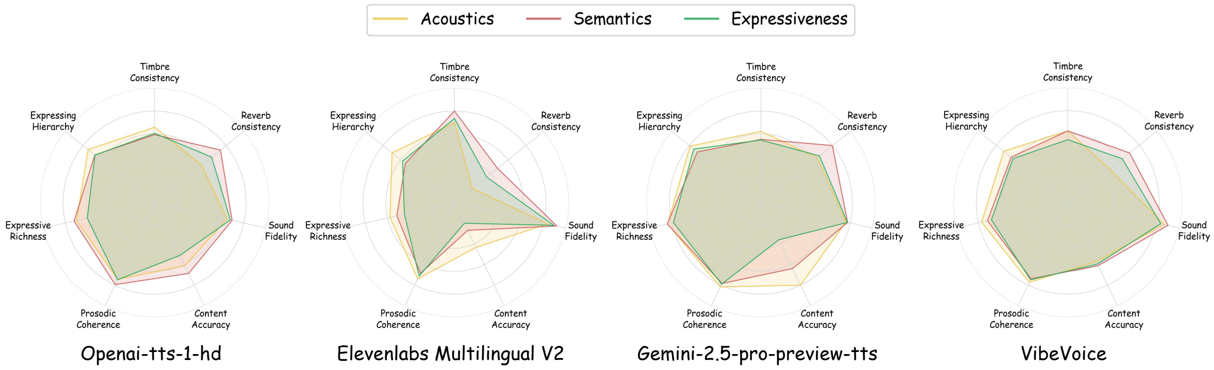



Figure 3: **LFS-Bench Results across Three Core Challenges.** For each chart, we plot the evaluation results across three core challenges. The results are normalized between 1 and 5 (larger is better) for visibility across challenges.

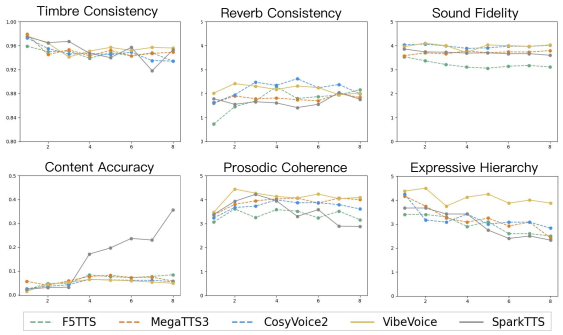



Figure 4: **Results on Sequence Length**. The horizontal axis represents the number of sentences in the text.

tent accuracy as sequence length increases, whereas NAR models maintain stability without significant degradation. Consequently, we propose that future long-form TTS architectures should evolve beyond this binary choice toward a Coarse-to-Fine Architecture (Kharitonov et al., 2023; Ju et al., 2024), thereby effectively reconciling long-range semantic coherence with local generation stability.

**Data Quality v.s. Data Quantity** While scaling laws have advanced speech synthesis by leveraging more data and bigger parameters (Du et al., 2025), our analysis suggests that relying solely on mainstream datasets presents three critical impediments to long-form audio generation: 1) **Fragmentation in open-source data** (Chen et al., 2021) induces a short-form bias that compromises discourse coherence. For instance, SparkTTS is trained on VoxBox, a dataset characterized by an average segment duration of less than 10 seconds. Consequently, the model exhibits significant degradation in both content accuracy and prosodic coherence as the generation length extends, as illustrated in Figure 4; 2) **Acoustic instability** in web-crawled data (He et al., 2024), such as variable noise and recording conditions, triggers acoustic drift. For example, CosyVoice3 utilizes extensive in-the-wild data for training. As a result, it significantly lags behind other models in reverb consistency, as shown in Table 2; and 3) The **averaging effect** of scaling enhances generalization but homogenizes expressiveness. As shown in Table 2, flagship models such as GLM-TTS and FishSpeech excel in acoustic metrics. However, they underperform in the expressiveness dimension despite their large scale. Consequently, they fail to capture the dynamic nuances required for narration. Therefore, the path forward requires a strategic shift towards prioritizing data quality and temporal continuity over raw quantity. We advocate for the adoption of curriculum-learning strategies (Wang et al., 2021) that progressively transition from sentence-level to paragraph-level training. By leveraging high-fidelity, long-context recordings, future models can more effectively capture the long-range dependencies essential for coherent and expressive narration.

## 6 Conclusion

In this work, we present LFS-Bench, a holistic benchmark tailored for evaluating long-form TTS models. LFS-Bench addresses three core challenges in long-form generation, encompassing 1,101 carefully curated instances across 17 downstream scenarios. To facilitate precise and automatic assessment, we propose a disentangled, human-aligned evaluation protocol featuring seven complementary metric dimensions. Through extensive benchmarking of over 20 models, we provide an in-depth analysis of current capabilities and limitations from the perspectives of model architectures as well as training data and strategy. We envision LFS-Bench as a standardized testbed for future research, propelling the development of more robust and immersive long-form speech synthesis.

## Limitations

We identify three limitations in this work. First, the linguistic scope of LFS-Bench is currently restricted to Chinese and English, leaving low-resource languages and diverse dialects or accents underexplored. Second, our investigation into semantics remains preliminary; while LFS-Bench's evaluation metrics prioritize acoustic coherence, we lack a robust automated framework to assess emotional and stylistic transitions grounded in deep semantic understanding of long-form text. Finally, the prompt speech utilized in our experiments is derived from only 20 speakers from open-source datasets. This limited speaker diversity may introduce evaluation bias, and we encourage the research community to contribute additional data to facilitate a more comprehensive assessment of model generalization.

## Ethical considerations

Although this work itself raises no immediate ethical concerns, two potential risks must be addressed when applying our benchmark. First, when utilizing our benchmark for evaluation, users must ensure that the prompt speech does not infringe upon the rights of the original voice actors. The use of audio from unverified sources or those restricted by regulations is strictly prohibited. Second, while our objective is to enhance the holistic performance of long-form synthesis, practitioners must ensure that models trained or evaluated using our methods are not deployed for generating disinformation, such as fabricated news reports or unauthorized political speeches. To mitigate these risks, we intend to implement strict usage guidelines upon open-sourcing the benchmark to prevent unethical and unauthorized applications.

## References


Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

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

Oleg Atamanenko, Anna Chalova, Joseph Coombes, Nikki Cope, Phillip Dang, Zhifeng Deng, Jimmy Du, Michael Ermolenko, Feifan Fan, Yufei Feng, et al. 2025. Tts-1 technical report. *arXiv preprint arXiv:2507.21138*.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 818–824.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu

Wei. 2024b. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

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

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. *arXiv preprint arXiv:1909.03965*.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, et al. 2025. Glm-tts technical report. *arXiv preprint arXiv:2512.14291*.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Google DeepMind. 2025. Gemini 3. https://deepmind.google/technologies/gemini/. Accessed: 2025-12-25.

Dake Guo, Xinfa Zhu, Liumeng Xue, Yongmao Zhang, Wenjie Tian, and Lei Xie. 2024a. Text-aware and context-aware expressive audiobook speech synthesis. *arXiv preprint arXiv:2406.05672*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hao-Han Guo, Kun Liu, Fei-Yu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024b. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890.

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

Hieu-Nghia Huynh-Nguyen, Ngoc Son Nguyen, Huynh Nguyen Dang, Thieu Vo, Truong-Son Hy, and Van Nguyen. 2025. Ozspeech: One-step zero-shot speech synthesis with learned-prior-conditioned flow matching. *arXiv preprint arXiv:2505.12800*.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22605–22623.

Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, Zhengtao Wang, Xu Tan, Xinyu Zhou, Tao Qin, and Xiangyang Li. 2025. Mooncast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

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

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36.

Zhipeng Li, Xiaofen Xing, Jingyuan Xing, Hangrui Hu, Heng Lu, and Xiangmin Xu. 2025. Long-context speech synthesis with context-aware memory. *arXiv preprint arXiv:2508.14713*.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*.

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

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

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

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The msp-conversation corpus. *Interspeech 2020*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Christoph Minixhofer, Ondřej Klejch, and Peter Bell. 2024. Ttsds-text-to-speech distribution score. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 766–773.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. 2025. Ttsds2: Resources and benchmark for evaluating human-quality text to speech systems. *arXiv preprint arXiv:2506.19441*.

Yuto Nishimura, Takumi Hirose, Masanari Ohi, Hideki Nakayama, and Nakamasa Inoue. 2024. Hall-e: hierarchical neural codec language model for minute-long zero-shot text-to-speech synthesis. *arXiv preprint arXiv:2410.04380*.

OpenAI. 2024. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2025-12-25.

OpenAI. 2025. Gpt-5. https://chagpt.com/. Accessed: 2025-12-25.

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

Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan. 2024. Long-form speech generation with spoken language models. *arXiv preprint arXiv:2412.18603*.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, et al. 2025. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2024. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *ICLR*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. 2025. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation. *arXiv preprint arXiv:2510.06961*.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE international conference on acoustics, speech and signal processing*.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.

Haoxu Wang and Biao Tian. 2025. Zipenhancer: Dual-path down-up sampling-based zipformer for monaural speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, pages 4555–4576.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, et al. 2025a. Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity. *arXiv preprint arXiv:2510.23541*.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. 2025b. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

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

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, et al. 2025a. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*.

Weicheng Zhang, Cheng-Chieh Yeh, Will Beckman, Tuomo Raitio, Ramya Rasipuram, Ladan Golipour, and David Winarsky. 2023. Audiobook synthesis with long-form neural text-to-speech. In *12th Speech Synthesis Workshop (SSW) 2023*.

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

Xingjian Zhao, Zhe Xu, Qinyuan Cheng, Zhaoye Fei, Luozhijie Jin, Yang Wang, Hanfu Chen, Yaozhou Jiang, Qinghui Gao, Ke Chen, et al. 2025. Moss-speech: Towards true speech-to-speech models without text guidance. *arXiv preprint arXiv:2510.00499*.

Siqi Zheng, Luyao Cheng, Yafeng Chen, Hui Wang, and Qian Chen. 2023. 3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. *arXiv preprint arXiv:2306.15354*.

Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, Xi Yang, et al. 2025a. Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 12524–12537.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025b. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Dong Zhang, Xin Zhang, et al. 2025a.

Zipvoice-dialog: Non-autoregressive spoken dialogue generation with flow matching. *arXiv preprint arXiv:2507.09318*.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025b. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*.

## Appendix Contents

The Appendix is structured as follows:

- Section A: Details of dataset construction, including the detailed explanation of scenarios as well as the complete process of data collection and refinement.
- Section B: Statistics of LFS-Bench.
- Section C: Details of Evaluation Protocols.
- Section D: Details of the validation of human alignment and the user study.
- Section E: The details of the experiment's setting.
- Section F: Ablation studies and experiments related to multi-speaker dialogue evaluation.
- Section G: More reuslts and analysis of the experiments.
- Section H: Limitations and future works.
- Section I: Potential social impact of LFS-Bench.

## A Details of LFS-Bench's Construction

### A.1 Explanation of Scenarios

LFS-Bench systematically categorizes the challenges inherent in current long-form speech generation into three primary dimensions: **Acoustics, Semantics, and Expressiveness**. To facilitate a more fine-grained and precise assessment, we curate a dataset of 1,101 audio samples aligned with these dimensions, encompassing 17 downstream scenarios such as audiobooks, podcasts, talk shows, and news broadcasting. In the following section, we comprehensively detail the audio scenarios and data selection criteria associated with each challenge category.

#### Scenarios for Acoustics Challenges

In the context of long-form TTS and dialogue generation tasks, the primary user concerns regarding acoustic performance are categorized as follows:

- **Audio Quality:** As a fundamental requirement, the generated audio must be devoid of background noise and electronic artifacts, ensuring high fidelity and clear auditory perception for the user.
- **Timbre Consistency:** In single-speaker settings, the speaker's timbre must remain perceptually consistent throughout the sequence, analogous to identity preservation in video generation tasks. In multi-speaker dialog scenarios, accurate speaker transitions are critical, requiring precise alignment between the dialogue script and the corresponding speaker identities.
- **Acoustic Environment Consistency:** The ability to maintain a stable sound field is a core capability in long-form speech generation. This requires unity across acoustic dimensions, such as the recording environment and sound imaging. Furthermore, in multi-speaker contexts, ensuring that different speakers appear to share a unified acoustic scene is a crucial objective.

Based on the above basic requirements, we select six audio downstream scenarios to construct test cases related to the acoustic dimension, which are specifically introduced as follows.

**Customer Service** Widely deployed in e-commerce, AI agents frequently deliver lengthy responses detailing policies and products. This scenario demands high-fidelity, artifact-free audio to maintain professional credibility and ensure a trustworthy user experience.

**Audiobooks** As a quintessential long-form scenario, audiobooks demand rigorous acoustic consistency. The synthesis must maintain timbre stability to mitigate "speaker drift," preserve a stationary acoustic environment to ensure immersion, and guarantee high-fidelity quality for prolonged listening comfort.

**Podcasts** This scenario focuses on multi-turn dialogue generation and natural interaction. Characterized by an informal or semi-formal conversational style, this domain places relatively lower demands on dramatic expressiveness; however, it imposes strict requirements on turn-taking transitions. Consequently, this scenario necessitates that TTS models not only execute accurate speaker switching but also synthesize appropriate and stable reverberation to reconstruct an authentic and vivid conversational atmosphere.

**Chat, Debate, and Interview** While lacking direct commercial applications, these real-world scenarios serve as benchmarks for acoustic modeling limits. The frequent speaker transitions inherent in these domains pose significant challenges to synthesis systems. Furthermore, the associated complex acoustic environments introduce additional layers of difficulty regarding background noise and channel variability.

### Scenarios for Semantics Challenges

In the semantic dimension, long-form speech generation is categorized into two sub-dimensions: accuracy and naturalness.

- **Content Accuracy**: Evaluates the alignment between the generated speech and the input text. In long-sequence generation, this metric primarily assesses the model's robustness against omissions, repetitions, and hallucinations, ensuring high content fidelity.
- **Prosodic Coherence**: Evaluates the consistency between prosodic structure and semantic logic. Beyond natural pausing, this includes the appropriate handling of stress and intonation, ensuring a fluent rhythm at the paragraph level and avoiding mechanical or disjointed delivery.

To rigorously evaluate model performance regarding semantic challenges, we construct test cases across five downstream scenarios, specifically targeting the two aforementioned dimensions.

**News and Popular Science** In these scenarios, content correctness is paramount, as users exhibit minimal tolerance for semantic deviations. Consequently, we curate instances featuring linguistic complexity, challenging pronunciations, and domain-specific knowledge to comprehensively assess model robustness.

**Lesson, Seminar, and Presentation** Beyond basic accuracy, these scenarios impose higher demands on naturalness. Speakers are expected to enhance auditory perception through appropriate stress and rhythmic cadence. Therefore, in addition to content complexity, we incorporated colloquial expressions and diverse prosodic structures to further evaluate the model's prosodic coherence.

### Scenarios for Expressiveness Challenges

Immersion and high expressiveness are the ultimate goals of audio synthesis. For long-form generation, given its temporal complexity, we decompose expressiveness into Richness and Hierarchy.

- **Expressive Richness**: Evaluates the overall expressive quality through the lenses of emotional resonance, character portrayal, and storytelling. Similar to sentence-level synthesis, this metric primarily focuses on the **average magnitude** of expressiveness maintained throughout the entire audio sequence.
- **Expressive Hierarchy**: Represents the fundamental distinction between paragraph-level and sentence-level generation. The extended context necessitates a focus on dynamic variations (e.g., shifts in emotion and volume) and the alignment between the acoustic evolution and the semantic scenario.

Guided by these evaluation dimensions, we curate test cases across six highly expressive downstream scenarios to rigorously probe the upper boundaries of model capabilities within LFS-Bench.

**Sportcast and Live Streaming**: These scenarios predominantly challenge Expressive Richness. Characterized by sustained high-intensity delivery and emotional saturation, they demand that the model maintain a consistently elevated energy level to match the fast-paced nature of the content.

**Speech, Host, Talkshow, and Drama**: These domains necessitate a synergy of both Richness and Hierarchy. Beyond high emotional fidelity, they require sophisticated control over dynamic evolution, such as tension building in drama or rhythmic variation in hosting, to ensure the acoustic delivery aligns seamlessly with the narrative arc.

## A.2 Details of Data Collection

In this section, we provide further elaboration on the data sources and processing pipeline of LFS-Bench.

### Online Text Corpora

For the Audiobook, News, Drama, and Host scenarios, we harvest long-form texts from diverse online resources, spanning classic literature, web novels, and TouTiao[4]. Following data acquisition via OCR or web crawling, we employ the `clean-text`[5] library to sanitize the raw corpus by removing artifacts such as URLs, emojis, and garbled characters. Subsequently, human annotators conduct rigorous quality assurance and enrich the dataset with metadata labels for scenario, topic, and speaker identity.

[4] https://app.toutiao.com/news_article/
[5] https://pypi.org/project/clean-text/

**Online Audio Media**

We extensively utilize online audio materials across various scenarios, with data sources including YouTube[6], Bilibili[7], Spotify[8], RedNote[9], and Apple Podcasts[10]. First, we crawl audio materials tailored to our target scenarios from these platforms. Subsequently, we denoise the raw audio using Zipenhancer (Wang and Tian, 2025) to ensure processing accuracy. After obtaining cleaner data, we filter out samples with low expressiveness and quality based on a DNS-MOS (Reddy et al., 2021) threshold of 3.5. We then perform speaker diarization using 3D-Speaker (Zheng et al., 2023) and transcribed the resulting audio segments via SenseVoice-Small[11]. Finally, human annotators are employed to proofread the machine-generated transcripts against the ground truth and update the metadata labels.

**LLM Generation**

In scenarios such as chat, presentations, and customer service, we leverage GPT-5 (OpenAI, 2025) to facilitate the generation of high-quality test cases. Specifically, we designe sophisticated prompts to guide the LLM in producing structured content that aligns with specific scenarios and topics while maintaining a certain level of generation complexity. Figure 5 illustrates a set of prompts used for generating presentation topics for computer science students. These structured prompts serve as customizable templates, allowing users to adapt them for generating diverse long-form data. After LLM generation, the generated content is mutually proofread by annotators.

We recruit three undergraduate students for data annotation and verification, compensated at a rate of $0.20 per instance. To ensure quality, all data samples are double-checked. The total expenditure for the data collection process amount to $220.

### A.3 Details of Data Refinement

**Semantic De-duplication**

To ensure data diversity, we perform topic-level deduplication on both crawled and generated test instances. Specifically, we utilized GPT-5 to extract topics, keywords, and summaries from each long-text instance. These elements are concatenated and encoded into embeddings using Sentence-BERT[12] (Reimers and Gurevych, 2019). We then filter out semantically redundant samples based on a cosine similarity threshold of 0.8 and replenish the dataset via LLM-based generation.

**Quality Evaluation**

We further employ GPT-5 to assess the quality of the de-duplicated samples. Specifically, we design prompts to evaluate textual expressiveness and content consistency, guiding the LLM to rate the suitability of each instance for long-form speech generation on a scale of 1 to 5. Only samples with recommendation scores exceeding 2 are retained. The specific prompt used for this quality assessment is in Figure 6.

**Privacy and Ethical Filtering**

To ensure the safety and anonymity of our dataset, we employ DeepSeek V3.2 (Liu et al., 2024a) to conduct a rigorous privacy and ethical assessment. Specifically, we design a prompt incorporating Chain-of-Thought (CoT) (Wei et al., 2022) reasoning to guide the model through a two-step analysis:

1. **Selective PII Anonymization**: The model is instructed to specifically identify and anonymize the names of **private individuals** (non-public figures). While the names of celebrities or public entities are retained to preserve contextual integrity, the names of ordinary citizens are replaced with generic placeholders or synthetic alternatives.

2. **Ethical Risk Assessment**: The model then scrutinizes the content for social and ethical risks, including hate speech, violence, sexual explicitness, and bias.

Based on this analysis, samples containing toxic content are discarded, while those with minor sensitivity issues are revised. The specific prompt used for this filtering is presented in Figure 7.

**Manual Review**

Following the automated filtering pipelines, we implement a three-stage human-in-the-loop review process to finalize the dataset. Expert annotators execute the following operations:

[6] https://www.youtube.com
[7] https://www.bilibili.com
[8] https://open.spotify.com/
[9] https://www.xiaohongshu.com/
[10] https://podcasts.apple.com/
[11] https://huggingface.co/FunAudioLLM/SenseVoiceSmall

[12] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

**Prompt for generating structured presentation data**

You are an expert computer science professor and content creator. Your task is to generate a high-quality, long-form presentation script on the topic: **[Insert Topic Here]**.
**Generation Requirements:** 1. **Complexity**: The content must be academically rigorous, suitable for computer science students. Include technical terminology and logical reasoning. 2. **Structure**: The speech should be coherent but segmented into logical paragraphs. 3. **Format**: You must strictly output a valid JSON object without any Markdown formatting.
**JSON Schema:**

```
{
  "content": [
    {
      "speaker": "Speaker1",
      "text": "The first paragraph of the speech..."
    },
    {
      "speaker": "Speaker1",
      "text": "The second paragraph of the speech..."
    }
  ],
  "num_speakers": 1,
  "theme": "[Insert Topic Here]",
  "source": "LLM Generation",
  "TLDR": "A one-sentence summary of the presentation."
}
```

Figure 5: Prompt template used for generating presentation topics for computer science students.

1. **Harmless Placeholder Infilling**: For samples that underwent privacy anonymization, the automated generic tags (e.g., `[NAME]`, `[LOC]`) are replaced with specific but fictitious entities. This step ensures the text remains natural and grammatically fluid while strictly maintaining the harmlessness and anonymity.

2. **Residual Error Purging**: Annotators then scrutinize the dataset to identify subtle logical inconsistencies, formatting errors, or context mismatches that might have evaded the automated filters. Samples deemed substandard or unnatural are strictly discarded.

3. **Dataset Replenishment**: To compensate for the discarded samples and maintain the volume, new instances are constructed. These replenished samples undergo the same process before being added to the final pool.

Five undergraduate students are enlisted for this manual review, receiving a compensation of \$0.30 per instance. The cumulative expenditure for the data collection process totaled \$330.

### A.4 Instructions for Use

The test set will be released on Hugging Face under the **CC BY-NC-SA 4.0** license, allowing for free non-commercial use. For evaluations involving additional voice profiles on our benchmark, users must strictly adhere to the specific licenses associated with those assets. Furthermore, the complete codebase for data processing and evaluation will be made publicly available on our GitHub repository.

## B Statistics of LFS-Bench

### B.1 Categorical Statistics

We present a comprehensive statistical analysis of the 1,101 samples in LFS-Bench across five key dimensions: language (Chinese/English), speaker configuration (single/dual/multi-speaker), core challenges (Acoustics, Semantics, Expressiveness), scenarios, and content topics, as illustrated in Figure 8. As observed, LFS-Bench maintains a strictly balanced language ratio, comprising 49.3% Chinese and 50.7% English samples. Regarding speaker configuration, while the dataset primarily focuses on single-speaker long-form speech and dual-speaker dialogue, we explicitly include 101 multi-speaker samples (involving 3 or 4 speakers) to facilitate the evaluation of multi-talker generation capabilities. Furthermore, the dataset exhibits a relatively even distribution across the three core challenges, with the Acoustics challenge accounting for the largest proportion at 34.5%. We also quantify the sample distribution across 17 specific downstream scenarios and generate a word cloud to visualize the topic diversity. This balanced sce-

**Prompt for the evaluation of long-form instances**

You are an expert linguist and data quality evaluator. Your task is to assess the suitability of the following text sample for **long-form speech generation**.
Please evaluate the text based on the following two criteria:
1. **Textual Expressiveness**: Assess the fluency, naturalness, and rhetorical quality of the text. Is the language vivid and rhythmically suitable for long-duration speech synthesis?
2. **Content Consistency**: Assess the logical coherence and semantic stability of the text. Is the narrative or argument consistent throughout without contradictions or abrupt topic shifts?
Rate each criterion on a scale of 1 to 5 (1 = Poor, 5 = Excellent). Based on these, provide an **Overall Score** (1-5) representing your recommendation for retaining this sample.
**Output Requirement:**
You must output the result strictly in the following JSON format:
{
"reasoning": "Provide a brief analysis explaining the scores, highlighting pros and cons.",
"textual_expressiveness_score": <integer between 1 and 5>,
"content_consistency_score": <integer between 1 and 5>,
"overall_score": <integer between 1 and 5>
}
**Text to Evaluate:**
[Insert Text Here]

Figure 6: Prompt template for the quality evaluation of test instances.

nario distribution, combined with a rich variety of content topics, minimizes potential bias during the evaluation process.

### B.2 Distributional Statistics

We also conduct a detailed analysis of the text length distribution within LFS-Bench, as illustrated in Figure 9. Specifically, text length is quantified by the number of characters for Chinese data and the number of words for English data, excluding non-phonetic elements such as punctuation. The results indicate that text lengths for both languages follow an approximate normal distribution, primarily concentrate within the interval $[80, 500]$, with mean lengths of 271.8 for Chinese and 174.6 for English. This distribution effectively supports the rigorous and realistic evaluation of long-form speech generation capabilities.

## C Details of Evaluation Protocol

### C.1 Timbre Consistency

To evaluate timbre consistency, we adopt a segment-based speaker similarity approach following prior zero-shot TTS studies (Du et al., 2024; Guo et al., 2024b).

Specifically, for a **single-speaker** long-form speech sample $w$, we apply a sliding window over the waveform to extract a sequence of speaker embeddings $\{\mathbf{e}_i\}_{i=1}^{n}$ by WavLM TDCNN [13], where $n$ denotes the number of windows. Given that speaker embeddings are sensitive to segment duration and verification models are typically optimized for 2–4s segments, **we employ a window length of 3s with a stride of 2s.** We then compute the pairwise cosine similarity between all distinct embeddings:

$$\mathrm{sim}_{i,j} = \cos\left(\frac{\mathbf{e}_i}{\|\mathbf{e}_i\|}, \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|}\right), \quad \forall i \neq j. \quad (2)$$

Finally, we utilize the average score of the resulting similarity sequence $\{\mathrm{sim}_{i,j}\}$ as the quantitative metric for timbre consistency.

Evaluating dual and multi-speaker scenarios is inherently more complex due to the involvement of speaker transitions. To ensure validity, we first utilize 3D-Speaker (Zheng et al., 2023) to verify the number of speakers, confirming that at least one successful speaker turn occurs. Subsequently, let $K$ denote the number of distinct speakers in the generated audio. We employ forced alignment to obtain sentence-level timestamps and concatenate speech segments belonging to each speaker $k \in \{1, \ldots, K\}$, yielding a speaker-specific audio stream $\tilde{w}_k$. We utilize a Paraformer-based Align Model[14] (Gao et al., 2022) for Chinese data and WhisperX[15] (Bain et al., 2023) for English data. Both models demonstrate alignment errors of less than 100ms on minute-level recordings, minimizing error accumulation. Finally, for each speaker-

[13] https://huggingface.co/docs/transformers/en/model_doc/unispeech-sat

[14] https://modelscope.cn/models/iic/speech_timestamp_prediction-v1-16k-offline

[15] https://github.com/m-bain/whisperX

**Prompt for Privacy and Ethical Filtering**

**Role:** You are an expert data safety and privacy compliance assistant. Your task is to review the input text for privacy leaks and ethical risks.
**Instructions:** Please analyze the input text following these steps (Chain-of-Thought):

1. **PII Detection (Selective):** Identify all person names.
   - If the name belongs to a **public figure** (celebrity, politician, historical figure), retain it to preserve context.
   - If the name belongs to a **private individual** (ordinary citizen), anonymize it using a placeholder (e.g., [NAME]).
2. **Ethical Risk Assessment:** Check for hate speech, explicit violence, sexual content, or severe bias.
   - If the risk is severe and cannot be mitigated, mark as invalid.
   - If the risk is minor or related to PII, provide a revised version.

**Output Format:** Output the result in a strict JSON format with the following keys:

- `"reasoning"`: A brief explanation of your analysis regarding PII and safety risks.
- `"valid"`: Boolean (true/false). Set to `false` only if the content contains unmitigable toxic content. Set to `true` if it is safe or has been successfully anonymized/revised.
- `"revised_text"`: The clean version of the text after anonymizing private names and removing minor risks. If invalid, return an empty string.

**Input Text:** [INPUT_TEXT]

Figure 7: The prompt template used for privacy and ethical filtering. It guides the LLM to selectively anonymize private individuals' names while retaining public figures, and outputs the decision in a structured JSON format.

specific stream $\tilde{w}_k$, we compute its similarity average $a_k$ following the single-speaker protocol defined above. The final metric is calculated as the average across all speakers:

$$\text{Score}_{\text{multi}} = \frac{1}{K}\sum_{k=1}^{K} a_k. \tag{3}$$

### C.2 Reverb Consistency

We employ the Speech-to-Reverberation Modulation Energy Ratio (SRMR) to quantify reverberation intensity, analyzing its temporal fluctuations to evaluate the model's ability to maintain a consistent acoustic environment.

Specifically, for a generated utterance $w$, we apply a sliding window to compute the SRMR for each segment using the SRMRpy toolkit[16]. To balance estimation reliability with the temporal resolution required to detect "reverberation drift", we adopt a window size of 3s and a stride of 2s, consistent with our timbre consistency evaluation.

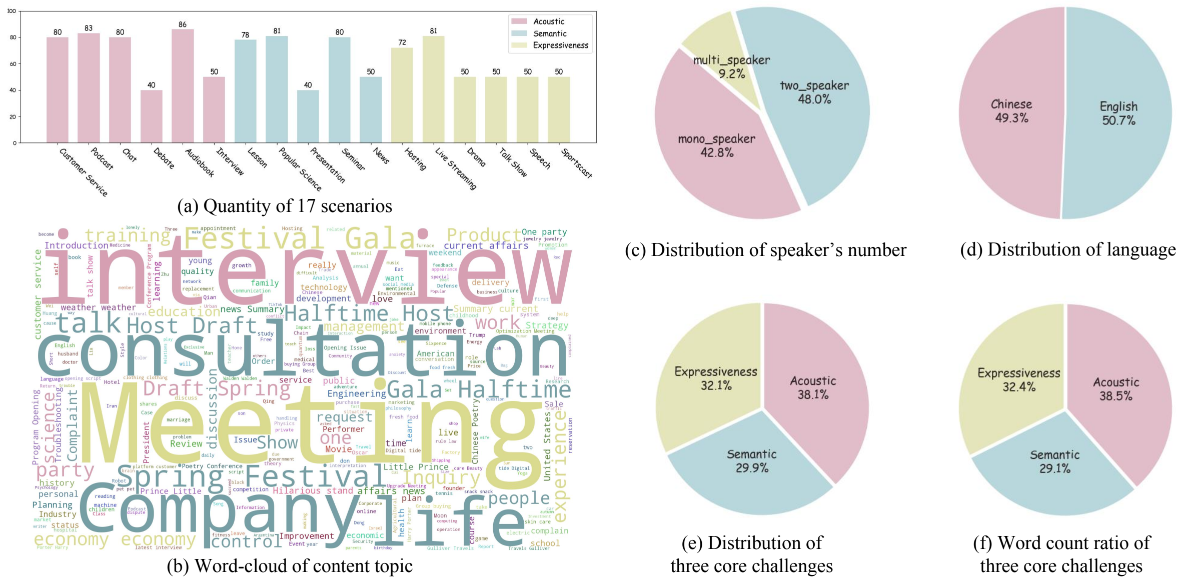
Furthermore, to mitigate the impact of non-speech segments (e.g., silence or noise) on the statistical analysis, we pre-filter each window using a Voice Activity Detection (VAD) model[17]. Any window containing more than 60% non-speech frames is discarded. This process yields a sequence of valid reverberation scores $\{r_i\}_{i=1}^{n}$, where $n$ denotes the number of effective windows.

Finally, we compute the standard deviation of this sequence as our Reverb Consistency metric; a lower value indicates a more stable reverberation pattern throughout the utterance.

It is important to note that this metric is predicated on the assumption that the **acoustic environment within a single long-form segment should remain stable**. We acknowledge that specific scenarios, such as *Outdoor Live Streaming*, may inherently require dynamic acoustic shifts for semantic correctness. However, for the majority of standard long-form synthesis tasks, acoustic stability serves as a critical indicator of generation robustness; therefore, we treat high variance as a penalty in this evaluation framework.

[16] https://github.com/jfsantos/SRMRpy

[17] https://modelscope.cn/models/iic/speech_fsmn_vad_zh-cn-16k-common-pytorch, https://github.com/snakers4/silero-vad



(a) Quantity of 17 scenarios

(b) Word-cloud of content topic

(c) Distribution of speaker's number

(d) Distribution of language

(e) Distribution of three core challenges

(f) Word count ratio of three core challenges

Figure 8: The categorical statistics of LFS-Bench across five key dimensions: language, speaker numbers, core challenges, content topics and scenarios.

### C.3 Sound Fidelity

To achieve a non-intrusive, reference-free assessment of audio fidelity, we directly utilize the SQUIM-PESQ metric via the official `Torchaudio` interface[18]. This metric yields scores ranging from -0.5 to 4.5, with values typically exceeding 1.0 for speech audio.

### C.4 Content Accuracy

To quantify content accuracy, we employ Character Error Rate (CER) for Chinese and Word Error Rate (WER) for English. The evaluation pipeline proceeds as follows: First, we obtain the transcribed text $T_{\text{pred}}$ from the generated audio using FunASR-Nano[19]. Subsequently, we perform rigorous normalization on both the ground truth $T_{\text{gt}}$ and the prediction $T_{\text{pred}}$. This process includes: 1) **Punctuation Removal**: eliminating punctuation via `string.punctuation` and `zhon.hanzi.punctuation`[20]; 2) **Whitespace Standardization**: trimming leading/trailing spaces and collapsing multiple spaces; and 3) **Character Normalization**: converting Traditional Chinese to Simplified using `zhconv`[21] while filtering non-ASCII characters in English text via `clean-text`[22]. Finally, following the methodology of F5-TTS (Chen et al., 2024c), we calculate the WER and CER using the `JiWER` library[23].

It is worth noting that our selected transcription system, FunASR-Nano, demonstrates exceptional performance on clean speech benchmarks, achieving a WER of 1.76% on Librispeech-clean (EN) and a CER of 2.56% on Fleurs-zh. These results are competitive with state-of-the-art models of similar parameter scale (Srivastav et al., 2025). Utilizing such a high-performance ASR model minimizes transcription-induced errors, ensuring that the reported metrics accurately reflect the content fidelity of the generated audio.

### C.5 Prosodic Coherence

For prosody evaluation, we utilize Speech-Judge (Zhang et al., 2025b), a fine-tuned Qwen2.5-Omni model specialized for audio assessment. To specifically target long-form modeling capabilities, we refine the original prompt design, decomposing the evaluation criteria into three granular dimensions: *Prosodic Coherence & Flow*, *Rhythmic Hierarchy & Layering*, and *Overall Naturalness*. Ratings are assigned on a scale from 1.0 to 5.0, as detailed in Figure 12. Furthermore, to mitigate the inherent variance of LALMs, we conduct 10 independent evaluations for each generated audio sample and calculate the mean to derive the final prosody score.

[18] https://docs.pytorch.org/audio/main/tutorials/squim_tutorial.html
[19] https://github.com/FunAudioLLM/Fun-ASR
[20] https://pypi.org/project/zhon/
[21] https://pypi.org/project/zhconv/
[22] https://pypi.org/project/clean-text/
[23] https://pypi.org/project/jiwer/

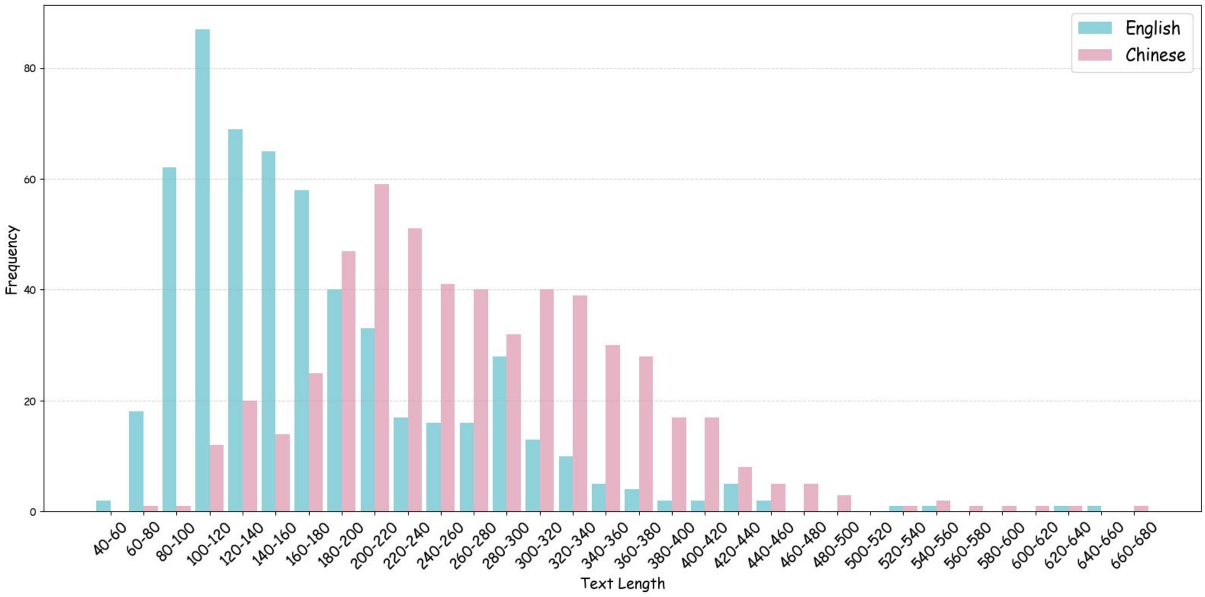



Figure 9: The statistics of the text length distribution within LFS-Bench. The red dashed line indicates the average text length of English, and the green dashed line indicates the average text length of Chinese.

### C.6 Expressive Richness

This dimension assesses the global expressive quality of the generated speech, representing the average level of expressiveness. Formally, we segment the audio waveform into a sequence of non-overlapping 10-second chunks $\{c_i\}_{i=1}^{M}$. An LALM is then employed to assign an expressiveness score $s_i$ to each chunk $c_i$. The final Expressive Richness metric is defined as the arithmetic mean of these segment scores:

$$\text{Score}_{\text{rich}} = \frac{1}{M}\sum_{i=1}^{M} s_i. \quad (4)$$

The 10-second segmentation window is selected to align with the typical generation duration of chunk-based long-form synthesis pipelines. This strategy effectively mitigates the confounding effects of inter-chunk inconsistencies, allowing for a more focused evaluation of intrinsic expressiveness. The prompt template used for this assessment is illustrated in Figure 14.

### C.7 Expressive Hierarchy

Complementing the local expressiveness defined above, paragraph-level expressive hierarchy is equally critical in long-form settings. Unlike the segment-based approach for **Expressive Richness**, we leverage the long-context understanding capabilities of modern LALMs to conduct a holistic assessment. Specifically, the entire audio sequence is fed into the model, which is instructed to evaluate the speech based on three dimensions: **Emotional Variation**, **Vocal Dynamics**, and **Scene Appropriateness**.

The prompt template used for this assessment is illustrated in Figure 13.

## D User Study

For the subjective evaluation, we recruit a balanced cohort of 10 expert listeners (5 male, 5 female) with diverse professional backgrounds, including audio engineers from the internet industry, live streaming specialists, and academic researchers (professors and PhD candidates) in signal processing. All participants possess extensive experience in audio quality assessment. In all subjective tests, we conduct Mean Opinion Score (MOS) evaluation. They are compensated at a rate of $1.00 per evaluation instance (either a single sample or a paired comparison), with the total expenditure for the user study amounting to $2,000.

### D.1 Validation of Timbre Consistency

In this experiment, we randomly select 50 samples from the test set for subjective evaluation. Listeners are instructed to rate the "Timbre Maintenance" capability using a Mean Opinion Score (MOS). They are explicitly required to focus exclusively on timbre stability, disregarding other acoustic factors (e.g., sound field, audio quality) and semantic dimensions (e.g., pronunciation, prosody). If the expressiveness of the audio does not affect the timbre, it can also be ignored.

We concurrently compute the objective Timbre Consistency score for each sample. The correlation analysis between the subjective MOS and our objective metric yields the following results: SRCC=0.75, PLCC=0.77, and KRCC=0.59. These results demonstrate that our proposed timbre consistency evaluation aligns closely with human perception.

Furthermore, the user study reveals several statistical thresholds regarding our objective metric:

1. **Score $< 0.85$**: Indicates significant timbre drift. In multi-speaker scenarios, this may also suggest inaccurate speaker transitions.

2. **Score $< 0.93$**: Demonstrates superior timbre maintenance, with performance comparable to ground truth recordings.

3. **Score $\in [0.85, 0.90]$**: Represents generally acceptable performance, typically characterized by minor local timbre mutations or artifacts.

Besides, the robustness of this metric presents room for improvement. Potential misclassifications may arise in specific edge cases, such as audio exhibiting periodic timbre variations (e.g., looping patterns). Since our metric relies on global averages, it may fail to penalize such rhythmic fluctuations, yielding a favorable score despite perceptual inconsistency. Future work will aim to incorporate temporal modeling to address these dynamic artifacts.

### D.2 Validation of Sound Fidelity

Considering that `SQUIM-PESQ` is trained on English sentence-level data, we select 50 samples from the test set to verify its generalization to Chinese and long-form scenarios. Listeners are instructed to rate "Clarity and Fidelity" using MOS. Specifically, they are required to focus exclusively on factors such as background noise, artifacts, and articulation, while disregarding prosody and expressiveness. We concurrently compute the `SQUIM-PESQ` scores for these samples. The correlation analysis between subjective MOS and `SQUIM-PESQ` yield an SRCC of 0.72, a PLCC of 0.47, and a KRCC of 0.53. These results demonstrate that the metric aligns closely with human perception.

Table 4: Human alignment comparison across different LALMs on Expressive Richness.

| Models | PLCC | SRCC | QWK | MAE |
|---|---|---|---|---|
| **UTMOS** | -0.0203 | -0.0433 | -0.0313 | 1.043 |
| **UTMOSv2** | -0.0745 | -0.0789 | -0.0827 | 0.9012 |
| **SQUIM-MOS** | -0.3145 | -0.2767 | -0.0825 | 1.3177 |
| **DNS-MOS** | -0.0243 | -0.0189 | -0.0034 | 0.8537 |
| **GPT-4o** | 0.1549 | 0.2002 | 0.1435 | 0.7982 |
| **Qwen3Omni-Flash** | 0.1464 | 0.1696 | 0.0812 | 1.0401 |
| **Qwen3Omni-Instruct** | 0.2245 | 0.2488 | 0.1172 | 1.0809 |
| **Gemini2.5-flash** | 0.4166 | 0.4079 | 0.2623 | 0.8123 |
| **Gemini2.5-pro** | 0.5085 | 0.5160 | 0.4242 | 0.7635 |
| **Gemini3-flash** | 0.5224 | 0.5266 | 0.5066 | 0.6562 |
| **Gemini3-Pro** | **0.7061** | **0.7080** | **0.6772** | **0.5879** |

### D.3 Validation of Prosodic Coherence

To validate the Prosodic Coherence metric, we adopt the methodology of SpeechJudge (Zhang et al., 2025b), conducting a human preference test to assess the model's evaluation performance. In addition to the robust correlation reported in Section 3.5, our analysis yields the following statistical insights:

1. **Score Divergence** $> 1$: A difference of more than 1 points indicates a substantial and perceptually obvious gap in prosodic quality between audio samples.

2. **Score** $\geq 4$: Audio samples achieving this threshold demonstrate competent basic prosody and rhythmic structure.

3. **Score** $\geq 4.5$: Performance at this level is considered virtually indistinguishable from ground truth recordings.

### D.4 Validation of Expressiveness

In this experiment, we curate a diverse set of 200 samples spanning all models and tasks for subjective evaluation. Listeners are tasked with rating the audio strictly adhering to the same prompt criteria provided to the LALMs.

Concurrently, we benchmark this 200-sample test set against 4 specialized MOS prediction models (UTMOS (Saeki et al., 2022), UTMOSv2 (Baba et al., 2024), SQUIM-MOS (Kumar et al., 2023), DNS-MOS (Reddy et al., 2021)) and 8 flagship LALMs (GPT-4o, Qwen3Omni-Instruct-30B-A3B (Xu et al., 2025b), Qwen3Omni-Flash, StepFun-Audio-R1 (Tian et al., 2025), Gemini-2.5-flash, Gemini-2.5-pro, Gemini-3-flash, Gemini-3-pro). Notably, due to context length constraints, only a subset of these LALMs is employed for the Expressive Hierarchy evaluation.

Table 5: Human alignment comparison across different LALMs on Expressive Hierarchy.

| Models | PLCC | SRCC | QWK | MAE |
|---|---|---|---|---|
| **GPT-4o** | 0.1328 | 0.1171 | 0.0803 | 0.7604 |
| **Qwen3Omni-Flash** | 0.3263 | 0.2496 | 0.2193 | 0.8426 |
| **Qwen3Omni-Instruct** | 0.1641 | 0.1181 | 0.0869 | 0.9130 |
| **Gemini2.5-flash** | 0.0421 | 0.0005 | 0.0256 | 0.8673 |
| **Gemini2.5-pro** | 0.3732 | 0.3744 | 0.2871 | 0.800 |
| **Gemini3-flash** | 0.406 | 0.3924 | 0.2032 | 1.1837 |
| **Gemini3-Pro** | **0.6041** | **0.6234** | **0.5452** | **0.7204** |

We examine the correlation between the mean listener ratings and the model-predicted scores, with results summarized in Table 4 and Table 5. Notably, **Gemini3-Pro** demonstrates superior performance, significantly outperforming other models across both metrics. It is also worth noting that all traditional MOS prediction networks exhibited poor correlation with human perception regarding expressiveness. This suggests that standard MOS training datasets likely lack a specific focus on expressive qualities.

Moreover, we conduct independent repeated trials on this test set to validate the stability of our selected evaluator, Gemini 3 Pro. Specifically, we perform five independent scoring iterations for each audio sample, where Gemini 3 Pro yields inconsistent scores for only 11 instances, demonstrating a level of robustness comparable to human evaluators. Consequently, we adopt **a single-pass evaluation strategy** for this metric.

Furthermore, to ensure consistency in the rating scales adopted by our recruited listeners, we computed the correlation between each individual rater and the mean score of the remaining raters. As shown in Table 6, the high inter-rater correlation confirms the reliability and validity of our evaluation protocol.

# E Implementation Detail

## E.1 Computational Resources and Environments

All inference and evaluation experiments for open-source models are conducted on a server equipped with 8 NVIDIA GeForce RTX 4090 GPUs and an Intel Xeon Gold 6530 CPU, running Ubuntu 22.04. For model inference, we strictly adhere to the environment specifications provided in the respective official repositories. The core dependencies for our evaluation pipeline include Python 3.10, PyTorch 2.8.0, Torchaudio 2.8.0, and Transformers 4.57.3.

## E.2 Selected Voice

For open-source models, we curate a set of 25 reference audio prompts from diverse datasets, including Emilia (He et al., 2024), AISHELL-3 (Shi et al., 2020), NCSSD (Liu et al., 2024b), LibriSpeech (Panayotov et al., 2015), MSPPodcast (Martinez-Lucas et al., 2020), and ChildMandarin (Zhou et al., 2025a), as well as reference voices provided in specific model repositories (see Table 7). We conduct extensive evaluations across these prompts and reported the results of the best-performing voice for each model. This strategy aims to minimize the impact of biases arising from training data discrepancies and inherent voice preferences.

For closed-source models, we selected official voices characterized by high fidelity, superior prosody, and rich expressiveness. Detailed specifications are provided in Table 8.

## E.3 Synthesis Strategy

For open-source models, we strictly adhere to the default configurations provided in their official repositories. Specific adjustments for MegaTTS3, CosyVoice3, and IndexTTS2 are detailed below:

**MegaTTS3** As the official VAE Encoder (Kingma et al., 2013) is not publicly available, we obtain the VAE latents for our reference prompt speech by contacting the model maintainers.

**IndexTTS2** To ensure a fair and objective comparison, we disabled the text sentiment analysis module by setting `use_emo_text` to `false`.

**CosyVoice3** We utilized the system text prompt "You are a helpful assistant" during generation, consistent with the official implementation.

For closed-source models, we similarly followed the default synthesis strategies without manually adjusting attributes such as emotion, pitch, or speaking rate.

All open-source models are evaluated in a zero-shot setting for long-form and dialogue generation, whereas closed-source models generated speech using designated voice profiles. Finally, all generated audio is resampled to 24kHz for consistent evaluation.

# F Supplementary Experiment

## F.1 Inference Speed

The capability to efficiently generate long-form speech is a pivotal performance criterion, garner-

Table 6: Correlation analysis among different evaluators (A denotes Annotator).

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **PLCC(↑)** | 0.8696 | 0.8426 | 0.9014 | 0.9035 | 0.9163 | 0.8766 | 0.9022 | 0.7080 | 0.8830 | 0.7623 |
| **SRCC(↑)** | 0.8711 | 0.8296 | 0.9028 | 0.9025 | 0.9143 | 0.8635 | 0.8945 | 0.7010 | 0.8820 | 0.7585 |
| **KRCC(↑)** | 0.7255 | 0.6804 | 0.7726 | 0.7678 | 0.7872 | 0.7238 | 0.7575 | 0.5399 | 0.7405 | 0.6011 |
| **QWK(↑)** | 0.8732 | 0.8330 | 0.9030 | 0.8984 | 0.9079 | 0.8544 | 0.8938 | 0.7002 | 0.8740 | 0.7596 |
| **MAE(↓)** | 0.3713 | 0.4398 | 0.3336 | 0.3452 | 0.3336 | 0.3994 | 0.3541 | 0.5800 | 0.3892 | 0.5402 |

Table 7: Sources and related information of the voice used in LFS-Bench for open-source models' inference.

| No. | Gender | Age Group | Language | Data Source |
|---|---|---|---|---|
| 1 | Female | Children | English | Emilia |
| 2 | Male | | English | Emilia |
| 3 | Female | | Chinese | ChildMandarin |
| 4 | Male | | Chinese | ChildMandarin |
| 5 | Female | Teenager | English | NCSSD_R_EN |
| 6 | Male | | English | NCSSD_R_EN |
| 7 | Female | | Chinese | AISHELL-3 |
| 8 | Male | | Chinese | NCSSD_R_ZH |
| 9 | Female | Youth-Adult | English | msppodcast |
| 10 | Male | | English | NCSSD_R_EN |
| 11 | Female | | Chinese | AISHELL-3 |
| 12 | Male | | Chinese | NCSSD_R_ZH |
| 13 | Male | | Chinese | VibeVoice Github |
| 14 | Female | | Chinese | VibeVoice Github |
| 15 | Male | | English | VibeVoice Github |
| 16 | Female | | English | VibeVoice Github |
| 17 | Female | Middle-Aged | English | LibriSpeech |
| 18 | Male | | English | Emilia |
| 19 | Female | | Chinese | NCSSD_C_ZH |
| 20 | Male | | Chinese | NCSSD_C_ZH |
| 21 | Male | | Chinese | SparkTTS Github |
| 22 | Female | Elderly | English | msppodcast |
| 23 | Male | | English | msppodcast |
| 24 | Female | | Chinese | NCSSD_C_ZH |
| 25 | Male | | Chinese | NCSSD_C_ZH |

ing widespread attention across both academia and industry. To assess this, we evaluate the computational efficiency of various open-source models using the Real Time Factor (RTF) metric. The RTF is defined as:

$$\mathrm{RTF} = \frac{T_{\text{inference}}}{T_{\text{audio}}}, \tag{5}$$

where $T_{\text{inference}}$ denotes the time required for generation and $T_{\text{audio}}$ represents the duration of the generated audio. The computational efficiency results for each model are summarized in Table 9 and Table 10. We observe that non-autoregressive models exhibit a significant advantage in generation speed compared to their autoregressive counterparts. This finding is consistent with the inherent parallel decoding mechanism of non-autoregressive architectures.

### F.2 Ablation on Window Size

The computation of both Timbre Consistency and Reverb Consistency may be sensitive to the sliding window configuration. To validate the rationality of our selected window size and stride, we conduct an ablation study across these two dimensions. The experimental results are in Table 11 and Table 12.

In the ablation study for **timbre consistency**, we observe that a window size of $\leq$ 2s results in real data exhibiting lower consistency than CosyVoice3, suggesting a misalignment with human perception. Conversely, window sizes of $\geq$ 4s gradually reduce the discrepancy between real and synthetic data, indicating that larger windows tend to average out transient timbre mutations. Regarding the stride, comparative experiments reveal no significant impact on the results. Consequently, to enhance evaluation efficiency and reduce computational overhead, we opt for a larger stride. Based on these findings, we select a window size of 3s and a stride of 2s.

In the ablation study for **reverb consistency**, a window size of 1s provides sufficient differentiation but proved unstable. Specifically, VibeVoice exhibit an excessively high standard deviation relative to its mean reverb score of 9.25, indicating hypersensitivity at this scale. Conversely, window sizes of $\geq$ 4s reduce the inter-model differences, implying that overly large windows overlook small-scale acoustic field mutations. Balancing computational efficiency and resource overhead, we similarly select a window size of 3s and a stride of 2s. Notably, our evaluation method demonstrates overall stability, as the relative rankings of the models remain consistent.

Table 8: the information of the voices selected in the evaluation for closed-source models.

| Provider | Language | Single Speaker | Two Speakers | Multi Speakers |
|---|---|---|---|---|
| **OpenAI** | General | Alloy | Onyx, Nova | Round-robin: [“alloy”, “echo”, “fable”, “onyx”, “nova”, “shimmer”] |
| **Gemini** | General | Puck | Puck, Aoede | Round-robin: [“Puck”, “Aoede”, “Charon”, “Kore”, “Fenrir”] |
| **ElevenLabs** | General | Rachel | Charlie, Rachel | Charlie, Rachel, George, Bella, Antoni |
| **Minimax** | English | male-qn-qingse | – | – |
| | Chinese | Chinese (Mandarin)_ Male_Announcer | – | – |
| **Seed-TTS** | English | BV503_streaming | – | – |
| | Chinese | BV005_streaming | – | – |
| **Seed-TTS-Podcast** | General | – | zh_male_dayixiansheng_v2_saturn_bigtts, zh_female_mizaitongxue_v2_saturn_bigtts | – |
| **Inworld** | English | Deborah, Alex | – | – |
| | Chinese | Jing, Yichen | – | – |

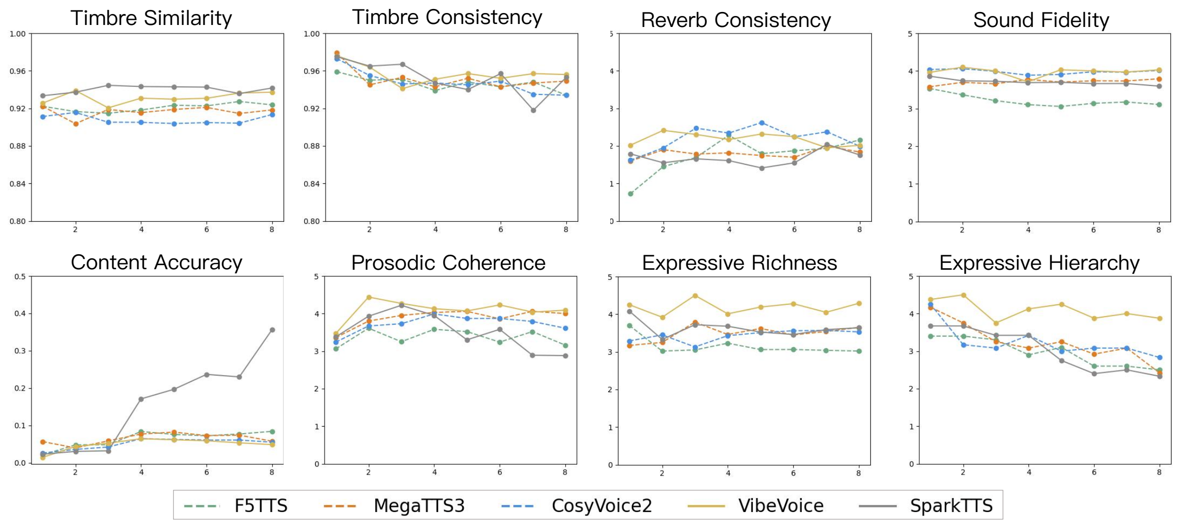



Figure 10: **Results on Sequence Length**. The horizontal axis represents the number of sentences in the text. Solid lines denote models using the End-to-End strategy, while dashed lines represent the chunked synthesis.

## F.3 Ablation on Generated Length

To further verify the impact of long-sequence modeling on acoustic, semantic, and expressive performance, we extend the analysis presented in Figure 4. Beyond the original six dimensions, we additionally track the evolution of Timbre Consistency and Timbre Similarity with respect to increasing generation length, as shown in Figure 10.

Regarding the Timbre Similarity metric, we adopt the methodology from prior works (Huynh-Nguyen et al., 2025). Specifically, the generated audio $w$ is segmented into a sequence $\{w_i\}_{i=1}^{n}$ using a window size of 3s and a stride of 2s. We then utilize WavLM TDCNN[24] to extract and normalize speaker embeddings for each segment $w_i$ and the reference audio $w_{ref}$, yielding the embedding sequence $\{e_i\}_{i=1}^{n}$ and the reference embedding $e_{ref}$. Finally, we calculate the average cosine similarity between the generated segment embeddings and the reference embedding to serve as the quantitative indicator of Timbre Similarity.

Overall, we observe a general performance decay across nearly all metrics as the generation duration increases. Specifically, Reverb Consistency, Prosodic Coherence, and Expressive Hierarchy exhibits the most significant degradation. These findings suggest that current models struggle to maintain acoustic field stability and effectively capture long-term dependencies in long-form settings. Conversely, Timbre Similarity and Timbre Consistency remained relatively stable compared to other acoustic dimensions. This stability highlights the effectiveness of “in-context learning” paradigms (Du et al., 2024; Jiang et al., 2025) in preserving speaker identity. Additionally, with the exception of Spark-TTS, most models demonstrate robust Content Accuracy. This can be attributed to the strong text understanding and alignment capabilities inherent

[24] https://huggingface.co/docs/transformers/en/model_doc/unispeech-sat

Table 9: The Real Time Factor of mono-speaker long form speech generation models.

| Models | RTF |
|---|---|
| *Autoregressive Models* | |
| CosyVoice-2 (0.5B) | $1.061 \pm 0.031$ |
| CosyVoice-3 (0.5B) | $0.747 \pm 0.048$ |
| FishSpeech (0.5B) | $1.351 \pm 0.131$ |
| GLM-TTS (1.5B) | $2.400 \pm 0.158$ |
| IndexTTS-2 (0.1B) | $1.065 \pm 0.037$ |
| SparkTTS (0.5B) | $2.046 \pm 0.212$ |
| VibeVoice (1.5B) | $3.801 \pm 0.317$ |
| *Non-Autoregressive Models* | |
| F5TTS (0.3B) | $0.198 \pm 0.006$ |
| MegaTTS3 (0.45B) | $0.172 \pm 0.002$ |
| ZipVoice (0.12B) | $0.338 \pm 0.013$ |

Table 10: The Real Time Factor of two-speaker dialogue generation models. MOSS-TTSD supports batch inference, thus we directly report the RTF of batch process( batchsize = 32)

| Models | RTF |
|---|---|
| FireRedTTS2 | $4.717 \pm 0.963$ |
| MoonCast (2.6B) | $5.219 \pm 0.048$ |
| MOSS-TTSD (1.7B) | $0.219 \pm 0.019$ |
| SoulX-PodCast (1.7B) | $2.143 \pm 0.169$ |
| VibeVoice (1.5B) | $4.092 \pm 0.305$ |
| ZipVoice-Dialog (0.12B) | $0.305 \pm 0.030$ |

Table 11: The Ablation study of window setting for timbre consistency. We select the representative models, CosyVoice3 and OpenAI-tts-1-hd, to conduct this ablation in single-spaeker settings.

| Window Setting | | CosyVoice3 | OpenAI | Real-Speech |
|---|---|---|---|---|
| Size (s) | Stride (s) | | | |
| 1 | 0.5 | 0.868 | 0.824 | 0.844 |
| 2 | 1 | 0.911 | 0.887 | 0.901 |
| 3 | 1 | 0.930 | 0.916 | 0.956 |
| 3 | 2 | 0.929 | 0.915 | 0.955 |
| 4 | 2 | 0.941 | 0.931 | 0.963 |
| 5 | 2 | 0.942 | 0.949 | 0.967 |
| 10 | 4 | 0.968 | 0.971 | 0.971 |

Table 12: The Ablation study of window setting for reverb consistency. We select the representative models, VibeVoice and Gemini-2.5-pro-preview-tts, to conduct this ablation in two-spaeker settings.

| Window Setting | | VibeVoice | Gemini | Real-Dialog |
|---|---|---|---|---|
| Size (s) | Stride (s) | | | |
| 1 | 0.5 | 6.40 | 4.99 | 3.87 |
| 2 | 1 | 4.27 | 3.62 | 3.20 |
| 3 | 1 | 3.58 | 3.17 | 2.67 |
| 3 | 2 | 3.59 | 3.17 | 2.74 |
| 4 | 2 | 3.20 | 2.85 | 2.51 |
| 5 | 2 | 2.95 | 2.61 | 2.41 |
| 10 | 4 | 2.23 | 1.88 | 1.60 |

in modern TTS architectures.

### F.4 Multi-Speaker Dialogue Generation

To facilitate future research in multi-speaker long-form speech synthesis, LFS-Bench incorporates 101 test cases specifically designed for 3- and 4-speaker dialog scenarios. Using this subset, we evaluate three closed-source models capable of multi-speaker generation: ElevenLabs Multilingual V2, Gemini-2.5-pro-preview-tts, and OpenAI-tts-1-hd. The experimental results are shown in Table 13.

## G More Analysis Based on LFS-Bench

### G.1 Detailed analysis on each metric

**Timbre Consistency** Although experimental results indicate that real data generally outperforms synthetic data in timbre consistency (single speaker: 0.96 vs. 0.93; two-speaker: 0.95 vs. 0.92), this gap is not significant. This suggests that the consistency performance of current models is acceptable. However, we offer two deeper insights. First, open-source models exhibit a relatively larger standard deviation compared to closed-source models, indicating that their stability still lags behind commercial solutions. Second, dialogue models demonstrate greater variance in timbre consistency than single-speaker long-form speech. Given that we have minimized error accumulation from forced alignment, this increased variance likely reflects that models are still hindered by speaker transitions.

**Reverb Consistency** In this dimension, single-speaker performance is comparable to human recordings. Apart from the CosyVoice series and ElevenLabs models, which underperform on this metric, other models maintain robust reverb consistency, demonstrating strong acoustic field preservation over extended durations. Conversely, in dialogue scenarios, all open-source models and the majority of closed-source models show a significant performance gap compared to real data (Open average: 3.45; Closed average: 3.36). Feedback from

Table 13: **Results of multi-speaker dialogue generation models across LFS-Bench's metrics**. The best results are in **bold** and the second best are underlined.

| Model | Acoustics | | | Semantics | | Expressiveness | |
|---|---|---|---|---|---|---|---|
| | Timbre(↑) | Reverb(↓) | Sound Fidelity(↑) | CER/WER(↓) | Prosody(↑) | Richness(↑) | Hierarchy(↑) |
| Elevenlabs Multilingual V2 | **0.93±0.030** | 4.72±0.69 | **3.19±0.37** | 0.183 / 0.12 | 3.28±0.87 | 3.23±0.54 | 3.52±0.82 |
| Gemini-2.5-pro-preview-tts | 0.92±0.012 | 3.28±0.75 | 3.04±0.17 | **0.077 / 0.102** | **3.92±0.36** | **3.86±0.46** | **4.05±0.62** |
| OpenAI-tts-1-hd | 0.92±0.011 | **1.91±0.38** | 2.29±0.17 | 0.106 / 0.104 | 3.78±0.63 | 2.93±0.60 | 3.77±0.84 |
| Average | 0.92 | 3.30 | 2.84 | 0.122 / 0.109 | 3.66 | 3.34 | 3.78 |

our user study further reveals inconsistencies in sound fields and volume between speakers in generated dialogues. This indicates a need to enhance the models' ability to disentangle prompt speech attributes. Consequently, future work should prioritize maintaining acoustic unity during speaker transitions.

**Sound Fidelity** Regarding this metric, the performance of generated speech aligns closely with that of real data. Notably, models such as FishSpeech and ElevenLabs achieve scores significantly surpassing the mean of real data. This suggests that contemporary models have largely resolved sound quality constraints. The fact that generated speech outperforms human recordings likely stems from the composition of the real data. Since the majority of real data is web-crawled rather than studio-recorded, it is susceptible to device and environmental noise, which compromises its fidelity.

**Content Accuracy** Prior studies indicate that metrics such as WER have reached saturation in sentence-level speech generation (Chen et al., 2024b). This finding extends to chunk-based in-context learning approaches, where models like CosyVoice2 and MegaTTS3 demonstrate exceptional performance. However, the metric remains relevant for autoregressive end-to-end architectures. For instance, SparkTTS exhibits suboptimal Content Accuracy in long-form generation. As in Figure 10, deeper ablation studies confirm that the character accuracy of such models declines as the text length increases.

**Prosodic Coherence** Regarding prosodic coherence, we observe a distinct gap between real and synthetic speech, suggesting that current models require further improvement in prosody modeling. Notably, closed-source models significantly outperform their open-source counterparts in this dimension. This indicates that while open-source models achieve parity with state-of-the-art systems in fidelity and content accuracy, they still lag in perceptual metrics such as prosodic naturalness.

**Expressive Richness** Experimental results identify expressiveness as the primary differentiator between real and synthetic audio. Specifically, open-source models trail real data by approximately 1.5 points in Expressive Richness. While closed-source models demonstrate marked improvement, they still exhibit a gap of nearly 1.0 point. Furthermore, our scenario-based analysis confirms that models underperform in high-expressiveness settings. These findings consistently underscore that generating realistic, highly expressive speech remains a pivotal challenge for achieving immersive audio generation.

**Expressive Hierarchy** Similar to Expressive Richness, real data outperforms synthetic speech in this metric, with closed-source models surpassing their open-source counterparts. Notably, in single-speaker tasks, models consistently achieve lower scores on Expressive Hierarchy compared to Expressive Richness. This indicates that capturing and modeling paragraph-level hierarchical structure remains a significant challenge. Furthermore, dialog models generally exhibit superior hierarchical performance compared to single-speaker models. We attribute this to the inherent semantic logic of dialog interactions, which likely provides stronger contextual cues that facilitate the learning of hierarchical patterns.

### G.2 Analysis based on the scenarios

We extend our analysis by providing scenario-based performance results, visualizing the metrics of closed-source models via a radar chart in Figure 11. These detailed findings corroborate our primary conclusion: most metrics exhibit varying degrees of degradation in high-expressiveness scenarios. A granular visualization reveals that challenging scenarios such as sportscast, host, and talk-show suffer the most severe performance decline. This further indicates that current models lack the capacity to effectively model highly dynamic prosody and intense emotional variations.

We provide a detailed explanation of the normalization procedures applied to the radar charts in Figure 11. For LALM-based metrics (Expressive Richness, Expressive Hierarchy, Prosodic Coherence), we directly utilize the original values as its definition is consistent with that of MOS scores. For Fidelity, quantified by `SQUIM-PESQ` (range $[-0.5, 4.5]$), we apply a linear shift of $+0.5$ for alignment. Regarding Timbre Consistency, Reverb Consistency, and Content Accuracy, we first identify the global maximum $s_{\max}$ and minimum $s_{\min}$ across all models in all scenarios. Then, we employ a mapping function $f$ that projects the range $[s_{\min}, s_{\max}]$ onto the interval $[1, 5]$. This transformation ensures that for all dimensions in the radar chart, a larger value consistently represents superior performance. The radar charts in Figure 3 and Figure 1 follow this identical normalization protocol.

### G.3 Analysis based on the Languages

We also present the experimental results for the evaluated models across the two covered languages, Chinese and English, as shown in Table 14 and Table 15.

We observe that although all evaluated models claim bilingual capabilities, the target language significantly impacts performance for the majority. For instance, despite utilizing identical voice profiles, ElevenLabs Multilingual V2 exhibits a marked disparity in Expressive Richness between Chinese and English (1.79 vs. 2.87). A similar divergence is evident in Seed-TTS-Podcast (Chinese: 4.19 vs. English: 3.49). In contrast, Gemini-2.5-pro-preview-tts stands out by not only delivering exceptional performance in prosody and expressiveness but also maintaining a consistent balance across both languages.

## H Future Works

While LFS-Bench provides a comprehensive evaluation framework for long-form speech generation, several challenges warrant further exploration:

**Dependency on Closed-source Models**: The evaluation of Expressiveness in LFS-Bench currently relies on closed-source models such as Gemini 3 Pro. The absence of open-source alternatives poses a risk to reproducibility due to potential updates in closed-source APIs. Future work will focus on distilling high-performance open-source evaluators using data derived from both human assessments and closed-source model outputs.

**Limited Language Coverage**: Our current dataset focuses exclusively on English and Chinese, omitting other languages, particularly low-resource ones. Future efforts should aim to expand the linguistic breadth of long-form speech generation evaluation.

**Timbre Sensitivity**: To ensure diversity, LFS-Bench utilizes over 20 reference voices spanning various genders and ages. However, as noted in prior work (Manku et al., 2025), model performance in expressiveness and prosody is highly sensitive to the reference voice. Our current selection may not be sufficiently diverse. Future research should investigate the impact of input voice characteristics on long-form synthesis more deeply.

**Instruction Following Capabilities**: LFS-Bench primarily evaluates models in zero-shot settings. However, recent advancements have introduced models capable of Instruct-based speech generation (Huang et al., 2025; Wang et al., 2025; Zhou et al., 2025b; Xu et al., 2025b). Developing long-form InstructTTS systems and evaluating their instruction-following capabilities in long-context settings represent significant avenues for future research.

## I Social Impacts

This work aims to advance immersive and robust long-form speech generation, facilitating superior downstream applications. However, enhanced generative capabilities inevitably heighten the risk of misuse, potentially violating ethical norms and legal regulations. These risks highlight the critical need for ethically aligned practices and sufficient oversight. To mitigate these concerns, we subjected our text data to rigorous ethical review and anonymization. We also verified that the accompanying audio samples are free of Personally Identifiable Information (PII). Additionally, we mandate that all researchers utilizing this benchmark strictly adhere to the CC BY-NC-SA 4.0 license. We hope that the progress in speech generation technology will benefit society through responsible and ethical deployment.

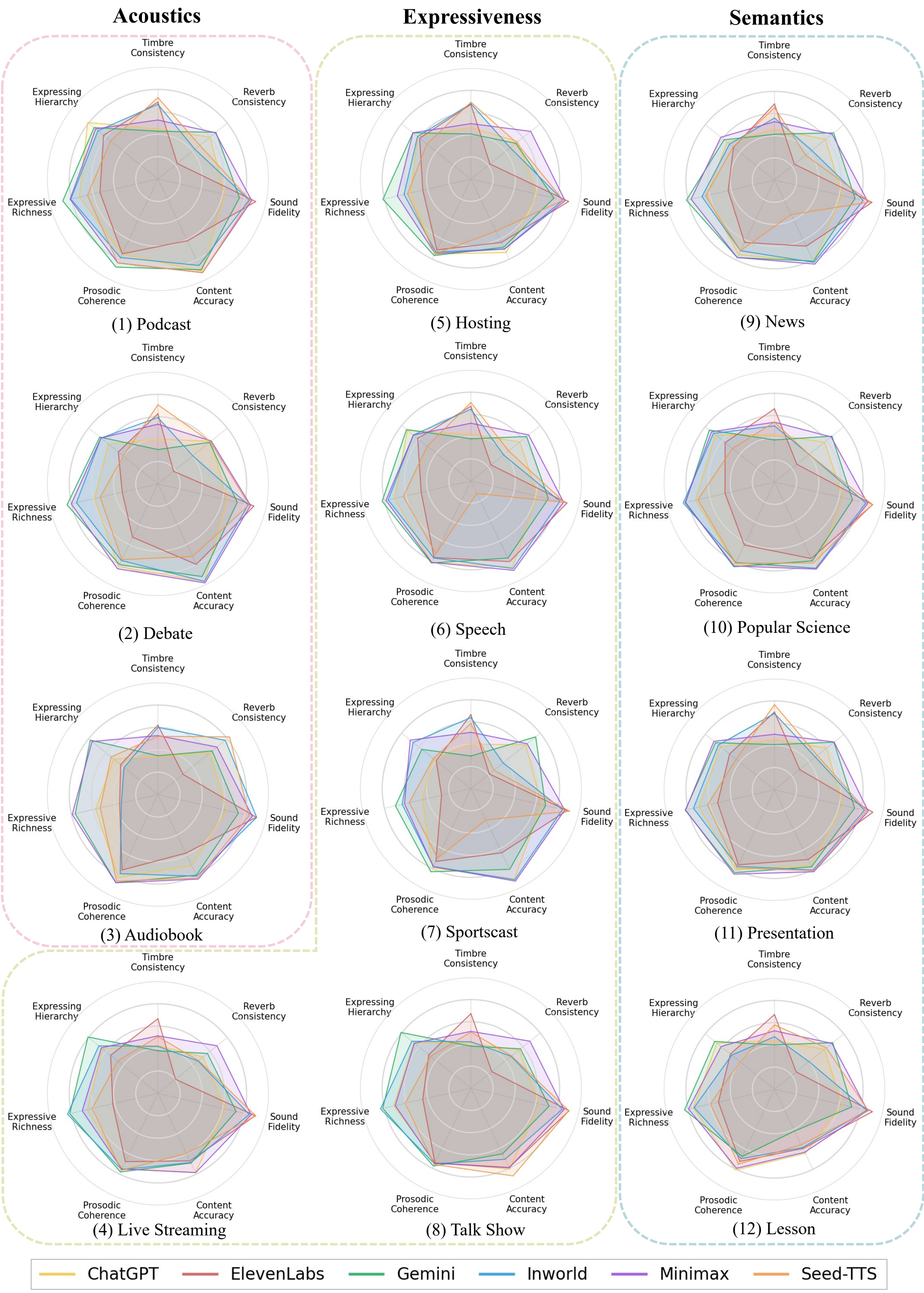



Figure 11: We visualize the performance of closed-source models in single-speaker long-form generation across various downstream scenarios using a radar chart. To ensure consistency, we normalize the metrics for Timbre Consistency, Reverb Consistency, and Content Accuracy within their respective minimum and maximum ranges. As a result, all metrics are presented such that higher values indicate better performance.

Table 14: **Evaluation results of long-form TTS models across two languages.** Metrics cover Acoustics (Timbre/Reverb Consistency, Fidelity), Semantics (Content Accuracy, Prosodic Coherence), and Expressiveness (Richness, Hierarchy). Closed-source models and open-source models are separately marked, with the best results in **bold** and the second best *italic*. Chinese results and English results are separately marked as well, with Chinese in black and English in red.

| Models | Languages | Acoustics | | | Semantics | | Expressiveness | |
|---|---|---|---|---|---|---|---|---|
| | | Timbre(↑) | Reverb(↓) | Fidelity(↑) | Content(↓) | Prosody(↑) | Richness(↑) | Hierarchy(↑) |
| *Open-Source Models* | | | | | | | | |
| SparkTTS | ZH | 0.90 | **0.79** | 3.47 | 0.329 | 2.37 | 3.29 | 2.11 |
| | EN | **0.95** | 2.96 | 3.70 | 0.240 | 2.78 | *3.64* | 2.65 |
| ZipVoice | ZH | 0.90 | 1.65 | 3.55 | 0.072 | 3.24 | 3.16 | 2.87 |
| | EN | 0.89 | 2.47 | 3.47 | 0.396 | 3.13 | 1.71 | 1.34 |
| GLM-TTS | ZH | 0.93 | 1.52 | *3.99* | 0.035 | **4.07** | 3.17 | 3.12 |
| | EN | *0.94* | *1.70* | 3.90 | 0.118 | 3.21 | 2.19 | 1.96 |
| CosyVoice2 | ZH | 0.90 | 1.74 | 3.57 | **0.032** | 3.62 | 3.47 | 3.13 |
| | EN | 0.93 | 2.95 | *4.02* | 0.168 | 2.84 | 2.56 | 2.39 |
| CosyVoice3 | ZH | *0.94* | 1.83 | 3.83 | *0.034* | 3.92 | 3.36 | 2.83 |
| | EN | 0.93 | 2.68 | 3.82 | 0.141 | 2.83 | 2.23 | 2.07 |
| MegaTTS3 | ZH | 0.93 | 2.12 | 3.67 | 0.035 | 3.92 | 3.02 | 2.88 |
| | EN | 0.93 | **1.50** | 3.43 | **0.108** | 3.30 | 2.60 | 2.17 |
| IndexTTS2 | ZH | **0.95** | 1.28 | 2.39 | *0.033* | 3.96 | **4.02** | *3.30* |
| | EN | 0.92 | 2.15 | 3.15 | 0.135 | 3.33 | 3.15 | 2.62 |
| FishSpeech | ZH | 0.92 | 1.76 | **4.06** | 0.043 | *4.03* | 3.25 | 3.16 |
| | EN | 0.93 | 1.81 | **4.13** | 0.113 | *3.56* | 2.06 | 2.63 |
| VibeVoice | ZH | 0.91 | 1.54 | 3.88 | 0.047 | 3.91 | 3.47 | **3.34** |
| | EN | **0.95** | 2.75 | 3.75 | *0.111* | **3.88** | **3.95** | **3.34** |
| F5TTS | ZH | 0.88 | *1.13* | 3.12 | 0.072 | 3.28 | *3.50* | 2.73 |
| | EN | 0.92 | 2.51 | 3.65 | 0.113 | 3.54 | 2.64 | *2.81* |
| Average | ZH | 0.92 | 1.54 | 3.55 | 0.073 | 3.63 | 3.37 | 2.95 |
| | EN | 0.93 | 2.35 | 3.70 | 0.164 | 3.24 | 2.67 | 2.40 |
| *Closed-Source Models* | | | | | | | | |
| gemini-2.5-pro-preview-tts | ZH | 0.90 | **1.38** | 3.13 | 0.059 | *4.13* | **4.20** | *3.53* |
| | EN | 0.92 | *1.49* | 3.19 | 0.169 | 3.69 | **4.07** | **3.48** |
| OpanAI-tts-1-hd | ZH | 0.91 | 1.65 | 2.69 | 0.043 | 4.00 | 3.20 | 3.07 |
| | EN | 0.92 | 1.82 | 2.60 | 0.119 | **3.82** | 3.71 | *3.43* |
| MiniMax-Speech-2.6-hd | ZH | 0.93 | *1.43* | 3.83 | **0.030** | **4.14** | *4.00* | **3.56** |
| | EN | 0.92 | **1.32** | 3.81 | 0.119 | *3.77* | 3.60 | 2.95 |
| Elevenlabs Multilingual V2 | ZH | **0.95** | 3.04 | **4.00** | 0.100 | 3.26 | 1.79 | 2.38 |
| | EN | **0.96** | 3.05 | **4.04** | *0.115* | 3.73 | 2.87 | 2.97 |
| Inworld-tts-1-max | ZH | *0.94* | 2.19 | 3.72 | 0.053 | 3.73 | 3.41 | 2.92 |
| | EN | 0.92 | 2.19 | 3.74 | **0.114** | 3.69 | *3.95* | 3.13 |
| Seed-TTS2 | ZH | *0.94* | 1.99 | *3.86* | 0.106 | 3.86 | 3.06 | 2.46 |
| | EN | *0.94* | 1.91 | *3.89* | 0.193 | 3.62 | 3.14 | 2.21 |
| Average | ZH | 0.93 | 1.95 | 3.54 | 0.065 | 3.85 | 3.28 | 2.99 |
| | EN | 0.93 | 1.96 | 3.55 | 0.138 | 3.72 | 3.56 | 3.03 |

Table 15: **Evaluation results of dialog generation models across two languages.** Metrics cover Acoustics (Timbre/Reverb Consistency, Fidelity), Semantics (Content Accuracy, Prosodic Coherence), and Expressiveness (Richness, Hierarchy). Closed-source models and open-source models are separately marked, with the best results in **bold** and the second best *italic*. Chinese results and English results are separately marked as well, with Chinese in black and English in red.

| Models | Languages | Acoustics | | | Semantics | | Expressiveness | |
|---|---|---|---|---|---|---|---|---|
| | | Timbre(↑) | Reverb(↓) | Fidelity(↑) | Content(↓) | Prosody(↑) | Richness(↑) | Hierarchy(↑) |
| *Open-Source Models* | | | | | | | | |
| ZipVoice | ZH | 0.90 | 3.15 | 2.65 | *0.069* | **4.01** | 3.01 | 2.87 |
| | EN | 0.91 | 3.91 | 2.67 | *0.114* | 3.34 | 2.24 | 2.72 |
| MoonCast | ZH | 0.89 | *3.11* | 2.56 | 0.313 | 3.25 | 2.58 | 2.60 |
| | EN | 0.91 | **3.01** | 2.68 | 0.125 | 3.08 | 2.78 | 2.79 |
| FireRedTTS2 | ZH | **0.92** | *3.32* | 3.16 | 0.075 | *3.57* | 3.16 | 3.03 |
| | EN | *0.93* | *3.64* | 2.08 | 0.131 | 2.91 | 2.29 | 2.58 |
| MOSS-TTSD | ZH | 0.90 | **3.02** | 3.13 | 0.148 | 3.10 | *3.66* | *3.26* |
| | EN | 0.91 | 4.07 | 2.64 | 0.239 | 2.47 | 2.75 | 2.71 |
| VibeVoice | ZH | *0.90* | 3.26 | *3.32* | 0.106 | 3.48 | **3.74** | 3.34 |
| | EN | 0.91 | 3.91 | *3.38* | 0.125 | *3.66* | **3.78** | *3.39* |
| SoulXPodcast | ZH | **0.92** | 3.31 | **3.94** | **0.061** | **4.01** | 3.69 | **3.82** |
| | EN | **0.94** | 3.70 | **3.98** | **0.090** | **4.00** | *3.18* | **3.59** |
| Average | ZH | 0.91 | 3.20 | 3.13 | 0.129 | 3.42 | 3.31 | 3.15 |
| | EN | 0.92 | 3.71 | 3.07 | 0.154 | 3.24 | 2.84 | 2.96 |
| *Closed-Source Models* | | | | | | | | |
| Gemini-2.5-pro-preview-tts | ZH | 0.91 | 3.07 | 3.05 | *0.086* | *4.12* | *4.10* | *4.11* |
| | EN | **0.93** | 3.26 | 2.96 | **0.092** | **4.00** | **4.02** | **3.93** |
| OpenAI-tts-1-hd) | ZH | *0.92* | *2.97* | 2.26 | 0.104 | 3.52 | 3.17 | 3.56 |
| | EN | **0.93** | **2.99** | 2.29 | *0.103* | 3.86 | 3.41 | *3.84* |
| Elevenlabs Multilingual V2) | ZH | **0.93** | 4.55 | *3.38* | 0.127 | 3.44 | 2.32 | 3.11 |
| | EN | **0.93** | 4.31 | *3.58* | 0.109 | *3.89* | 3.36 | 3.81 |
| Seed-TTS-Podcast | ZH | *0.92* | **2.48** | **3.90** | **0.063** | **4.16** | **4.19** | **4.26** |
| | EN | *0.91* | *3.22* | **3.88** | 0.108 | 3.70 | *3.49* | 3.42 |
| Average | ZH | 0.92 | 3.27 | 3.15 | 0.095 | 3.81 | 3.45 | 3.76 |
| | EN | 0.93 | 3.45 | 3.18 | 0.103 | 3.86 | 3.57 | 3.75 |

## Prompt for Prosody Coherence

**Role:** Senior Linguistic Expert & Prosody Analyst. You are an expert in assessing speech naturalness, with a hyper-sensitivity to prosodic coherence, rhythmic hierarchy, and robotic artifacts.
**Input Data:**

- **Target Text:** The reference text script that needs to be synthesized.
- **Audio Output:** The speech audio generated by the TTS model (labeled as Output A).

**Generation Requirements:**

1. **Core Task**: Evaluate the audio's naturalness by analyzing its **prosodic structure** and **coherence** against the target text, rather than just audio quality.
2. **Dimension 1 - Prosody Coherence & Flow**: Assess the smoothness of the speech stream. Check for unnatural pauses, abrupt disjoints between words/phrases, and the logical flow of intonation across sentence boundaries.
3. **Dimension 2 - Rhythmic Hierarchy & Layering**: Evaluate the structural stress patterns. Does the speaker correctly emphasize content words while de-emphasizing function words? Is there a natural "melody" (intonation contour) rather than a flat or repetitive beat?
4. **Dimension 3 - Overall Naturalness**: Check for presence of human-like micro-prosody (e.g., breathiness, slight pitch variations).
5. **Format**: Strictly output a valid JSON object. No other text.

**Scoring Guidelines (1.0–5.0, step of 0.5):**

- **5.0 (Human-Parity):** Indistinguishable from a professional human speaker; perfect coherence and rich prosodic hierarchy.
- **4.0 (Natural):** Very smooth and pleasant; minor prosodic flaws only noticeable to experts; good structural layering.
- **3.0 (Acceptable):** Intelligible and decent flow; but lacks depth (flat hierarchy) or contains audible TTS artifacts.
- **2.0 (Mechanical):** Disjointed flow; unnatural pauses; wrong stress placement (e.g., stressing every word equally).
- **1.0 (Robotic):** Completely lifeless; broken prosody; difficult to listen to.

**JSON Schema:**

```
{
    "Overall_Impression": "[Brief summary of naturalness and flaws]",
    "Detailed_Analysis": {
        "Coherence_and_Flow": "[Critique the smoothness and connection...]",
        "Hierarchy_and_Layering": "[Analyze stress patterns and intonation curves...]",
        "Naturalness": "[Comments on naturalness]"
    },
    "Score": [Number 1.0-5.0],
}
```

Figure 12: Structured prompt for evaluating long-form audio's performance in Prosody Coherence.

## Prompt for Expressive Hierarchy

**Role:** Senior Voice Director & Audio Engineer (Long-Form Specialist). You are an expert in long-form narration (audiobooks, documentaries), hyper-sensitive to monotony, repetitive patterns, and lack of structural progression.
**Generation Requirements:**

1. **Core Task**: Analyze how the performance **evolves over time**, focusing on "Layering and Hierarchy".
2. **Dimension 1 - Emotional Variation & Arc**: Evaluate progression from beginning to end, distinction between climax and exposition, and avoidance of "one-note" acting.
3. **Dimension 2 - Vocal Dynamics**: Check for macro/micro dynamics (volume/tempo shifts).
4. **Dimension 3 - Scene Appropriateness & Structural Fit**: Assess contextual adaptation to content structure and long-term engagement.
5. **Format**: Strictly output a valid JSON object. No other text.

**Scoring Guidelines (1.0–5.0, step of 0.5):**

- **5.0 (Masterful):** A journey with rich variety; no repetitive patterns; perfect for long listening.
- **4.0 (Strong):** Good dynamics and clear emotional shifts; avoids obvious monotony.
- **3.0 (Acceptable but Static):** Pleasant but lacks progression; risks boring the listener over time.
- **2.0 (Repetitive):** Clear signs of "looping prosody"; same intonation for every sentence.
- **1.0 (Robotic):** Lifeless; no dynamic range or emotional change; raw TTS-like.

**JSON Schema:**

```
{
  "Overall_Impression": "[A brief summary of the long-form experience]",
  "Hierarchy_Analysis": {
    "Emotional_Arc": "[Describe the emotional progression...]",
    "Dynamics_and_Rhythm": "[Critique the pacing and prosody...]",
    "Scene_Fit": "[How well does it adapt to the structure?]"
  },
  "Score": [Number 1.0-5.0],
  "Final_Recommendation": "[Highly Recommended / Recommended with Reservations / Not Recommended]"
}
```

Figure 13: Structured prompt for evaluating long-form audio performance, focusing on expressive hierarchy.

**Prompt for Expressive Richness**

**Role:** You are a Senior Voice Director and Audio Engineer with standards equivalent to a top-tier animation studio. Your task is to meticulously evaluate a voice recording and determine if it meets professional standards.
**Evaluation Dimension: Performance & Expressiveness**

- **Emotional Resonance:** Genuine, layered emotion vs. flat/forced.
- **Character Portrayal:** Believable, consistent character; tone/age/personality coherence.
- **Storytelling & Immersion:** Narrative flow, atmosphere, and engagement.

**Exclusions:** Ignore sudden stop, audio quality, timbre consistency, and pronunciation accuracy.
**Scoring Guidelines (1.0–5.0):**

- **5.0 (Outstanding):** Richly expressive, immersive, and artistically elevated.
- **4.0 (Strong):** High expressiveness, close to professional but lacks fine nuance.
- **3.0 (Adequate):** Meets basic requirements; emotions may be somewhat generic.
- **2.0 (Flat):** Unconvincing, weak emotional expression, clearly subpar.
- **1.0 (Mechanical):** Synthetic/lifeless, no emotional color or dynamics.

**JSON Schema:**

```
{
    "Overall_Impression": "A brief, one-sentence summary of the audio.",
    "Expressiveness": "Detailed professional analysis of the performance dimension.",
    "Expressiveness_Score": [Number between 1.0 and 5.0 in 0.5 increments],
    "Final_Recommendation": "[Highly Recommended / Recommended with Reservations / Not Recommended]"
}
```

Figure 14: The structured prompt used for professional voice performance and expressiveness assessment.