

Evaluating Expressive Multilingual TTS

ZTTS1-Eval

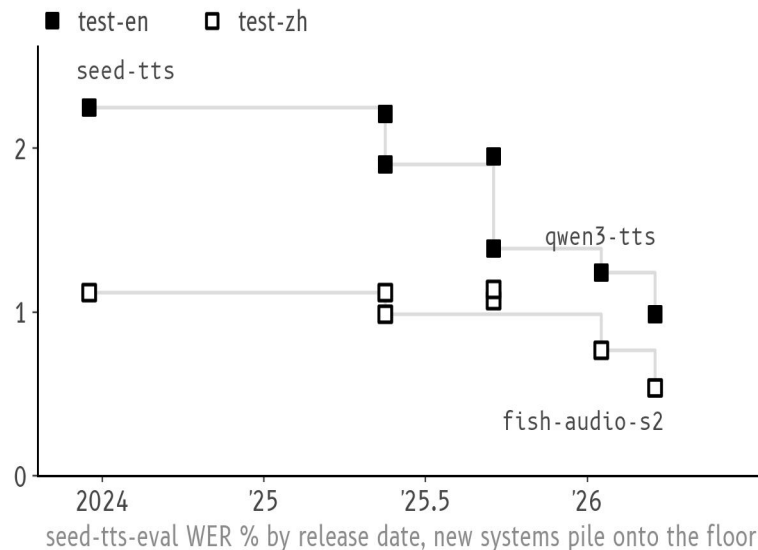
Sofian Mejjoute, Cartesia Research Talk, August 21st 2026

Contents

- Previous Work: Seed-TTS Eval, CV3-Eval, Others
- ZTTS1-Eval: Clean Set, In-the-wild Set, Updated Metrics, Prosody
- Future Work: Sample size, Improving upon TTSDS2, More filtering, Longform/Multi-turn

Benchmark saturation

- Two languages, Chinese and English only
- Dated scoring models: Whisper-Large, Paraformer, WavLM
- Read speech only: Common Voice, DiDiSpeech
- WER alone no longer separates the strongest systems



Previous Work

Seed-TTS Eval

- Languages: Chinese, English
- Data: DiDiSpeech-2, Common Voice
- Size: 2,000 zh, 1,000 en (+ zh hard cases); 3 h
- Intelligibility: Paraformer-zh, Whisper-large-v3
- Speaker similarity: WavLM-large (SV), cosine
- Issues: only zh/en; outdated metric models; read speech, varied recording environments

Seed-TTS Eval: metric models

- Paraformer-zh: WER, Chinese (outdated)
- Whisper-large-v3: WER, English (outdated)
- WavLM-SV: speaker similarity, cosine (outdated)
- Audio: Common Voice (en), DiDiSpeech (zh), read speech

CV3-Eval

- Languages: Chinese, English, Japanese, Korean, German, French, Russian, Italian, Spanish
- Data: Common Voice, FLEURS (unfiltered noise and silence)
- Intelligibility: Paraformer-zh, Whisper-Large
- Speaker / emotion / quality: ERes2Net, emotion2vec, DNSMOS
- Hard cases: zh/en rare words, tongue twisters, domain terms
- Subtasks: multilingual cloning (used), cross-lingual, emotion, expressive, accented (unused)

CV3-Eval: metric models

- Paraformer-zh, Whisper-Large: WER (outdated)
- ERes2Net: speaker similarity
- emotion2vec: emotion (emo2vec-large-plus on the emotion-cloning subset)
- DNSMOS: quality

Others

- Minimax Multilingual Testset: 24 languages, two read Common Voice prompts each, Seed-TTS-Eval scoring
- EmergentTTS-Eval: 1,645 LLM-evolved cases, six scenarios, audio-LLM judge
- Audio Turing Test: 1,000-sample Chinese corpus, trap utterances, “does it sound human?” instead of MOS
- Long-TTS-Eval: long-form generation quality, en/zh, six content categories

ZTTS1-Eval

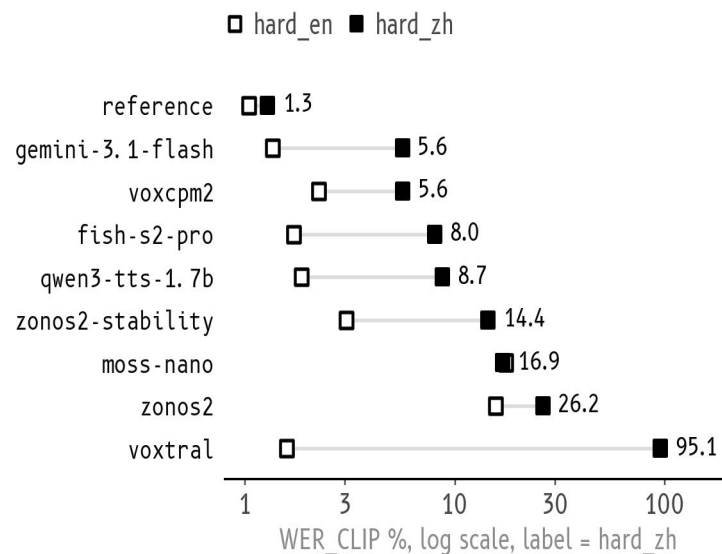
Two new evaluation sets as well as an updated and extended evaluation protocol.

Clean Set

- Source: FLEURS-R
- Languages: en, zh, de, es, fr, it, ja, ko, ru
- Size: 13 h, 500 utterances per language
- Text: structured read-aloud prepared speech
- Hard Set: hard portions for complex en/zh

Hard Set

- Languages: English, Chinese
- Origin: CV3-Eval hard-case texts, retained
- Size: 64 en, 60 zh utterances
- Reference WER: hard_en 1.04, hard_zh 1.28

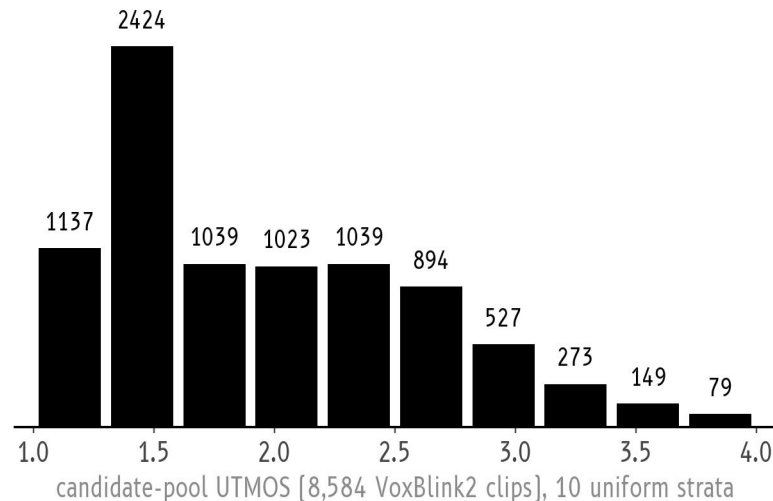


In-the-wild Set

- Source: VoxBlink2 (data CC-BY-NC-SA 4.0)
- Size: 1,618 utterances, ~2.86 h, 17 languages, 84–120 per language
- Speech: conversational, natural, spontaneous
- Pairing: prompt and target from different videos of the same speaker
- MOS Stratification: 10 UTMOS bins, 1–4

MOS Stratification

- Quality score: UTMOS on prompt audio
- Strata: 10 uniform bins spanning UTMOS 1–4
- 84–120 prompts per language
- Purpose: clean prompts must not dominate the distribution



The two sets

	Clean	In-the-wild
Source	FLEURS-R	VoxBlink2
Size	13 h, 500 utt per language	~2.86 h, 84–120 utt per language
Languages	9	17
Speech	read-aloud prepared	spontaneous conversational
Pairing	same recording session	two different videos, same speaker
Hard splits	hard_en 64, hard_zh 60	—

Updated Metrics

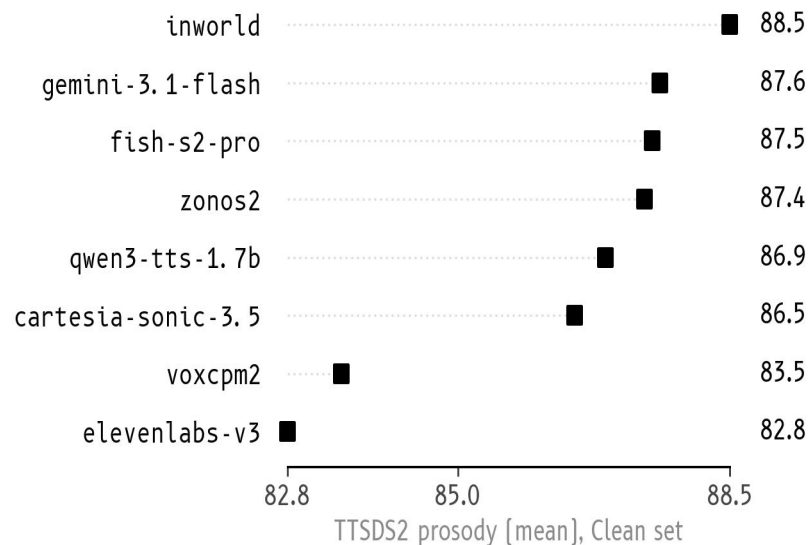
	Previous	ZTTS1-Eval
Intelligibility	Whisper-large-v3, Paraformer-zh	Qwen3-ASR-1.7B
Speaker similarity	WavLM-large / ERes2Net	ReDimNet
Quality	— / DNSMOS	MSR-UTMOS
Emotion	emotion2vec	—
Prosody	—	TTSDS2
Diversity	—	DS-WED

Updated Metrics: models

- Intelligibility: Qwen3-ASR-1.7B (WER_CLIP)
- Speaker similarity: ReDimNet2 (cosine x 100)
- Quality: MSR-UTMOS
- Prosody and diversity: TTSDS2, DS-WED
- WER_CLIP caps per-utterance WER at 100% before averaging
- Ground truth scored through the same pipeline as a reference floor

Prosody: TTSDS2

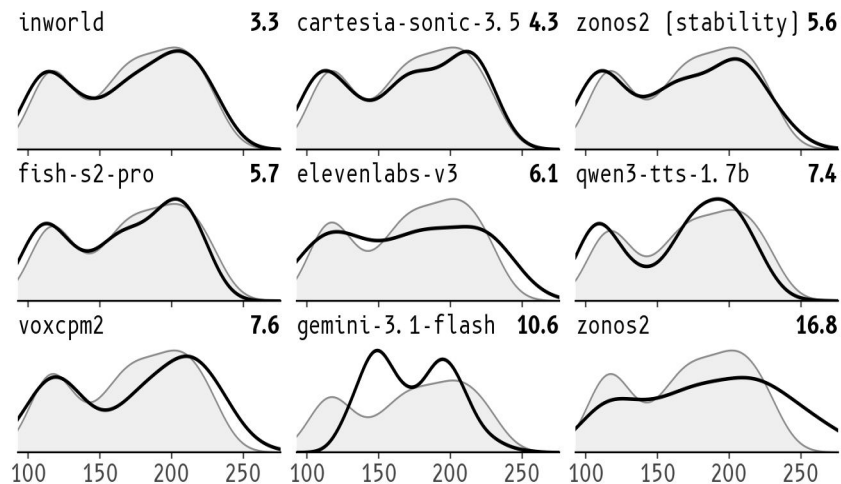
- Target: prosody distribution matching
- Avg Pitch: mean F0 (WORLD)
- Hubert Unit Rate: semantic units per second
- Allosaurus Phoneme Rate: phonemes per second
- MPM: Masked Prosody Model
- Scoring: distributional distance, normalized by real-to-noise



Avg Pitch

- Definition: mean F0 distribution per utterance
- Extractor: WORLD vocoder pitch contour
- Scoring: 2-Wasserstein vs real reference

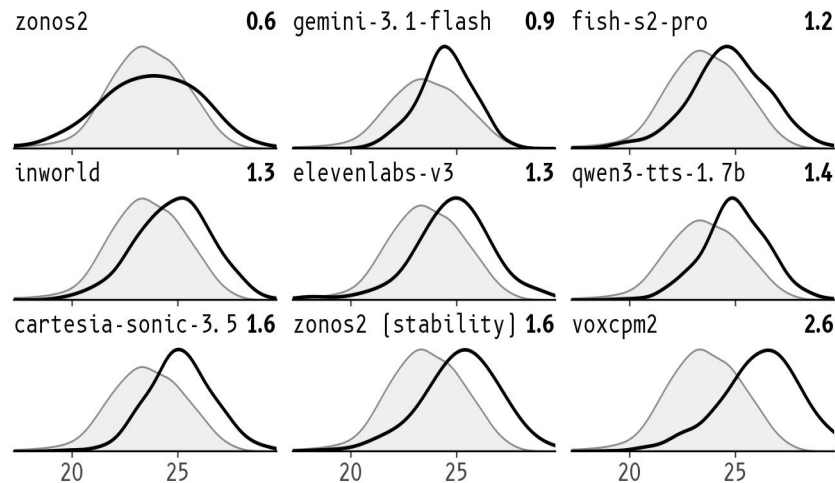
mean F0 [Hz], gray = reference, number = W1 to reference



Hubert Unit Rate

- Definition: deduplicated HuBERT tokens per second (speaking rate)
- Extractor: mHuBERT-147 (multilingual)
- Scoring: 2-Wasserstein vs real reference

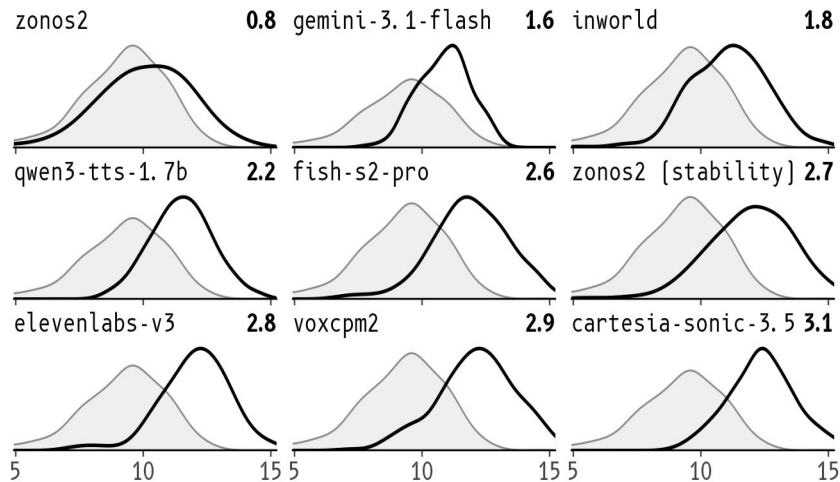
HuBERT units / s, gray = reference, number = W1 to reference



Allosaurus Phoneme Rate

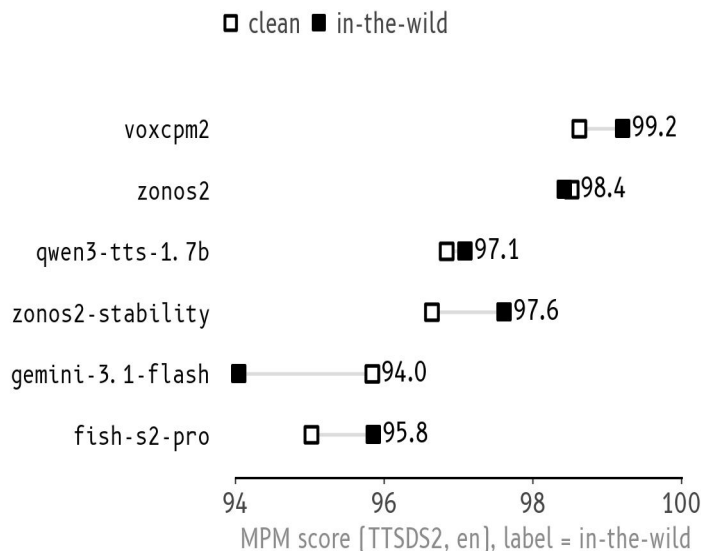
- Definition: phones per frame, speaking rate from universal phone recognition
- Extractor: Allosaurus
- Scoring: 2-Wasserstein vs real reference

phonemes / s, gray = reference, number = W1 to reference



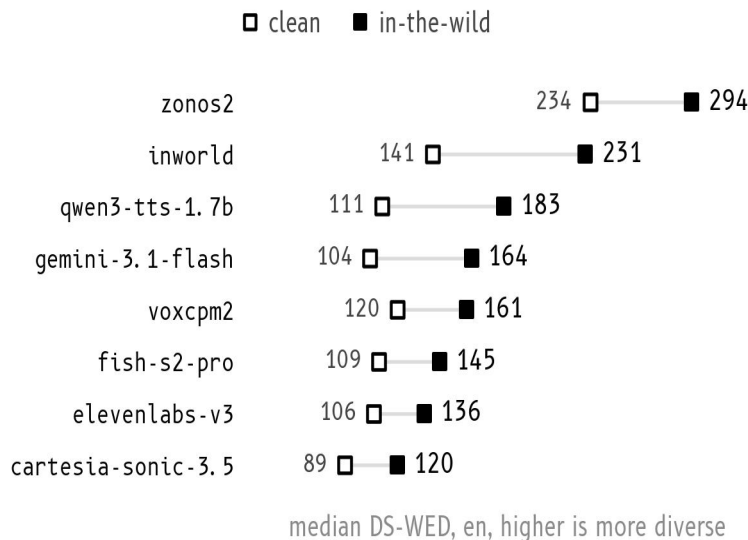
MPM

- Definition: masked-prosody-model embedding distributions
- Extractor: MPM (introduced with TTSDS)
- Scoring: Frechet distance vs real reference
- Scores cluster at 94–99, least discriminative of the four features



DS-WED

- Discretized Speech Weighted Edit Distance
- Target: prosodic variation between paired generations
- Definition: weighted edit distance over semantic tokens (SSL + k-means)
- Validated on ProsodyEval (human-annotated); real-time factor 0.110

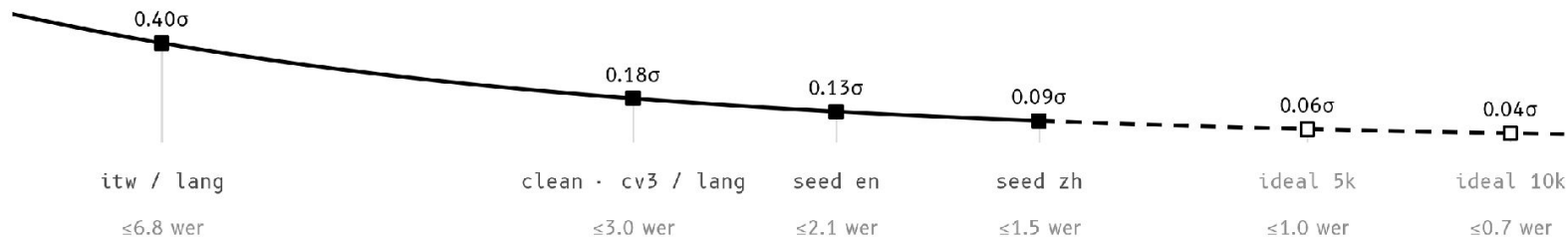


Future Work

Sample size

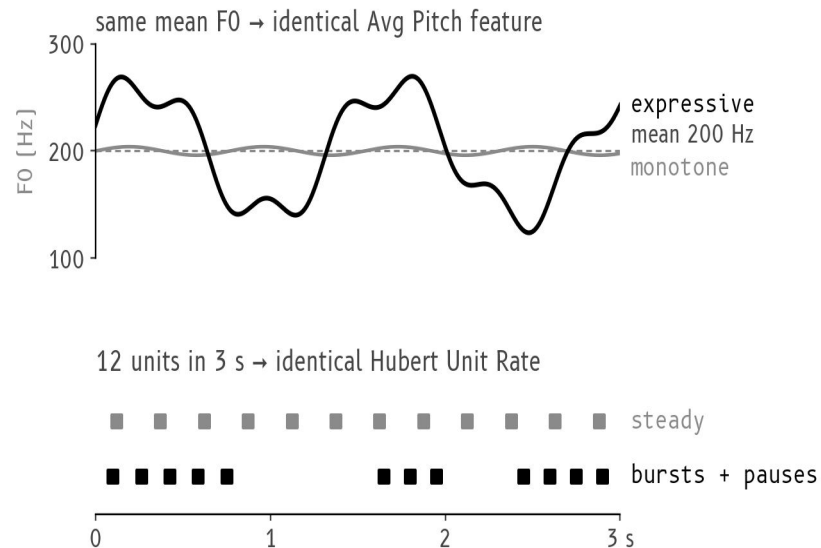
- Today: Clean 500 utt per language, hard_en 64, hard_zh 60, ITW 84–120 per language
- $MDE = 2.80 \times \sigma \times \sqrt{2/n}$ (t-test, $\alpha=.05$, $\text{power}=.80$)
- $n=500$ detects 0.18σ ; halving the detectable effect costs 4x the samples

smallest detectable difference, in σ of the per-utterance metric, t-test, $\alpha=.05$, $\text{power}=.80$, $MDE = 2.80 \cdot \sigma \cdot \sqrt{2/n} \cdot \text{wer}$: $\sigma \leq \sqrt{\mu[100-\mu]} = 17.1$ @ $\mu=3\%$



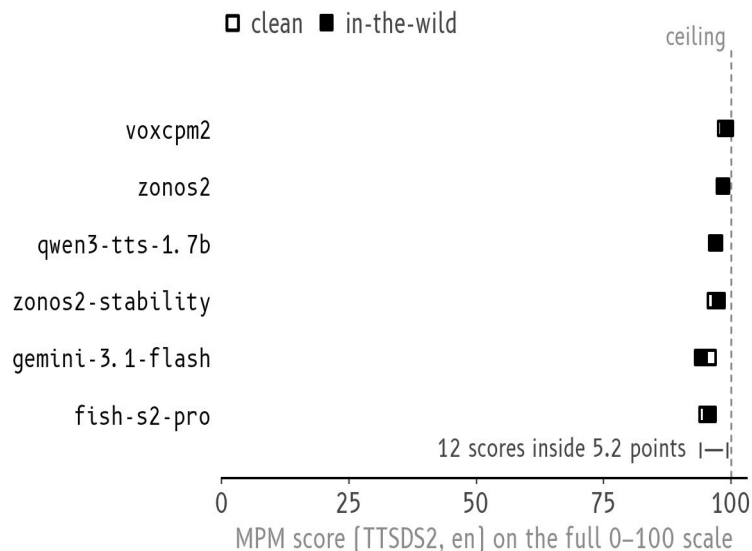
Improving upon TTSDS2 (1/2)

- Pitch Contour/Signature: utterance means hide the contour
- Rhythm/Pace: mean rates hide bursts and pauses



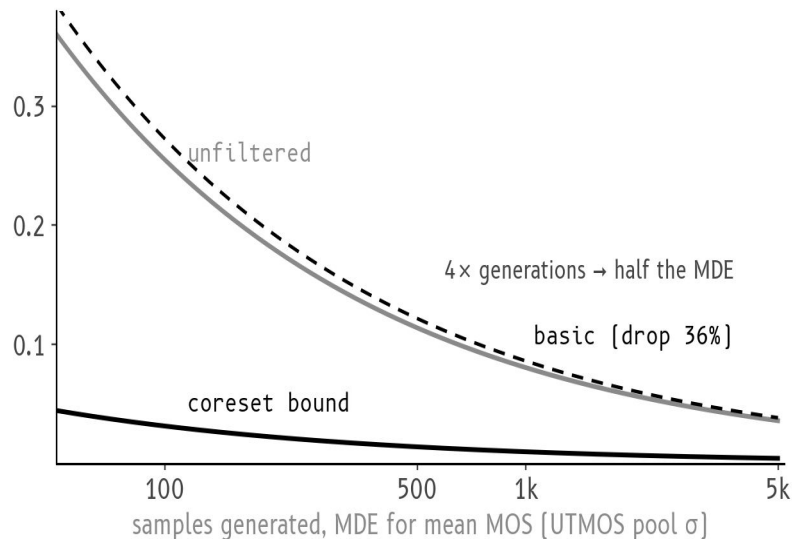
Improving upon TTSDS2 (2/2)

- MPM: today masked-prosody-model embeddings
- On the full 0–100 scale all scores sit within 5.2 points of the ceiling



More filtering

- Basic: drop 36% of samples, sigma 0.64 to 0.55, a wash for precision
- Coresets: stratified bound sigma 0.08, about 65x effective samples



Longform/Multi-turn

- Speaker Similarity Across Time
- Error Accumulation Over Time
- Contextual Appropriateness
- Cross-turn consistency

Summary

- Clean Set, In-the-wild Set, Updated Metrics, Prosody
- 17 languages, 16 h, 5 metric axes
- “Distinguishing the most expressive, robust, and multilingual models at the top.”
- github.com/Zyphra/ZTTS1-Eval

Thank you

github.com/Zyphra/ZTTS1-Eval

[arXiv:2606.24320](https://arxiv.org/abs/2606.24320)

ztts1slides.sofian.dev

mejjoute.sofian@gmail.com