

# Hearing Between the Lines: Unlocking the Reasoning Power of LLMs for Speech Evaluation



Arjun Chandra<sup>1\*</sup> Kevin Miller<sup>1\*</sup> Venkatesh Ravichandran<sup>2</sup>

Constantinos Papayiannis<sup>2</sup> Venkatesh Saligrama<sup>1</sup>

<sup>1</sup>Boston University

<sup>2</sup>Amazon AGI



EACL 2026

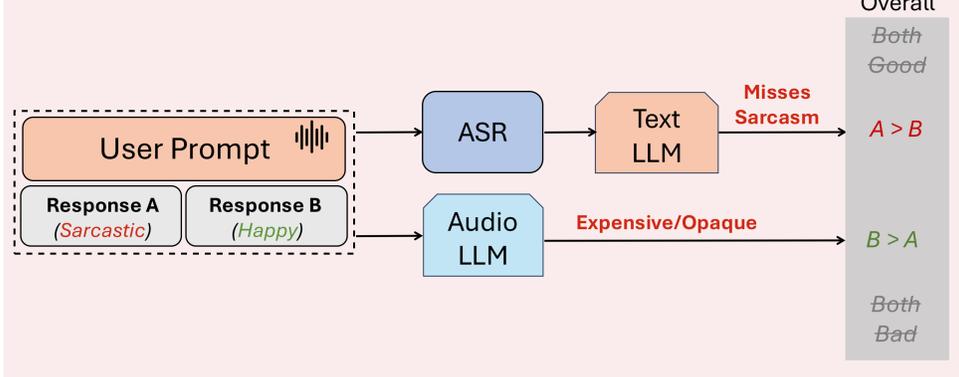
RABAT · MOROCCO

Mars · March 24-29, 2026 · مارس

## Problem Statement

Can we introduce a new paradigm for automatic evaluation of speech-to-speech models that is *cheap, interpretable and audio-aware*?

### A. Existing Speech-to-Speech Evaluation

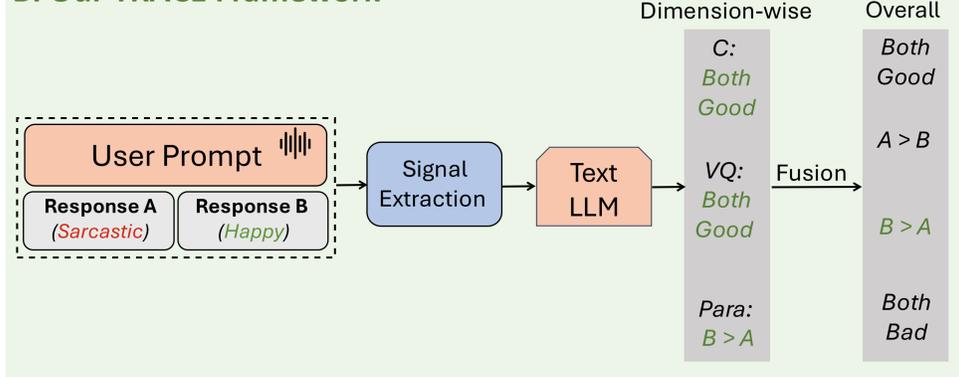


### Limitations of Existing Evaluation Paradigms.

- Labeling protocols either (1) force winners or (2) don't allow for typed-ties
- Text LLM evaluators are blind to paralinguistics (only use transcripts)
- Audio LLM evaluators are expensive, opaque, and often still miss paralinguistics

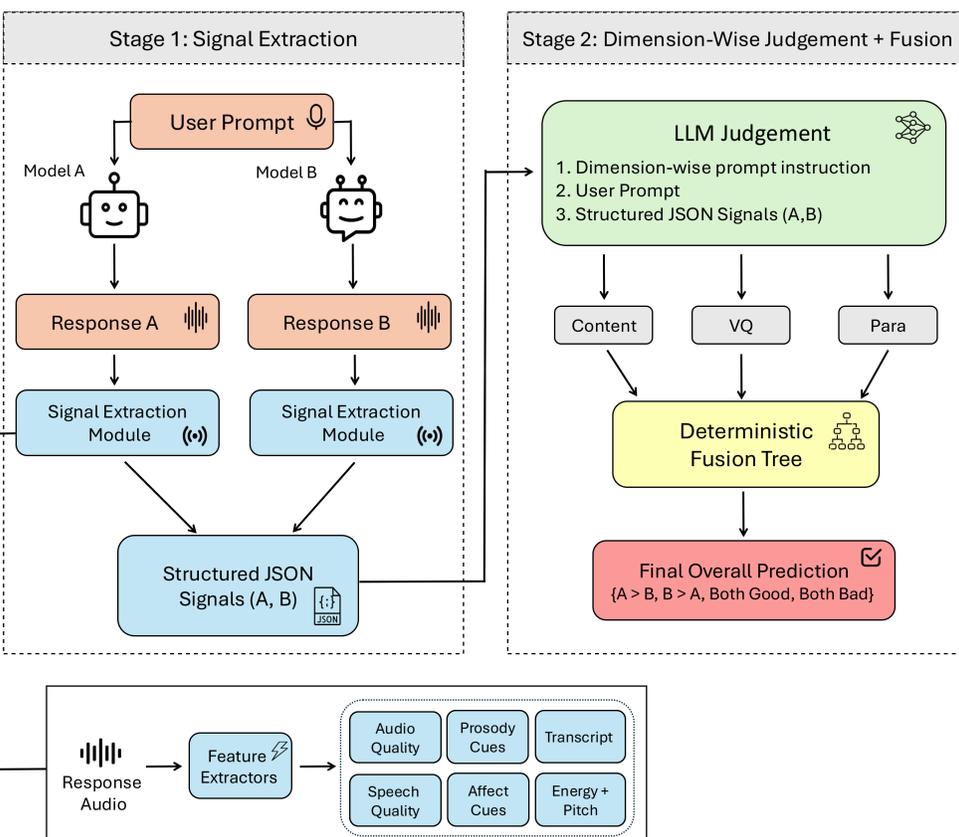
## Proposed Evaluation Framework

### B. Our TRACE Framework



### Key Innovations.

- Human Chain-of-Thought (HCoT):** A dimension-first labeling protocol with typed-ties
  - Content:** Does the response provide relevant, appropriate, and factual information?
  - Voice Quality:** Does the response sound clear and human-like?
  - Paralinguistics:** Does the response acknowledge and use the appropriate tone of voice, intonation, emotion etc. ?
- Textual Reasoning over Audio Cues for Evaluation (TRACE):** A new automatic evaluator which enables the reasoning power of text LLMs for speech evaluation



## Results - Quantitative

Dataset	Judge	Content	Voice Quality	Paralinguistics	Overall
SPEAKBENCH	Random Guess	25.0	25.0	25.0	25.0
	Audio Judge	62.5	45.6	21.4	61.1 (56.7-65.4)
	LLM Judge	60.4	39.8	29.8	62.7 (58.2-67.0)
	<b>TRACE</b>	<b>63.2</b>	<b>50.4</b>	<b>39.6</b>	<b>68.6</b> (64.3-72.7)
	Human-human agreement	76.0	60.0	82.0	60.0
S2S-ARENA	Random Guess	25.0	25.0	25.0	25.0
	Audio Judge	<b>58.9</b>	52.5	37.7	47.5 (42.4-52.7)
	LLM Judge	57.0	<b>55.1</b>	35.4	45.9 (40.4-51.3)
	<b>TRACE</b>	58.0	51.6	<b>48.1</b>	<b>57.0</b> (51.6-62.4)
	Human-human agreement	73.3	48.3	75.0	75.0

Table 1. Per-dimension and overall accuracy vs. HCoT labels (Gemini 2.5 Flash backbone).

Judge	Content	VQ	Para	Overall	Cost	Audio Judge	LLM Judge	TRACE
Audio Judge	51.4	39.1	17.2	53.4	GPU (\$)	0.00	0.25	0.42
LLM Judge	<b>58.8</b>	32.1	29.8	60.6	API (\$)	12.53	2.51	3.73
<b>TRACE</b>	58.0	<b>50.0</b>	<b>36.2</b>	<b>62.1</b>	<b>Total (\$)</b>	<b>12.532</b>	<b>2.763</b>	<b>4.158</b>

Table 2. Backbone ablation (GPT-4o).

Table 3. Cost analysis (SpeakBench w/ GPT-4o).

### Observations & Takeaways.

- TRACE achieves the **strongest agreement** with HCoT labels (Tab. 1)
- TRACE gains **generalize** across different backbone models (Tab. 2)
- TRACE is **significantly cheaper** than Audio LLM judges (Tab. 3)

## Results - Qualitative

### Example — Capturing Paralinguistic Nuance (SpeakBench)

**User Prompt.** Recite the alphabet slowly and deliberately, using a deep, authoritative voice as if you are a movie trailer narrator.

**Model 1 (Gemini 2.0 Flash).** [Deep, authoritative, human-like voice] A, B, C ... Z. Is there anything else you need help with?

**Model 2 (Cascade — GPT-4o + TTS).** [Robotic, unnatural voice] a hum in a deep dramatic voice, A, B, C ... Z. This fall, letters will never be the same.

### HCoT-Human

	C	VQ	P	Overall
label both_good	1	1	1	1

### Judge-Predictions

	C	VQ	P	Overall
Audio Judge	2	both_good	2	2
LLM Judge	2	2	2	2
<b>TRACE</b>	1	1	1	1

### Example — User Tone Handling (S2S-Arena)

**User Prompt.** [Disgusted tone of voice] Dogs are sitting by the door.

**Model 1 (GPT-4o).** [Neutral/happy tone of voice] It sounds like they're ready for a walk or eagerly awaiting someone's return. Dogs have a way of knowing when something exciting is about to happen. Are they waiting for you to take them out?

**Model 2 (Cascade — GPT-4o + TTS).** [Neutral tone of voice] Maybe they're waiting for someone to take them for a walk.

### HCoT-Human

	C	VQ	P	Overall
label both_good	1	both_bad	both_bad	both_bad

### Judge-Predictions

	C	VQ	P	Overall
Audio Judge	1	1	both_good	1
LLM Judge	1	both_good	both_good	1
<b>TRACE</b>	1	1	both_bad	both_bad

### Observations & Takeaways.

- TRACE is the only automatic evaluator to mirror **overall human preference**
- TRACE achieves strong **paralinguistic fidelity** — a key dimension where both baseline evaluators fail
- TRACE is able to reason about complex **input and response dynamics**

## Conclusion – Limitations & Future Directions

We propose a *cheap, interpretable, and audio-aware* framework for speech-to-speech evaluation.

- HCoT:** A new labeling protocol which explicitly decouples speech evaluation dimensions
- TRACE:** A new auto-rater which enables the strong textual reasoning of LLMs for speech evaluation

### Limitations.

- Experiments are limited to English datasets and cultural norms of expressivity
- TRACE relies on a manually designed set of audio features in Signal Extraction
- TRACE is subject to error propagation from upstream feature extraction models

### Future Work.

- A data-driven schema induction to adapt feature extraction step dynamically
- Enabling the LLM with tool-calling to only collect the necessary audio features

[1] Manakul, Potsawee, et al. "Audiojudge: Understanding what works in large audio model based speech evaluation." arXiv preprint arXiv:2507.12705 (2025).

[2] Jiang, Feng, et al. "S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information." arXiv preprint arXiv:2503.05085 (2025).