SPEECHCRITIC: LEARNING A DIAGNOSTIC SPEECH JUDGE FROM LIMITED HUMAN PREFERENCES

Anonymous author

ABSTRACT

Human speech conveys rich perceptual information, such as emotion and speaker identity, yet most automatic speech quality judges reduce it to a single naturalness score. We study *diagnostic* speech judges: given two candidates, a diagnostic judge decides which is better, along which perceptual dimensions (e.g., timbre, emotion, timing) they differ, and which audible cues support its decision. Learning such judges is challenging: expert annotation is costly, and simply prompting a frontier audio-language model to produce labels is unreliable: our probing reveals substantial errors and unstable instruction following. We introduce SPEECHCRITIC, which learns a diagnostic judge in a reference-conditioned cross-lingual setting from only about 300 human-labeled comparisons. Rather than replacing the frontier model, SPEECHCRITIC calibrates it with these labels: for each dimension, it selects the acoustic measurements that agree with human judgments, maps them to A/TIE/B probabilities, and passes these to the model as non-binding hints alongside the audio. Compared with the same model labeling without hints, this raises dimension-level agreement with humans by 6.3 points and cuts the mismatch with human TIE rates by 10.4 points. We then train a 7B judge on this supervision and find that different training signals shape different judge behaviors: supervised fine-tuning establishes the task, on-policy distillation transfers the teacher's dimension-level strengths and weaknesses, and reinforcement learning helps most on clear-cut comparisons where human raters agree. Notably, human listeners also find that reinforcement learning makes rationales cite more specific, localized acoustic cues, although it never directly rewards rationale text. Finally, we show that the pipeline is language-pair agnostic by instantiating it on both English–Japanese and English–Spanish. Together, these results demonstrate a path from limited human preferences to a diagnostic speech judge. Code and Demo: `https://anonymous.4open.science/w/SpeechCritic-35DD/`

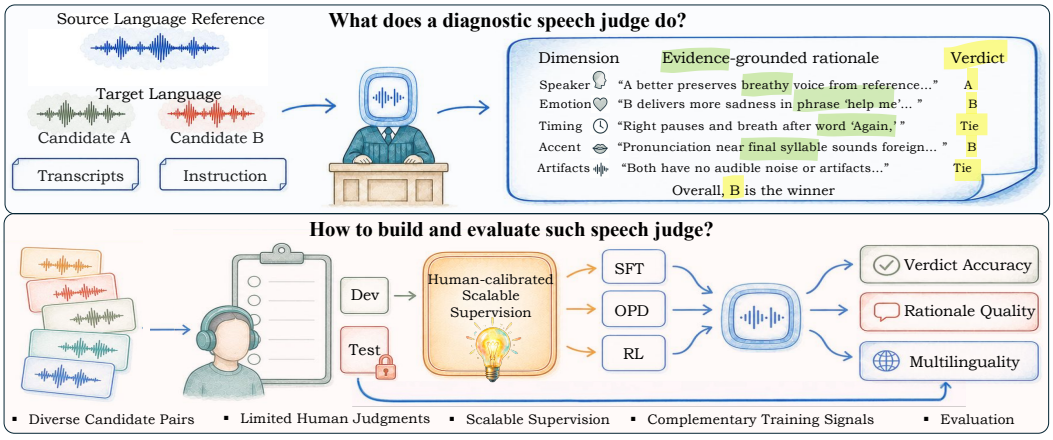


Figure 1: In a reference-conditioned cross-lingual comparison, a diagnostic speech judge produces multidimensional verdicts and evidence-grounded rationales that identify audible differences between two candidates. Our SPEECHCRITIC framework covers the end-to-end data, training, and evaluation pipeline. The central human-calibration method is introduced in Section 4.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent multimodal large language models can directly process speech, making automatic speech evaluation increasingly plausible. Yet speech perception is inherently multidimensional, and a useful evaluator should explain its preference with evidence rather than merely select a winner. We study diagnostic speech judging for pairwise speech comparison. A diagnostic judge listens to two speech candidates, whether natural or synthesized, and produces an Overall verdict, judgments across multiple perceptual dimensions, and evidence-grounded rationales supported by audibly verifiable cues. We instantiate this general task in a reference-conditioned cross-lingual setting, where the judge compares two target-language candidates against a source-language reference in terms of speaker characteristics, emotion, timing, pronunciation or accent, and audio artifacts.

Building such a judge is challenging because human listeners may reasonably disagree on subtle perceptual differences, while collecting consistent dimension-level judgments and audio-grounded explanations requires substantial expert effort. Existing approaches either train speech judges on large human preference corpora that are costly and time-consuming to collect (Zhang et al., 2025) or directly prompt frontier audio-language models as evaluators (Manakul et al., 2026). To examine the latter option, we probed three proprietary and five open-source models. Their results reveal weak human agreement, skewed candidate or TIE preferences, and inconsistent dimension-level judgments, making their outputs unreliable as training supervision. This raises our central question:

> *Can limited human judgments be used to build a scalable, specialized diagnostic speech judge without collecting human labels at training scale?*

Our key idea is to use human judgments to calibrate machine supervision rather than replace it. SPEECHCRITIC follows a select–calibrate–scale pipeline. Select: limited human judgments determine which automatic acoustic measurements provide reliable evidence for each supported dimension. Calibrate: each retained signal is mapped to uncertainty-aware probabilities over A, TIE, and B. Scale: the frozen mappings provide non-binding hints to a machine labeler, which still listens to the original speech and generates verdict-and-rationale supervision for more than 10,000 comparisons. Unlike conventional weak supervision, these measurements neither determine the labels nor replace the raw audio (Ratner et al., 2017). Unlike approaches based on human-written principles (Bai et al., 2022) or recursively generated model judgments (Wang et al., 2024), humans directly determine which acoustic evidence the machine labeler should trust. This calibration improves agreement with human dimension-level judgments by 6.3 percentage points and reduces the mismatch in TIE behavior by 10.4 points.

Building scalable supervision is only part of the problem. We further study how that supervision should be used by comparing supervised fine-tuning (SFT), on-policy distillation (OPD), and reinforcement learning (RL). SFT establishes the diagnostic task, but its gains are not monotonic in the amount of supervision, and adapting the audio path improves overall decisions without uniformly strengthening dimensional diagnosis. OPD transfers the teacher's judgment profile, with teacher behavior and conditioning mattering more than student initialization. RL sharpens the verdict behavior encoded by its reward, with gains concentrated on high-consensus comparisons and a trade-off in dimensional diagnosis. Human evaluation further shows that RL can improve acoustic grounding without direct rationale rewards. Separately, persuasive rationales can still accompany incorrect verdicts. Finally, both the SPEECHCRITIC supervision pipeline and the resulting judges transfer across Japanese and Spanish, demonstrating that the framework extends beyond a single target language.

Figure 1 summarizes our SPEECHCRITIC framework. Our contributions are:

- **Diagnostic task and human evaluation.** We formulate diagnostic speech judging with a five-dimensional rubric, analyze human subjectivity and agreement, and systematically probe zero-shot audio-language models.
- **Human-calibrated scalable supervision.** We introduce a select–calibrate–scale pipeline that turns roughly 300 human-labeled comparisons into more than 10,000 verdict-and-rationale training comparisons.
- **Learning and evaluation.** We characterize how SFT, OPD, and RL learn differently from imperfect supervision, evaluate the evidence grounding and reliability of their rationales with human listeners, and study how the framework transfers across target languages.

2 Related Work

Audio-language model judges. Recent work follows two main routes: prompting existing audio-language models and training dedicated evaluators. AudioJudge directly prompts off-the-shelf audio models (Manakul et al., 2026), while TRACE converts acoustic cues into textual descriptions for reasoning by a text LLM (Chandra et al., 2026). Dedicated evaluators and supporting resources cover instruction-driven assessment, low-level speech quality, multilingual judging, spoken dialogue, and multi-task evaluation (Zhang et al., 2026; Wang et al., 2025; 2026a; Ji et al., 2025; Wang et al., 2026b). SpeechJudge and GSRM further train specialized judges from large-scale human preferences or expert ratings (Zhang et al., 2025; Shen et al., 2026). Much of this literature emphasizes naturalness or low-level acoustic quality. Our task instead jointly evaluates richer perceptual relationships between speech candidates, using limited human judgments to calibrate scalable machine supervision.

Scalable evaluator training. The broader LLM literature has explored several ways to reduce large-scale human preference annotation. Prometheus and JudgeLM distill model-generated feedback into trainable text evaluators (Kim et al., 2024; Zhu et al., 2023), while Auto-J and Critique-out-Loud generate explicit critiques before final judgments or rewards (Li et al., 2024; Ankner et al., 2024). RLAIF replaces human preferences with AI feedback (Lee et al., 2023); Constitutional AI guides that feedback with human-written principles (Bai et al., 2022); and Self-Taught Evaluators bootstrap evaluator reasoning from synthetic data (Wang et al., 2024). These studies show that evaluator supervision can be generated or distilled. In speech modality, however, the supervising model must perceive the acoustic signal and ground both its verdicts and rationales in audible evidence, making the reliability of machine-generated supervision itself a central concern.

Weak supervision and domain evidence. Speech evaluation has long relied on specialized measurements, including speaker-embedding similarity (Wang et al., 2023), emotion representations (Wagner et al., 2023; Ma et al., 2024), temporal and pronunciation measures, and neural quality predictors (Saeki et al., 2022; Reddy et al., 2021; Mittag et al., 2021). These tools estimate individual or narrowly defined perceptual properties, but do not jointly produce comparative verdicts and audio-grounded rationales. Our use of such measurements is related to programmatic weak supervision, where systems such as Snorkel combine noisy heuristic labeling functions into probabilistic training labels (Ratner et al., 2017). Our measurements, however, do not directly label comparisons or replace the raw speech. Although TRACE and GSRM also expose structured acoustic evidence to evaluators, we screen each candidate measurement against held-out human judgments for the corresponding dimension, calibrate retained signals into probabilistic A/Tie/B hints, and provide them only as non-binding guidance to a raw-audio machine labeler.

3 Diagnostic Speech Judging: Task and Human Evaluation

Task formulation. We study *reference-conditioned cross-lingual comparison*. Each comparison contains a source-language reference utterance x^{src} and two target-language speech candidates x^A and x^B. Our primary experiments use an English reference and Japanese candidates. The judge also receives source and target transcripts with the matched semantic meaning (t^{src}, t^{tgt}) and an instruction specifying the rubric r. We denote the full input by

$$x = \left(x^{\mathrm{src}}, x^A, x^B, t^{\mathrm{src}}, t^{\mathrm{tgt}}, r\right),$$

The judge outputs a structured natural language response following the instruction, which contains a verdict v_k for each of five perceptual dimensions, and a binary Overall verdict v_{overall}. The output also contains rationales that support any verdict it makes, identifying localized and audibly verifiable observations, such as a word, phrase, pause, or change in prosodic delivery. Such rationales can help move automated evaluation closer to the actionable feedback of an expert human listener.

$$v_k \in \{A, B, \mathrm{Tie}\}, \quad k = 1, \ldots, 5 \qquad\qquad v_{\mathrm{overall}} \in \{A, B\}.$$

The five dimensions are Speaker, Emotion, Timing, Pronunciation/Accent, and Audio Artifacts. This rubric is empirically sufficient for the task: a simple equal-weight aggregation of human dimension-level votes recovers human Overall preferences with 93.0% out-of-fold accuracy after a one-feature logistic calibration on the development set.

Constructing diverse data. We constructed a heterogeneous speech pool using multiple off-the-shelf speech synthesis systems. Their distinct behaviors and failure modes create natural variation across perceptual dimensions. For example, cross-lingual synthesis produces varying degrees of accent leakage, yielding candidate pairs that differ in Pronunciation/Accent quality. Appendix B describes the pool in detail. We constructed more than 10,000 comparisons in total.

Human preference benchmark. We recruited 20 native Japanese listeners who also understand English. Human listeners are given input x and asked to produce verdicts v_k and v_{overall}, without the need to write down any explanations. Each comparison receives three human judgments. In subsequent analyses, comparisons with unanimous v_{overall} votes (3–0) are marked as *high-consensus comparisons* and those with split votes (2–1) as *low-consensus comparisons*.

Human annotation yields 315 development comparisons and 290 test comparisons. Only the development set is used to calibrate the following supervision pipeline; the test set is never used for calibration, training, or model selection. Throughout the paper, we evaluate the binary Overall verdict using accuracy over $\{A, B\}$ and the five dimension-level verdicts using macro-F1 over $\{A, B, \text{Tie}\}$, averaged across dimensions. To contextualize the specific task's subjectivity, we additionally collect all 20 listeners' preference for 30 anchor comparisons. Appendix A provides detailed annotation instructions, dimension definitions, and human agreement analysis.

On the full English–Japanese test set, the expected agreement of a randomly sampled panel listener with the panel-majority label is 85.8% for v_{overall} and 78.8% for averaged dimensional macro-F1. We treat these values as label-reliability references rather than model-performance upper bounds.

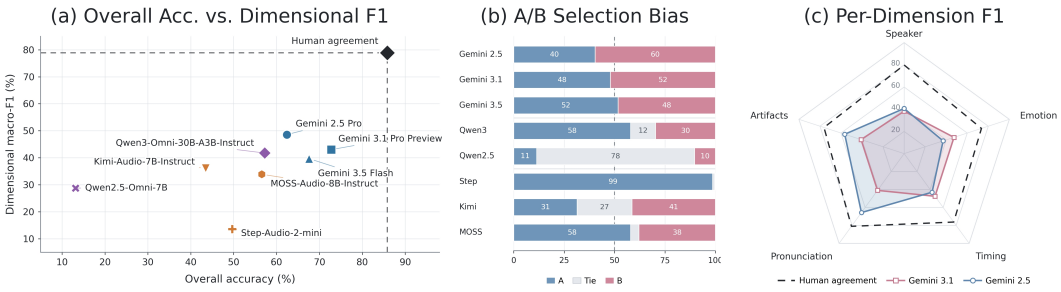


Figure 2: **Zero-shot probing** of three proprietary and five open-source audio-language models on our constructed English–Japanese test set (conducted only after all model choices were fixed.) The results reveal a clear gap to reliable human-aligned diagnostic judging.

Zero-shot probing. We evaluated proprietary and open-source audio-language models zero-shot on the task. All models received the same instruction. As shown in Figure 2, three patterns emerge. First, both Overall accuracy and dimensional macro-F1 remain below human agreement. Second, several models exhibit strongly skewed A/B output distributions, while Qwen2.5-Omni-7B (Xu et al., 2025a) predicts the invalid Overall TIE on 78.3% of comparisons, violating our task formulation. Third, no model performs consistently well across all five perceptual dimensions. Off-the-shelf models are therefore not reliable zero-shot diagnostic judges, motivating our human-calibrated supervision and specialized judge training.

4 METHOD: SCALING SUPERVISION FROM LIMITED HUMAN JUDGMENTS

Gemini-3.1-Pro (Google, 2026) achieves the highest zero-shot Overall accuracy in Figure 2, but its raw supervision remains systematically misaligned with human judgments. Humans choose TIE for 55.1% of dimension-level judgments, compared with only 19.8% for Gemini, whose rationales can also contain unsupported or post-hoc acoustic claims. We therefore use limited human judgments to calibrate the evidence supplied to the machine labeler through three steps: select, calibrate, and scale (Figure 3).

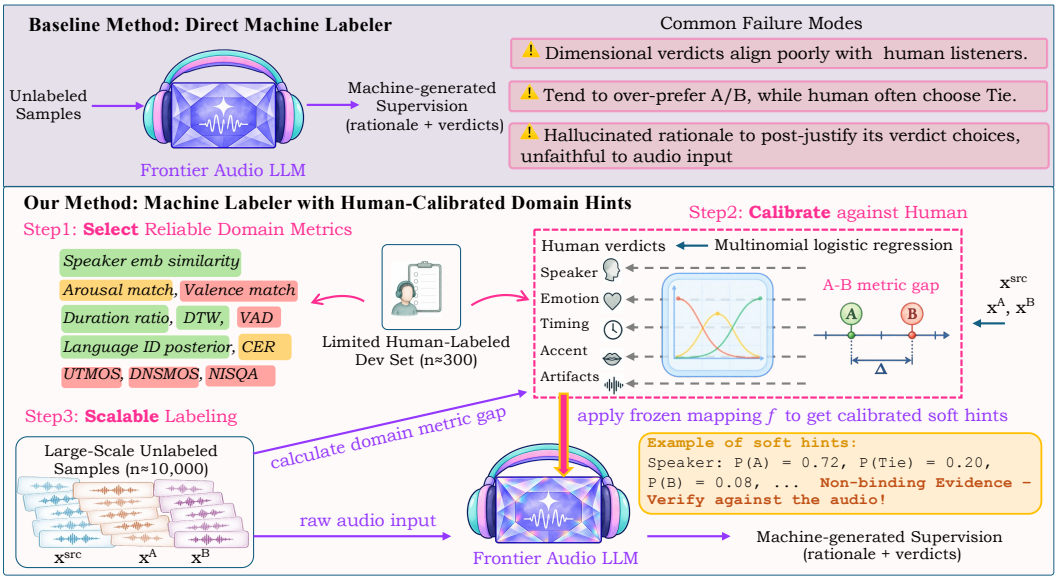


Figure 3: **Human-calibrated domain hints.** Direct machine labeling exhibits systematic disagreement with human. Our select–calibrate–scale pipeline uses limited human judgments to convert reliable domain metrics into probabilistic hints for scalable and acoustically grounded supervision.

4.1 Selecting Domain Metrics

We first use the human-labeled development set to determine which domain metrics provide reliable evidence for each diagnostic dimension instead of assuming they are all reliable. Let $m_{kj} \in \mathcal{M}_k$ denote candidate measurement j for dimension k. For development comparison i, we orient each measurement so that a larger value favors that candidate and compute its signed metric gap:

$$\Delta_{ik}^{(j)} = m_{kj}(A_i) - m_{kj}(B_i), \qquad z_{ik} \in \{A, \text{Tie}, B\},$$

where z_{ik} is the human-majority dimension-level verdict. We evaluate each candidate through five-fold grouped out-of-fold prediction, keeping comparisons from the same source group in one fold. Within each fold, standardization and a temporary probability mapping are fitted on the other four folds, and their held-out predictions are compared with z_{ik} using macro-F1. The class-frequency baseline uses the A/Tie/B frequencies estimated from the same four training folds. These fold-specific mappings are used only to select m_k and are then discarded. Table 1 summarizes the retained measurements and their out-of-fold performance. No candidate outperforms the baseline for Audio Artifacts, so it receives no domain hint. Appendix B.2 reports the complete screening results.

Table 1: Retained domain metrics and grouped OOF macro-F1 on the human-labeled dev set.

Dimension	**Speaker**	**Emotion**	**Timing**	**Pronunciation**	**Artifacts**
Retained metric(s)	Speaker similarity	Arousal mismatch	Duration deviation + envelope DTW	Character error rate or Language ID*	None
OOF macro-F1 (%)	20.9 → 29.1	19.2 → 36.0	18.6 → 54.7	26.7 → 50.6	No improv.

*OOF validation found the VoxLingua target-language posterior predictive only for English–Spanish, so it is used only for that language.

4.2 Calibrating the Selected Metrics

After selecting m_k, we refit one final mapping f_k on all development comparisons. Let $j_k^\star$ denote the retained candidate, and write $m_k = m_{kj_k^\star}$ and $\Delta_{ik} = \Delta_{ik}^{(j_k^\star)}$. We fit an ℓ_2-regularized multinomial logistic mapping

$$\mathbf{p}_{ik} = f_k(\Delta_{ik}) = \text{softmax}\left(\boldsymbol{\alpha}_k + \boldsymbol{\beta}_k \frac{\Delta_{ik} - \mu_k}{\sigma_k}\right) = \big(\Pr(A), \Pr(\text{Tie}), \Pr(B)\big).$$

Table 2: Human-calibrated domain hints improve the dimensional alignment and TIE-rate calibration of Gemini 3.1 Pro as a machine labeler, with no detectable change in Overall accuracy.

	Dev			Test		
Strategy	Overall Acc. % ↑	Dim. Macro-F1 % ↑	TIE MAE* ↓	Overall Acc. % ↑	Dim. Macro-F1 % ↑	TIE MAE ↓
Direct labeling	72.6	44.1	29.0	**71.3**	42.3	35.2
+ Calibrated hints	**75.8**	**50.9**	**22.1**	69.9	**48.5**	**24.8**

*The mean absolute difference between model and human TIE rates across dimensions; machine labelers tend to force an A/B choice more often than humans.

The mapping is learned from human dimension-level verdicts using cross-entropy. We then freeze its standardization parameters (μ_k, σ_k) and regression parameters $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$. The resulting distribution represents uncertainty in the calibrated human-choice mapping rather than forcing a hard label.

4.3 GENERATING SCALABLE DIAGNOSTIC SUPERVISION

For each unlabeled comparison x, we compute the selected signed metric gap and apply the frozen calibration mapping f_k:

$$\widehat{\mathbf{p}}_k(x) = f_k(m_k(A_x) - m_k(B_x)) = \big(\widehat{\Pr}(A), \widehat{\Pr}(\text{Tie}), \widehat{\Pr}(B)\big).$$

The machine labeler receives calibrated distributions for the four supported dimensions alongside the raw speech, transcripts, and rubric; Audio Artifacts is marked unknown, and no Overall hint is provided. As illustrated in Figure 3, these hints are non-binding: the model verifies them against the audio and may override them before generating five dimension-level verdicts with rationales and an Overall verdict.

Using roughly 300 human-labeled development comparisons, this pipeline guides machine-generated verdict-and-rationale supervision for more than 10,000 unlabeled comparisons. Table 2 evaluates this supervision before any student model is trained. On the held-out test set, paired reference-cluster bootstrap intervals (20,000 resamples) show that calibrated hints improve dimensional macro-F1 by 6.3 points (95% CI: $[3.0, 9.4]$) and reduce TIE-rate MAE by 10.4 points (95% CI: $[7.4, 13.4]$ reduction); both intervals exclude zero. Overall accuracy changes by -1.5 points, with an interval crossing zero (95% CI: $[-7.0, 4.0]$), indicating no detectable change. Although TIE still remains under-predicted compared to human preference (Appendix A.4, Table 7), the hints make the same Gemini model a better source of machine-generated diagnostic supervision.

5 TRAINING SIGNALS FOR DIAGNOSTIC SPEECH JUDGES

Our scalable supervision specifies both verdicts and rationales, but remains imperfect because it inherits errors from the machine labeler. We therefore study not only how to generate supervision, but how a student should learn from it. All student judges are initialized from Qwen2.5-Omni-7B; implementation details are provided in Appendix C. Unless otherwise stated, results are reported as the mean and standard deviation over three independent training runs.

SFT establishes the diagnostic task. Zero-shot Qwen2.5-Omni-7B does not reliably follow the diagnostic rubric (Figure 2): its responses often omit required verdicts or violate the binary Overall decision, yielding only 13.10% Overall accuracy. We therefore begin with SFT on fixed verdict-and-rationale targets, which provides token-level supervision over the complete structured response. As shown in Panel A of Table 3, with LoRA applied only to the language model, SFT raises Overall accuracy to 69.54% and establishes the basic rubric following, dimensional distinctions, and output structure required by subsequent training. SFT therefore serves as the task-learning stage, although it necessarily remains exposed to errors in the fixed machine-generated targets.

OPD transfers the teacher's conditional judgment profile. Fixed-target SFT never evaluates the student on prefixes generated by its own policy. OPD instead samples a student trajectory and minimizes token-level Jensen–Shannon divergence from a frozen teacher along it Agarwal et al. (2024). Let p_θ and q_ϕ denote the student and teacher next-token distributions. Vanilla OPD gives both models x; privileged OPD additionally gives the teacher an evidence brief c derived from

Table 3: **Main results:** SFT, OPD, and RL produce distinct training signals to shape a diagnostic speech judge, rather than a uniform accuracy ladder.

Configuration	Overall Acc. (%)	High-Consensus Acc. (%)	Low-Consensus Acc. (%)	Dim. Macro-F1 (%)
Baseline				
Qwen2.5-Omni-7B (zero-shot)	13.10	13.53	12.50	28.76
A. SFT				
SFT	69.54 ±0.80	81.76 ±1.02	52.22 ±3.37	52.37 ±0.10
B. OPD: teacher choice				
Privileged: SFT teacher	71.72 ±1.58	85.10 ±1.89	52.78 ±3.85	52.15 ±1.26
Privileged: Qwen3-30B teacher	73.79 ±0.91	**87.25** ±0.68	54.72 ±2.93	44.61 ±2.79
Vanilla: Qwen3-30B teacher	56.78 ±1.44	58.04 ±3.02	55.00 ±0.83	37.51 ±0.93
C. RL: reward and algorithm (initialized from SFT)				
DAPO — Overall only	69.43 ±0.87	83.14 ±1.89	50.00 ±0.83	52.79 ±0.83
DAPO — Dimensional only	71.15 ±1.05	84.71 ±1.56	51.94 ±0.48	45.72 ±1.39
DAPO — Overall + dimensional	70.57 ±0.20	84.12 ±1.56	51.39 ±1.92	50.19 ±2.28
DAPO — Gated dimensional reward	71.84 ±1.05	85.69 ±1.48	52.22 ±0.48	47.64 ±1.01
GRPO — Same gated reward	70.34 ±0.91	83.73 ±0.34	51.39 ±1.73	49.36 ±1.64
D. Composed training				
OPD initialization + gated RL	**73.91** ±0.40	85.88 ±0.59	**56.94** ±1.27	**52.87** ±1.74

scalable supervision, while the student receives only x (Zhao et al., 2026):

$$\mathcal{L}_{\mathrm{OPD}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{JSD}(p_\theta(\cdot \mid x, y_{<t}) \,\|\, q_\phi(\cdot \mid x, c, y_{<t}))\,.$$

Panel B of Table 3 shows that the teacher's input and behavior, rather than model capacity alone, determine what OPD transfers. Vanilla Qwen3 (Xu et al., 2025b) supervision produces only 56.78% Overall accuracy, whereas privileged conditioning raises it to 73.79%. Yet the resulting dimensional macro-F1 is 44.61%, well below the 52.15% obtained from the SFT teacher, showing that the student inherits the teacher's judgment profile rather than a universally better judge. Holding the privileged Qwen3 teacher fixed, changing the student initialization among base, SFT, and RL checkpoints shifts Overall accuracy only within 72.07–73.79% and yields no consistent improvement. Teacher choice and conditioning therefore dominate initialization.

RL improves the verdict behavior encoded by its reward. SFT and OPD both supervise the full token sequence, potentially transferring errors in the target or teacher rationale. Instead, we employ RL to evaluate only the parsed verdicts of each generated response, allowing the policy to improve decision behavior without treating any target rationale as ground truth. Our primary reward grants dimensional credit only after the Overall verdict is correct:

$$R = r_{\mathrm{overall}} + \lambda \mathbf{1}\{r_{\mathrm{overall}} = +1\} r_{\mathrm{dim}}.$$

Here $r_{\mathrm{overall}} \in \{-1, +1\}$, while r_{dim} averages correctness over dimensions with decisive A/B targets. TIE-labeled dimensions and rationale text receive no direct reward.

Panel C of Table 3 compares Overall-only, dimension-only, additive, and gated rewards under DAPO, together with GRPO using the same gated reward. Overall-only training provides no gain, while the gated reward performs best, raising Overall accuracy from 69.54% to 71.84%. Nearly all of this improvement occurs on high-consensus cases (+3.9 points), with no gain on low-consensus cases. At the same time, dimensional macro-F1 falls from 52.37% to 47.64%, and decreases further under the dimension-only reward. Because TIE targets receive no dimensional reward, these predictions can drift. RL therefore sharpens the decision behavior explicitly encoded by the reward rather than uniformly improving diagnostic quality.

These results show that the three signals do not form a uniform accuracy ladder: SFT establishes the task, OPD transfers the teacher's judgment profile, and RL sharpens rewarded verdict behavior. Additional prompt and supervision-format ablations, student-initialization controls, dual-teacher distillation, and OPD–RL interactions are consolidated in Table 9 of Appendix D.

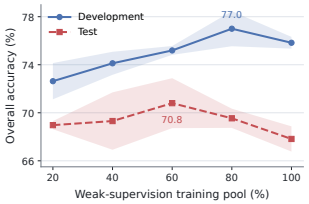


Figure 4: More supervision is not always helpful for SFT.

Table 4: **Trainable modules for SFT.** #M denotes trainable parameters in millions.

Trainable modules	#M	Overall Acc. (%)	Dim. Macro-F1 (%)
Audio projector only	1	14.48 ±1.38	29.92 ±0.40
LLM LoRA only	323	69.54 ±0.80	**52.37** ±0.10
LLM LoRA + audio projector	324	70.69 ±2.15	52.03 ±2.06
LLM LoRA + audio encoder	417	71.95 ±1.90	50.79 ±1.68
LLM LoRA + audio encoder + projector	418	**72.41** ±1.50	52.07 ±0.84

6 Understanding Diagnostic Judge Behavior

6.1 RQ1: How Should Scalable Supervision Be Used?

How much? More supervision is not always better. Human-calibrated domain hints improve the machine labeler, but its development-set labels still differ from human judgments and remain imperfect training targets (Table 2). This makes the optimal amount of supervision non-obvious: more comparisons increase coverage but may also introduce less reliable targets. Figure 4 varies the amount of machine-generated supervision used for SFT while holding the training recipe and epochs fixed, so larger fractions receive proportionally more updates. Performance initially improves, then declines when the full pool is used, consistent with fixed-target SFT imitating noise or inconsistencies in the additional comparisons. We therefore use the 80% regime (about 8,000 comparisons) for subsequent English–Japanese SFT comparisons and downstream experiments initialized from SFT.

Where? Audio-path adaptation improves overall decisions, not dimensional diagnosis. Scalable supervision can be applied only to the language model, where it primarily teaches the rubric and response structure, or also to the audio projector and encoder, where it can adapt the speech representations used for judging. Holding the SFT data and training recipe fixed, Table 4 shows that the audio projector alone is insufficient, whereas extending LLM adaptation to the audio path raises Overall accuracy from 69.54% to 72.41%, with most of the gain coming from the encoder. Dimensional macro-F1 nevertheless remains unchanged, motivating audio-path objectives designed specifically for fine-grained diagnostic perception.

6.2 RQ2: Does Better Judgment Imply Better Diagnostic Reasoning?

Verdict rewards can improve evidence-grounded rationales. A diagnostic judge must do more than predict verdicts: it should explain them with audibly verifiable evidence localized to specific words or phrases. We ask human listeners to compare paired SFT and SFT+RL rationales after hearing the audios. To isolate rationale quality, we include only cases in which both models produce the same Overall verdict. On audible-evidence grounding, listeners prefer SFT+RL in 11 of 28 judgments, prefer SFT in one, and report no meaningful difference in 16 (two-sided exact sign test, $p = 0.006$). This suggests that verdict rewards can improve the specificity and accuracy of cited acoustic evidence without directly scoring rationale text. Figure 5 shows one example; Appendix E provides the full protocol and criteria. This emergent effect may arise because sequence-level updates alter the rationale and verdict jointly, as also explored in text-only LLMs (DeepSeek-AI, 2025).

SFT

- B successfully captures the weary resignation of the reference, particularly in the delivery of 'ほんとうまくいかない'
- A's pacing is severely disrupted by unnatural pauses and inserted sounds, breaking the natural flow of the dialogue
- A contains severe synthesis artifacts, including hallucinated speech, inserted humming, and a loud, unnatural laugh at the end of the clip

SFT + RL 👍 Preferred by human

- B successfully captures the weary, resigned frustration of the reference, particularly in 'だめだな' and the final phrase 'もう遅いかも'
- A's pacing is ruined by unnatural, drawn-out vocalizations inserted between the phrases, completely breaking the natural rhythm
- A suffers from severe synthesis artifacts, including loud, glitchy groans, unnatural syllables, and robotic distortion throughout the entire clip

Figure 5: **RL can improve rationale grounding without direct rationale rewards.** Given identical verdicts, SFT+RL cites more specific and better-localized acoustic evidence than SFT.

Table 5: **Cross-lingual transfer of diagnostic speech judges.** All models adapt only the LLM with LoRA. Monolingual judges transfer strongly across target languages, as highlighted in red.

Training	SFT data (comparisons)	JA Overall Acc. (%)	JA Dim. Macro-F1 (%)	ES Overall Acc. (%)	ES Dim. Macro-F1 (%)
JA only	7,840	69.54 ±0.80	52.37 ±0.10	65.91 ±2.17	48.23 ±2.03
ES only	8,000	70.92 ±0.80	50.52 ±1.17	66.02 ±1.32	50.00 ±0.89
JA+ES combined	15,840	71.38 ±1.72	50.10 ±1.30	66.67 ±1.35	49.75 ±0.56

Persuasive Rationales Can Still Support Wrong Verdicts. In the main OPD comparison (Panel B of Table 3), the Qwen3-teacher model achieves the highest Overall accuracy, but its teacher is used zero-shot and has never learned our diagnostic rubric through task-specific training. Its rationale style is also quite different: its responses are substantially longer and more detailed than those of the other trained judges (about 488 tokens on average, compared with 322 for SFT-teacher OPD). Human listeners also prefer these rationales more often than those from SFT-teacher OPD, often citing their apparent evidence grounding. Yet this advantage appears primarily when at least one dimension-level verdict is wrong or both models choose the wrong overall winner, a pattern consistent with a possible verbosity or persuasiveness bias rather than greater diagnostic reliability. Detailed diagnostic language can therefore make an erroneous judgment appear credible: rationale fluency or persuasiveness is not diagnostic reliability. This reminds us that rationale quality must be evaluated against the audio and for consistency with the verdict, not from textual plausibility alone.

6.3 RQ3: Does SpeechCritic Scale Across Languages?

We evaluate cross-lingual scalability at two levels: whether the full SpeechCritic supervision pipeline can be instantiated for a new target language, and whether judges trained on one or both languages transfer across them. Starting from our primary English–Japanese setting, we repeat the supervision pipeline in Section 4 for English–Spanish comparisons and train monolingual and joint judges. On the Spanish test set, zero-shot Qwen2.5-Omni-7B and Gemini-3.1-Pro reach 8.14% and 62.87% Overall accuracy, respectively, providing reference points for task-specific specialization.

Table 5 shows that each monolingual judge transfers strongly to the other language, while joint training incurs no clear loss relative to either monolingual model. Thus, both the supervision framework and the resulting diagnostic behavior scale across the two target languages studied, and the latter can be consolidated into a single judge. Appendix D.1 further reports results on the public VOX-DUB benchmark (Toloka team, 2025). Because its dimensions and protocol differ from ours, we treat it as an external reference rather than a directly comparable benchmark.

7 Conclusion

We introduced SpeechCritic, a framework for learning diagnostic, acoustically grounded speech judges from limited human preferences. Its *select–calibrate–scale* pipeline identifies reliable acoustic evidence, maps it to uncertainty-aware human-choice probabilities, and uses the frozen mappings to guide a machine labeler that also listens to the original audio. This approach scales roughly 300 human-labeled comparisons into more than 10,000 verdict-and-rationale training comparisons, improving agreement with human dimension-level judgments and better reflecting when listeners consider two candidates tied.

Our results show that constructing scalable supervision and learning from it jointly shape a judge's decisions and explanations. SFT establishes task competence, although more supervision is not always better and audio-path adaptation does not improve every diagnostic criterion. OPD transfers the teacher's judgment profile, while RL sharpens the verdict behavior encoded by its reward. Human evaluation shows that RL can also improve rationales' grounding in specific, localized acoustic cues, even without directly rewarding rationale text. Yet persuasive rationales can still accompany incorrect verdicts, underscoring the need to evaluate verdict correctness and acoustic grounding separately, rather than treating fluent explanations as evidence of reliable judgment. Both the supervision pipeline and the resulting judges also transfer across Japanese and Spanish. Together, these findings demonstrate a path from limited human preferences to diagnostic speech judges.

AI USE STATEMENT

We used generative AI tools to synthesize dataset, refine manuscript writing, create partial graphical elements in the figures, prepare the interactive demonstration page, generate code for bootstrap-based statistical analysis, assist with code review, and run the auxiliary text-only rationale check reported in the appendix. The model and training implementation was adapted from established open-source frameworks, not generated by AI. The auxiliary check was used only to assess the limitations of automatic rationale evaluation, not as evidence of acoustic grounding. We did not use generative AI to formulate the research ideas, design the methodology or experiments, conduct qualitative analysis, interpret results, execute experiments, or record their outcomes. The authors reviewed and verified all AI-assisted content and take responsibility for the final manuscript.

ETHICS STATEMENT

Both human studies—preference annotation and rationale evaluation—were conducted through a professional third-party organization with trained, compensated native-speaker participants. The study materials excluded violent or otherwise harmful content and audio that could pose a hearing risk. This work does not involve the release of a dataset.

REPRODUCIBILITY STATEMENT

Training objectives, implementation details, and hyperparameters are provided in Appendix C.1. We also released the training and data preparation code.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Arjun Chandra, Kevin Miller, Venkatesh Ravichandran, Constantinos Papayiannis, and Venkatesh Saligrama. Hearing between the lines: Unlocking the reasoning power of llms for speech evaluation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2895–2916, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

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

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Google. Gemini 3.1 Pro Preview. Gemini API model documentation, 2026. URL `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`.

Google DeepMind. Gemini 3.5 Flash model card. Model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-5-flash/`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, et al. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*, 2025.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, et al. Generative judge for evaluating alignment. In *International Conference on Learning Representations*, volume 2024, pp. 27547–27574, 2024.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024. doi: 10.18653/v1/2024.findings-acl.931.

Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Barr Held, and Diyi Yang. AudioJudge: Understanding what works in large audio model based speech evaluation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3644–3663, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.168. URL https://aclanthology.org/2026.eacl-long.168/.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pp. 2127–2131. ISCA, August 2021. doi: 10.21437/interspeech.2021-299. URL http://dx.doi.org/10.21437/Interspeech.2021-299.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, November 2017. ISSN 2150-8097. doi: 10.14778/3157794.3157797. URL http://dx.doi.org/10.14778/3157794.3157797.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 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)*, pp. 6493–6497, 2021. doi: 10.1109/ICASSP39728.2021.9414878.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: Utokyo-sarulab system for voicemos challenge 2022, 2022. URL https://arxiv.org/abs/2204.02152.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Maohao Shen, Tejas Jayashankar, Osama Hanna, Naoyuki Kanda, Yancheng Wang, Kateřina Žmolíková, Ruiming Xie, Niko Moritz, Anfeng Xu, Yashesh Gaur, Gregory Wornell, Qing He, and Jilong Wu. GSRM: Generative speech reward model for speech rlhf, 2026. URL https://arxiv.org/abs/2602.13891.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report, 2026. URL https://arxiv.org/abs/2601.21337.

Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

Toloka team. VOX-DUB: a new benchmark that puts AI dubbing to the test. https://toloka.ai/blog/ai-dubbing-benchmark/, September 9 2025.

Jörgen Valk and Tanel Alumäe. VOXLINGUA107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658, 2021. doi: 10.1109/SLT48900.2021.9383459.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023. doi: 10.1109/TPAMI.2023.3263585.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096626.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. SpeechLLM-as-judges: Towards general and interpretable speech quality evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7675–7700, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.349. URL https://aclanthology.org/2026.acl-long.349/.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23588–23609, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1150. URL https://aclanthology.org/2025.acl-long.1150/.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators, 2024. URL https://arxiv.org/abs/2408.02666.

Yuanyuan Wang, Dongchao Yang, Yayue Deng, Zhiyong Wu, Steven Y. Guo, Helen M. Meng, and Xixin Wu. UniSRM: A unified speech reward model for reasoning-based fine-grained assessment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46346–46366, San Diego, California, United States, July 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2150. URL https://aclanthology.org/2026.acl-long.2150/.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report, 2025. URL https://arxiv.org/abs/2507.16632.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Donghua Yu, Jun Zhan, Kang Yu, Kexin Huang, Liwei Fan, Mingshu Chen, Qinyuan Cheng, Ruixiao Li, Shimin Li, Songlin Wang, Xingjian Zhao, Yang Gao, Yitian Gong, Yiyang Zhang, Zhe Xu, and Xipeng Qiu. Moss-audio technical report, 2026. URL https://arxiv.org/abs/2606.01802.

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

Leying Zhang, Bowen Shi, Haibin Wu, Bach Viet Do, and Yanmin Qian. JASTIN: Aligning llms for zero-shot audio and speech evaluation via natural language instructions, 2026. URL https://arxiv.org/abs/2605.04505.

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT:a scalable lightweight infrastructure for fine-tuning, 2025. URL https://arxiv.org/abs/2408.05517.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. 2023.

Table 6: Human benchmark used in our experiments. Counts include only development and test comparisons with a strict majority for Overall. Consensus counts are computed among standard three-rater comparisons: high-consensus is 3–0 and low-consensus is 2–1.

	English–Japanese		English–Spanish	
	Dev	**Test**	**Dev**	**Test**
Comparisons	315	290	307	307
High-consensus	180	160	139	136
Low-consensus	122	116	153	156

A Human Annotation and Benchmark Analysis

A.1 Annotation Protocol and Prompt

Each annotation page presented an English reference utterance and two target-language candidate utterances, denoted A and B. Raters first selected the better candidate overall, with no overall TIE option. They then compared the candidates along the five dimensions defined in Section 3: Speaker, Emotion, Timing, Pronunciation/Accent, and Audio Artifacts. Each dimension-level question allowed A, TIE, or B, and the page additionally provided an optional comment field. The deployed interfaces were bilingual. We reproduce the English instructions below, replacing the specific target-language name with [*target language*] so that the protocol applies to both of our language settings.

Instructions for Human Preference Annotation

- First listen to the English reference. Then compare the two [*target language*] speech candidates and choose which is better overall.
- After choosing the Overall winner, judge the five dimensions below independently. A candidate may win Overall while losing or tying on an individual dimension.
- Use TIE only for the five dimension-level judgments when there is no clear difference. Overall requires A or B.
- Do not evaluate translation accuracy or whether the [*target language*] text is a literal translation of the English. Both candidates use the same target text. Focus only on audible speech characteristics and technical audio quality.
- Comments are optional.

Overall verdict. Using the English utterance as the reference, which [*target language*] speech candidate is better overall? Choose one winner.

Dimension definitions.

D1 Speaker. Match the reference speaker's vocal timbre, pitch range, vocal weight, and speaking style.

D2 Emotion. Match the type and intensity of emotion in the English reference; do not simply reward more emotion.

D3 Timing. Pauses, duration, speaking rate, rhythm, and natural [*target language*] delivery all match the English reference.

D4 Pronunciation/Accent. Clear, correct, authentic [*target language*] articulation and pronunciation, without unnatural foreign-accent leakage.

D5 Audio Artifacts. Only noise, clipping, cutoffs, glitches, missing/extra sounds, or audio artifacts. Do not penalize emotion, timing, or pronunciation again here.

A.2 Annotation Results and Benchmark Composition

Table 6 reports only the development and test comparisons used in our experiments. Most received three independent annotations. To assess whether three-person majorities remain stable with a larger panel, we also collected 20 annotations for 30 shared anchor comparisons. These anchors are included in the benchmark totals but not in the high- and low-consensus counts, which refer strictly to 3–0 and 2–1 outcomes among the standard three-rater comparisons. Comparisons without a strict majority for Overall are excluded from the benchmark.

A.3 Do the Five Dimensions Explain Overall Preference?

We test whether the five dimensions form a coherent account of holistic human preference rather than an arbitrary diagnostic checklist. For comparison i and dimension d, let c_{id}^A and c_{id}^B be the numbers of votes for A and B among R_i raters, and define the signed preference score

$$z_{id} = \frac{c_{id}^A - c_{id}^B}{R_i}, \qquad s_i = \sum_{d=1}^{5} z_{id}. \tag{1}$$

A dimension-level TIE contributes zero to z_{id}. For the primary check, we fit a one-feature logistic calibration, $P(y_i = A \mid s_i) = \sigma(\alpha + \beta s_i)$, where y_i is the human Overall majority label. On the 302 standard three-rater English–Japanese development comparisons, five-fold out-of-fold evaluation keeps comparisons derived from the same source group in the same fold and gives 93.0% accuracy, 0.209 log loss, and a 0.059 Brier score. A more flexible logistic model with dimension-specific preference and TIE-rate features performs no better (92.1% accuracy, 0.228 log loss, and 0.065 Brier score). The same equal-weight analysis reaches 82.6% accuracy on 293 English–Spanish development comparisons, whose annotations exhibit lower agreement. These results show that the five perceptual judgments form a coherent rubric: together, they capture most holistic choices without requiring a learned dimension hierarchy. They do not establish that the dimensions are exhaustive or that Overall preference is causally determined by their equal-weight sum.

A.4 Agreement Analysis

Why report human agreement? Multi-dimensional speech-quality judgments are intrinsically subjective: two attentive listeners can hear the same candidates yet disagree about which difference should determine the Overall verdict or whether a perceptual difference is large enough to avoid a TIE. We therefore report a human agreement reference to contextualize model accuracy, not as a strict upper bound on achievable performance. A model near 70% accuracy should be interpreted differently when human labels themselves exhibit substantial disagreement than when the task has nearly deterministic labels.

Primary estimator: single rater versus panel gold. For comparison i, let $c_i(y)$ denote the number of panel votes for label y, let R_i be the number of raters, and let g_i be the panel-majority gold label. Our reference accuracy is

$$\frac{1}{N} \sum_{i=1}^{N} \frac{c_i(g_i)}{R_i}, \tag{2}$$

which is the expected accuracy of drawing one panel rater uniformly at random and scoring that judgment against the panel gold. Every comparison receives equal weight. For dimension-level judgments, we form the corresponding expected confusion matrix over comparisons with a unique panel-majority label, matching the model evaluation, compute three-class macro-F1 for each dimension, and average across the five dimensions. TIE recall analogously averages the fraction of raters choosing TIE on dimensions whose panel gold is TIE. For consensus-stratified human agreement, we use only standard three-rater comparisons: agreement is 100% on high-consensus comparisons and 66.67% on low-consensus comparisons; the separate 20-rater anchors are excluded from these two estimates.

This estimator is useful because it has the same interpretation and metric scale as the reported model scores. It is mildly optimistic—the sampled rater also contributes to the majority label—so we call it a human agreement *reference*, rather than an independent-rater ceiling. On the English–Japanese test set it yields 85.82% Overall accuracy, 78.84% dimensional macro-F1, and 86.07% dimensional TIE recall. These values quantify the subjectivity of the benchmark; they do not excuse model errors or imply that human disagreement is irreducible.

TIE behavior on the constructed test set. Table 7 characterizes how often each system uses TIE on all 290 comparisons in the English–Japanese test set. The relatively high human-majority TIE prevalence describes the composition of this benchmark, not a universal property of speech evaluation. Direct Gemini is substantially more decisive than the human panel; calibrated Gemini moves its TIE profile toward the human distribution, and the OPD judge moves closer still overall, although

Table 7: **Dimensional TIE prevalence on our constructed English–Japanese test set.** Human values are the percentages of comparisons whose panel-majority label is TIE; model values are the percentages predicted as TIE. The OPD judge uses an SFT-trained teacher; its results are averaged over three seeds.

Dimension	Human-majority TIE (%)	Direct Gemini TIE (%)	Calibrated Gemini TIE (%)	OPD judge TIE (%)
D1: Speaker	52.03	9.59	18.08	34.32
D2: Emotion	25.38	3.08	4.62	4.36
D3: Timing	39.63	7.41	12.59	4.57
D4: Pronunciation and accent	70.42	21.48	46.83	72.89
D5: Audio artifacts	87.89	57.44	64.71	87.77

the effect is strongly dimension-dependent. In particular, the OPD judge closely matches Pronunciation/Accent (D4) and Audio Artifacts (D5) but remains highly decisive on Emotion (D2) and Timing (D3). Matching marginal TIE prevalence also does not imply comparison-level correctness, which is why dimensional macro-F1 remains the primary diagnostic metric. Here, calibrated Gemini uses human-calibrated domain hints on the 272 synthesized-candidate comparisons and the audio-only policy on the 18 natural-versus-synthesized controls. Accordingly, Table 2 evaluates the labeling-policy comparison on the locked 272-comparison primary subset, whereas the dimension-wise rates below describe all 290 comparisons; its aggregate TIE MAE therefore cannot be reconstructed by averaging the full-set rates below.

Alternative agreement summaries. Raw pairwise agreement measures how often two raters select the same label, and Krippendorff's α additionally corrects for agreement expected from the label marginals. Both are valuable descriptions of annotation reliability, but neither is directly comparable to model accuracy or macro-F1. Leave-one-rater-out agreement is more independent in principle, but is severely downward biased for three-rater panels: removing one majority voter from a low-consensus comparison leaves a 1–1 tie among the remaining raters. Our 20-rater anchors permit a less biased leave-one-out diagnostic, but they are too few to represent the full test distribution. We therefore use single-rater-versus-gold agreement in the main table and treat pairwise, chance-corrected, and large-panel estimates as complementary reliability analyses.

B DATA CONSTRUCTION AND HUMAN-CALIBRATED DOMAIN HINTS

B.1 CONSTRUCTING DIVERSE SPEECH COMPARISONS

Candidate construction. For data construction, we partition the in-house reference utterances into training, development, and test sets using an 8:1:1 split. All utterances from the same speaker are assigned to the same split, preventing closely related speech from crossing evaluation boundaries. We use publicly available synthesis systems, including Qwen3-TTS-12Hz-1.7B-Base and CosyVoice2-0.5B (Hu et al., 2026; Du et al., 2024), to construct candidate pools with diverse perceptual characteristics and failure modes. For Speaker, we perform voice conversion using same-speaker and different-speaker conditioning references, creating variation in the preservation of speaker characteristics. For Pronunciation/Accent, we contrast target-language voice cloning with cross-lingual conditioning, which can introduce foreign-accent leakage. For Emotion, Timing, and Audio Artifacts, we sample 12 outputs from each of two multilingual TTS architectures. This multi-output pool captures both within-model stochastic variation and systematic differences between synthesis families. Within each target dimension, candidates are paired randomly to form the speech comparisons. Domain measurements are used to characterize and audit the resulting pool, not to select the winner within a comparison. We then deterministically randomize their assignment to positions A and B and balance the retained pool across target dimensions and candidate positions. This procedure produces comparisons ranging from clear to subtle while reducing shortcuts based on candidate position or synthesis system.

Human benchmark and scalable pool. From the development and test splits, we sample dimension-balanced subsets for human annotation. We also add a small natural–synthesized con-

Table 8: Candidate domain metrics, human validation, and use in the final soft hints. OOF denotes grouped out-of-fold evaluation against human development-set verdicts; pilot listening studies were conducted by linguist experts. **Retained** metrics enter the probability mapping, *auxiliary* metrics support construction or auditing, and rejected metrics are not exposed to the machine labeler.

Dimension	Metric implementation	Human validation and use
Speaker	WeSpeaker `w2vbert2_mfa` embedding cosine to the target-language reference (Wang et al., 2023)	**Retained.** OOF macro-F1 improves from 20.9 to 29.1.
	The same WeSpeaker embedding cosine to the reference	Rejected as the calibrated feature; retained only as a sanity check.
Emotion	Arousal mismatch from audEERING wav2vec2 MSP-DIM (Wagner et al., 2023)	**Retained.** Strongest screened emotion signal; OOF macro-F1 improves from 19.2 to 36.0.
	Valence mismatch from the same MSP-DIM model	Rejected; weaker than arousal in feature screening.
	`emotion2vec+ large` embedding similarity (Ma et al., 2024)	Rejected; unreliable on target-language speech in expert listening.
	F0/pitch-contour match	Rejected; no improvement over arousal in a 30-pair expert pilot.
Timing	Composite duration–envelope score: absolute reference-to-candidate duration-ratio deviation plus scaled amplitude-envelope dynamic time warping (DTW) (Sakoe & Chiba, 1978)	**Retained.** OOF macro-F1 improves from 18.6 to 54.7; adding envelope DTW improves an expert pilot from 66.7% to 74.1% on 27 pairs.
	Silero-VAD speech/pause alignment (Team, 2024)	Rejected; no improvement over duration in expert pilots.
Pronunciation/Accent	Qwen3-ASR-1.7B character error rate (CER) (Shi et al., 2026)	**Retained for Japanese.** OOF macro-F1 improves from 26.7 to 50.6; unsuitable for phonetic Spanish inputs.
	VoxLingua107 ECAPA target-language posterior (Valk & Alumäe, 2021)	**Retained for Spanish**; auxiliary for Japanese pool auditing.
Audio Artifacts	UTMOS22-Strong (Saeki et al., 2022)	Rejected; no OOF improvement over the class-frequency baseline.
	DNSMOS P.835 (SIG, BAK, OVRL) (Reddy et al., 2021)	Rejected; no consistent improvement over UTMOS in a 29-pair expert pilot.
	NISQA-TTS (Mittag et al., 2021)	Rejected; below chance in the same expert pilot.

trol set to test whether the judge can compare candidates with different provenance rather than only candidates produced by synthesis systems. These controls are a secondary test rather than the primary benchmark construction. The resulting benchmark composition is reported in Table 6. Human development judgments are used to select and calibrate candidate metrics; development-set Overall accuracy separately selects trained checkpoints as described in Appendix C. Test judgments remain untouched until final evaluation. The much larger training split receives the resulting soft hints and machine-generated verdict-and-rationale supervision. Thus, the construction procedure creates diverse comparisons, human labels determine which measurements deserve trust, and the held-out test set plays no role in either decision.

B.2 Selecting Domain Metrics for Human-Calibrated Hints

This section provides the metric-selection evidence behind the calibration step in Section 4.2; the retained mappings are subsequently used for scalable supervision as described in Section 4.3. We screened 14 prespecified candidate metrics using grouped out-of-fold macro-F1 against a class-frequency baseline, supplemented by small expert listening pilots when automatic validation was inconclusive; the strongest supported metric for each dimension was retained. Table 8 summarizes the candidate measurements we examined.

C Training and Evaluation Setup

C.1 Training Objectives and Implementation Details

Unless stated otherwise, the student is Qwen2.5-Omni-7B, computation uses bfloat16, the audio encoder and projector are frozen, and LoRA (Hu et al., 2021) is applied to all linear layers of the language model. We implement SFT and RL with ms-swift (Zhao et al., 2025) and OPD with a custom PyTorch/PEFT trainer. The principal SFT and RL experiments use eight NVIDIA A100 80 GB GPUs, while OPD uses eight NVIDIA H200 GPUs. Checkpoints are selected by development-set Overall accuracy. All reported three-seed comparisons use seeds 42, 123, and 456.

The English–Japanese machine-labeling stage covers 11,824 comparisons. After format and confidence filtering, 9,797 enter the SFT scaling pool; percentage conditions use rounded subset sizes, so the 80% setting contains 7,840 comparisons. The English–Spanish release separately contains 8,000 comparisons.

Supervised fine-tuning (SFT). We train with the ms-swift autoregressive SFT implementation on fixed machine-generated responses containing the dimension-level verdicts, rationales, and Overall verdict. The principal setting uses 7,840 comparisons (80% of the scalable supervision pool) and trains for six epochs; every data-size condition uses the same number of epochs. LoRA has rank 128, scaling factor 256, and dropout 0.05. We use a per-device batch size of 2 and four gradient-accumulation steps, giving an effective batch size of 64 across eight GPUs. The learning rate is 5×10^{-5} with cosine decay and 5% warmup; weight decay is zero. The maximum sequence length is 5,120 tokens, and training uses DeepSpeed ZeRO-2. The audio-path ablations retain this recipe while extending LoRA to the 4.6-million-parameter audio-to-LLM projector, to the audio tower's attention and feed-forward layers, or to both. The remaining audio-tower parameters stay frozen.

On-policy distillation (OPD). For each training input, the student greedily generates a trajectory of at most 1,024 tokens. The frozen teacher is then evaluated on the same trajectory, and we minimize the mean full-vocabulary Jensen–Shannon divergence between their next-token distributions at every generated position. Vanilla OPD gives both models the ordinary judge input, whereas privileged OPD additionally gives the teacher the evidence brief associated with that comparison. The student is initialized either from Qwen2.5-Omni-7B or from SFT-trained weights. Teachers are a frozen base or task-adapted Qwen2.5-Omni-7B, or Qwen3-Omni-30B-A3B-Instruct. The principal runs use 7,840 comparisons and LoRA with rank 128, scaling factor 256, and dropout 0.05 on the language-model query, key, value, output, gate, up, and down projections. AdamW uses learning rate 5×10^{-5} and weight decay 0.01. The per-device batch size is 1 with eight accumulation steps, giving an effective batch size of 64 across eight GPUs. We train for at most 750 optimizer steps.

Reinforcement learning (RL). We train with DAPO (Yu et al., 2025) as implemented in ms-swift and compute rewards directly from parsed verdicts; no learned reward model or rationale reward is used. Controlled reward comparisons contain 651 labeled prompts and initialize from SFT- or OPD-trained weights. LoRA has rank 64, scaling factor 128, and dropout 0.05. The Overall reward is $+1$ for the correct A/B verdict and -1 for an incorrect, missing, unparseable, or Tie Overall verdict. The dimensional reward is the mean ± 1 score over dimensions whose reference verdict is A or B; reference Tie cases are excluded. We compare Overall-only, dimension-only, additive, and Overall-gated dimensional rewards, using dimensional weight $\lambda = 0.3$ for the primary gated configuration; we separately evaluate GRPO (Shao et al., 2024) as an algorithmic control. We sample eight completions per prompt at temperature 1.0, capping each at 2,048 tokens and the full sequence at 4,096 tokens. The learning rate is 5×10^{-6}, the per-device batch size is 1, gradient accumulation is 32, and training runs for at most 300 optimizer steps.

C.2 Code Release

We release configuration-driven training code for SFT, OPD, and RL. SFT and RL build on ms-swift (Zhao et al., 2025), while OPD uses a custom distributed PyTorch/Transformers/PEFT implementation inspired by OPSD (Zhao et al., 2026). Users can launch each stage through `train/run.sh` after specifying the data, model or adapter, and output paths in the provided YAML configurations. The release includes an example JSONL manifest showing the required

audio paths, transcripts, verdicts, and rationales. Trained checkpoints are not included; users can initialize from the specified public base models or compatible checkpoints.

C.3 Evaluation Metrics and Statistical Protocol

An output is successfully parsed only when its Overall verdict and all five dimension-level verdicts can be extracted; otherwise it is counted as incorrect. Because the Overall gold label is binary, a predicted TIE is also incorrect. For each dimension, comparisons without a strict human-vote majority are excluded from that dimension's F1 calculation. Three-class macro-F1 is computed on the remaining comparisons and then averaged equally across the five dimensions. For consensus-stratified model results, high consensus means a majority share of at least 0.8, and low consensus means a strict majority below 0.8. This rater-count-agnostic rule places the 14 twenty-rater English–Japanese test anchors into 10 high- and four low-consensus comparisons; together with the standard three-rater counts in Table 6, the reported model slices therefore contain 170 and 120 comparisons, respectively. Reported seed variation is the sample standard deviation across independently trained runs, not a confidence interval. Where a paired-bootstrap interval is explicitly reported, we jointly resample reference clusters for both systems over 20,000 bootstrap replicates and report the percentile 95% interval of their metric difference.

C.4 LANGUAGE-AGNOSTIC JUDGE SYSTEM PROMPT

We use the following system prompt for the language-agnostic diagnostic judge.

System Prompt for the Diagnostic Speech Judge

You are an expert diagnostic evaluator of reference-conditioned cross-lingual speech comparisons. You receive three audio clips:

- **Reference**: speech in the source language;
- **Candidate A**: a speech candidate in the target language; and
- **Candidate B**: a speech candidate in the target language.

You may also receive the source-language reference transcript and the target-language transcript. Compare Candidates A and B based on what you hear and return the Overall verdict.

Important constraints.

- Do not invent numerical measurements, timestamps, or signal-processing facts.
- Do not judge translation meaning. Assume both candidates express the intended content unless an audible speech error is clearly present.
- Use transcripts only to localize audible evidence, such as a word, phrase, pause, emphasis, emotional turn, breath, laugh, or the beginning or end of the utterance.
- Prefer concrete audible observations over generic claims. For example, describe the audible speaker characteristics rather than merely saying that they match.
- Select `[[A]]` or `[[B]]` only when there is a concrete audible difference. Otherwise select `[[tie]]` and briefly explain why.
- Judge each dimension independently before deriving the Overall verdict. Include a "TIE note" only when the dimension-level verdict is `[[tie]]`.

Conflicting evidence. Describe trade-offs honestly: one candidate may be better on some dimensions and worse on others. Base the Overall verdict on the most consequential audible differences rather than counting dimension-level winners. When the trade-off is subtle, state the uncertainty while still returning the required Overall A/B verdict.

Dimension definitions.
D1 Speaker. Use the reference as the anchor. When audible, first characterize its vocal timbre, pitch range, vocal weight, and speaking style; then identify which candidate better preserves these attributes. Exact cross-language voice identity is not required. Overall similarity in audible speaker characteristics is more important than identical timbre. Localize the observation to a phrase or change in delivery when possible.
D2 Emotion. Use the reference as the anchor. Compare emotional type and intensity, pitch movement, stress, hesitation, warmth, urgency, restraint, and emotional arc. Matching may mean remaining calm when the reference is calm, rather than being more expressive. Name the audible emotion or attitude when possible. Keep emotion mismatch or limited expressiveness here; pronunciation and accent belong to D4.
D3 Timing. Use the reference as the anchor while also considering absolute naturalness. Compare rhythm, speaking rate, pause placement, duration fit, breath timing, and the naturalness of the delivery. Do not reward rushed or stretched speech merely because its duration matches the reference. Localize observations to words, phrase boundaries, pauses, breaths, or utterance boundaries.
D4 Pronunciation/Accent. Judge primarily from the candidates. Listen for clear target-language articulation, natural phoneme realization, source-language accent leakage, or pronunciation that sounds non-native or strained. Do not penalize natural regional pronunciation or favor one regional standard. Separate pronunciation and accent from breathy or nasal voice quality, emotional stiffness, and audio artifacts. Localize issues to words, syllables, phones, or phrase-level intonation when possible.
D5 Audio Artifacts. Judge primarily from the candidates and consider only technical defects: background noise or hiss; clipping, cutoff, dropped endings, or missing audio; glitches, clicks, pops, discontinuities, or unstable or distorted audio; unexpected extra sounds; and robotic or metallic artifacts that affect the signal itself. Do not use D5 for speaker characteristics, emotion, timing, pronunciation, or accent. If neither candidate has clear audio artifacts, select `[[tie]]`. Do not treat context-appropriate room tone or reverberation as a defect. Localize any defect to where it occurs.

Required output format.
[Reference Anchor]

- Describe the speaker profile.
- Describe the emotion or intent.
- Describe salient pacing, pauses, breaths, or other audible cues.
- Give a transcript-localized cue useful for comparing the candidates.

For each dimension, use:
[Dimension Name]
Candidate A: `<audible observation>`
Candidate B: `<audible observation>`
TIE note: `<include only for [[tie]]>`
Dimension-level verdict: `[[A]] / [[B]] / [[tie]]`

[Overall Verdict]
Summarize the main audible reasons for the Overall verdict in one to three sentences. Explicitly mention important trade-offs.
Overall verdict: `[[A]] / [[B]]`

D ADDITIONAL EXPERIMENTAL RESULTS

Table 9 consolidates secondary baselines and controls omitted from the main text.

A. Zero-shot baselines. The zero-shot comparison covers proprietary and open audio-language models of different sizes. Their Overall and dimensional performance varies widely, and no model provides consistently balanced diagnostic behavior without task-specific training. We evaluate the Gemini (Gemini Team, 2025; Google DeepMind, 2026), Qwen Omni (Xu et al., 2025a;b), Step-Audio (Wu et al., 2025), Kimi-Audio (Ding et al., 2025), and MOSS-Audio (Yang et al., 2026) model families.

B. Amount of scalable supervision. Holding the SFT recipe and number of epochs fixed, performance improves over the smaller fractions but declines when the full machine-generated supervision pool is used. This non-monotonic trend confirms that more imperfect supervision is not necessarily better.

C. Target and input controls. These targeted ablations vary whether SFT receives rationales, a reference summary, transcripts, or language-parametrized instructions. The results show that performance does not hinge on any single textual field; the additional natural-speech control also checks that the learned comparison is not confined to pairs of synthesized candidates. Because these are single-run controls, we treat their differences as sensitivity evidence rather than model rankings.

D. OPD conditioning and initialization. Privileged task information changes Qwen3 from an ineffective teacher into the strongest teacher for Overall accuracy, whereas changing the student initialization has a smaller effect. Updating the audio encoder can shift the Overall–dimensional balance, but does not remove the teacher-specific profile transferred by OPD.

E. Dual teachers and target language. The task-adapted Qwen2.5 teacher better preserves the dimensional rubric, whereas Qwen3 provides stronger Overall-verdict supervision, motivating us to test whether their signals are complementary. *MeanBlend* averages the two next-token distributions before distillation; *CorrectGated* uses only teachers whose Overall verdict matches the training target, falling back to the Qwen2.5 teacher if neither does; and *SumJSD* computes a separate distillation loss for each teacher and averages the two losses. MeanBlend is evaluated over three seeds, while the other two are targeted single-run controls. None uniformly dominates the stronger single-teacher configurations; the Spanish rows further test the two teacher choices on a second target language.

F. RL and composed training. Reward definition and RL algorithm both change which behavior improves: verdict-gated RL gives the largest Overall gain from the LLM-only SFT initialization, whereas dimension-only rewards incur the largest loss in dimensional macro-F1. Audio-path adaptation and gated RL each improve Overall accuracy separately, but their combination reaches 71.03%, below the audio-adapted SFT model (72.41%) and the LLM-only SFT model followed by RL (71.84%). In the seed with complete comparison-level predictions, the two interventions correct only five of the same errors (Jaccard overlap $= 0.13$), so the non-additivity is not explained by redundant corrections alone. Starting RL from OPD gives the strongest composed result, while updating the encoder during this stage offers no consistent additional gain.

Test-time self-consistency. For five representative judges, we also sample 20 responses for each test comparison at temperature 0.7 and aggregate the dimension-level verdicts by majority vote (one training seed per judge). Voting usually improves over an average sampled response but does not reliably outperform greedy decoding; a paired bootstrap finds a significant gain in Overall accuracy only for SFT, while the strongest OPD judge slightly declines. Agreement across draws is more useful as a confidence signal (correctness AUC 0.58–0.68), suggesting that test-time sampling exposes uncertainty more reliably than it adds diagnostic capability.

D.1 External Cross-Lingual Evaluation

We additionally evaluate on the public VOX-DUB benchmark (Toloka team, 2025), which contains pairwise candidate comparisons from commercial cross-lingual speech systems, with three human judgments per pair. Its five attributes map approximately to our rubric: voice, emotion, naturalness, pronunciation, and sound quality correspond to Speaker, Emotion, Timing, Pronunciation/Accent, and Audio Artifacts, respectively. The naturalness–Timing correspondence is the weakest, and VOX-DUB provides no Overall verdict. We therefore treat the benchmark as an out-of-distribution validity check rather than a second version of our primary evaluation.

For the English-to-Spanish subset, 21 utterances form 126 system comparisons. Because each commercial system occupies a fixed candidate slot in the released data, we evaluate every comparison in both A/B orders and pool the resulting 252 predictions. Table 10 reports three-class macro-F1. Human lower and upper references score a rater against the majority of the other raters or of all raters, respectively; with only three annotations, these form a more honest bracket than a single

Table 9: **Additional experimental results and controls.** Unless marked ES, results use the English–Japanese test set. Three-run results are mean ± standard deviation across training seeds; single-run results are zero-shot evaluations or targeted ablations. Trained checkpoints are selected by development-set Overall accuracy.

Configuration	Eval.	Runs	Overall Acc. (%)	High-Consensus Acc. (%)	Low-Consensus Acc. (%)	Dim. Macro-F1 (%)
A. Zero-shot audio-language models						
Gemini-2.5-Pro	JA	1	62.41	67.65	55.00	48.58
Gemini-3.5-Flash	JA	1	67.59	74.12	58.33	39.67
Qwen3-Omni-30B	JA	1	57.24	61.18	51.67	41.75
Qwen2.5-Omni-7B	JA	1	13.10	13.53	12.50	28.76
Step-Audio-2-mini	JA	1	49.66	50.00	49.17	13.58
Kimi-Audio-7B	JA	1	43.45	37.65	51.67	36.11
MOSS-Audio-8B	JA	1	56.55	63.53	46.67	33.87
B. SFT: amount of scalable supervision						
20% of the supervision pool	JA	3	68.97 ±0.34	79.61 ±1.48	53.89 ±1.73	49.45 ±2.75
40% of the supervision pool	JA	3	69.31 ±2.39	80.78 ±4.34	53.06 ±1.27	47.92 ±2.52
60% of the supervision pool	JA	3	70.80 ±2.08	82.55 ±1.48	54.17 ±3.00	51.95 ±1.09
80% of the supervision pool	JA	3	69.54 ±0.80	81.76 ±1.02	52.22 ±3.37	52.37 ±0.10
100% of the supervision pool	JA	3	67.82 ±1.05	81.57 ±0.90	48.33 ±1.44	51.14 ±0.44
C. SFT: target and input ablations						
Full verdict-and-rationale supervision	JA	1	70.69	84.12	51.67	51.49
Verdicts only	JA	1	70.69	84.12	51.67	52.96
Without the reference summary	JA	1	74.14	84.71	59.17	48.93
Without transcripts	JA	1	74.14	86.47	56.67	52.45
Language-parametrized instructions	JA	1	72.41	85.88	53.33	53.74
Additional natural-speech controls	JA	1	72.07	83.53	55.83	47.63
D. OPD: teacher conditioning and student initialization						
Vanilla OPD; audio-adapted SFT teacher; LLM only	JA	3	70.11 ±0.80	81.96 ±2.23	53.33 ±2.50	48.83 ±1.53
Vanilla OPD; Qwen3 teacher; LLM only	JA	3	56.78 ±1.44	58.04 ±3.02	55.00 ±0.83	37.51 ±0.93
Vanilla OPD; Qwen3 teacher; LLM + encoder	JA	3	54.83 ±1.50	57.06 ±2.35	51.67 ±1.44	37.63 ±3.78
Privileged OPD; base student; LLM only	JA	3	73.79 ±0.91	87.25 ±0.68	54.72 ±2.93	44.61 ±2.79
Privileged OPD; SFT student; LLM only	JA	3	72.99 ±0.40	85.29 ±1.56	55.56 ±3.15	45.53 ±0.45
Privileged OPD; RL student; LLM only	JA	3	72.07 ±0.34	85.49 ±0.90	53.06 ±0.96	43.32 ±1.47
Privileged OPD; base student; LLM + encoder	JA	3	74.14 ±2.60	86.47 ±1.02	56.67 ±5.00	47.21 ±1.89
Privileged OPD; audio-adapted SFT student; LLM + encoder	JA	3	71.38 ±1.58	85.29 ±1.02	51.67 ±2.89	45.98 ±1.16
E. OPD: dual teachers and language settings						
Dual teacher (MeanBlend); LLM + encoder	JA	3	72.07 ±1.50	86.27 ±1.22	51.94 ±2.55	52.74 ±6.40
Dual teacher (CorrectGated); LLM + encoder	JA	1	71.03	85.29	50.83	49.53
Dual teacher (SumJSD); LLM + encoder	JA	1	70.34	85.88	48.33	56.39
English–Spanish OPD; SFT teacher	ES	1	68.40	76.47	61.99	47.30
English–Spanish OPD; Qwen3 teacher	ES	1	67.10	80.15	56.73	45.62
F. RL: reward, algorithm, and composed-training controls						
LLM-only SFT	JA	3	69.54 ±0.80	81.76 ±1.02	52.22 ±3.37	52.37 ±0.10
LLM-only SFT + gated RL	JA	3	71.84 ±1.05	85.69 ±1.48	52.22 ±0.48	47.64 ±1.01
Audio-adapted SFT	JA	3	72.41 ±1.50	83.53 ±0.59	56.67 ±4.33	52.07 ±0.84
Audio-adapted SFT + gated RL	JA	3	71.03 ±1.38	84.90 ±1.22	51.39 ±1.73	51.71 ±0.25
LLM-only SFT + dimension-only RL	JA	3	71.15 ±1.05	84.71 ±1.56	51.94 ±0.48	45.72 ±1.39
LLM-only SFT + gated RL (GRPO)	JA	3	70.34 ±0.91	83.73 ±0.34	51.39 ±1.73	49.36 ±1.64
LLM-only SFT + gated RL; audio path updated during RL	JA	3	70.00 ±1.19	83.53 ±1.76	50.83 ±1.44	50.96 ±3.51
OPD initialization + gated RL; audio path frozen	JA	3	73.91 ±0.40	85.88 ±0.59	56.94 ±1.27	52.87 ±1.74
OPD initialization + gated RL; audio encoder updated	JA	3	72.76 ±1.50	85.10 ±1.36	55.28 ±1.73	53.99 ±1.13

Table 10: External evaluation on English-to-Spanish VOX-DUB. Values are dimensional macro-F1 (%); D3 (Naturalness) is an approximate match to our Timing dimension.

System	D1 Voice	D2 Emotion	D3 Naturalness*	D4 Pronun.	D5 Audio	Mean
Human upper reference	72.3	78.4	75.8	77.8	79.7	76.8
Human lower reference	50.0	56.4	49.0	64.3	65.5	57.0
Always TIE	29.5	11.4	22.6	25.9	19.9	21.8
Random	25.8	31.1	31.9	29.6	32.5	30.2
Gemini 3.1 Pro	26.1	45.1	35.9	45.0	48.2	**40.0**
EN–JA audio-adapted SFT	19.8	38.9	36.2	51.8	22.5	33.8
English–Spanish SFT	28.7	39.1	33.4	**53.8**	38.4	38.7
Joint-language SFT	25.4	**45.5**	31.4	51.6	**40.3**	38.8

human ceiling. Always-TIE and random baselines are important because the benchmark is highly TIE-heavy, particularly for voice.

The English–Spanish and joint-language judges reach mean macro-F1 of 38.7 and 38.8, close to Gemini-3.1-Pro's 40.0, and both exceed Gemini on pronunciation, the dimension most directly targeted by our supervision. No evaluated model beats the Always-TIE baseline on voice, whose gold labels are 80% TIE and have low inter-rater reliability; the Voice dimension (D1) therefore supplies little discriminative signal on this subset. These results support transfer beyond our constructed test sets while also showing why benchmark-specific label distributions and rubric differences must remain visible.

E HUMAN EVALUATION OF DIAGNOSTIC RATIONALES

E.1 PROTOCOL

Because a diagnostic judge should support its verdicts with audibly verifiable evidence localized to relevant words, phrases, or moments in the speech, we treat evidence grounding and localization as central evaluation criteria. We sample 244 unique rationale comparisons from the English–Japanese test set of 290 comparisons and add 22 repeated assignments for quality control. The two models in every comparison predict the same Overall verdict, preventing a rater from preferring an explanation merely because it accompanies the more accurate Overall verdict. The evaluation set includes 73 comparisons where both responses match the human Overall label and all scored dimension-level labels, 122 where both match the Overall label but at least one dimension-level label differs, and 49 where both make the same incorrect Overall prediction.

Ten Japanese-speaking raters first listen to the English reference and both Japanese candidates, then read two anonymous rationales in randomized order. Model identities and automatic scores are hidden, and the instructions explicitly separate rationale quality from the rater's preferred speech candidate. Of the resulting 266 assignments, one is incomplete, leaving 265 complete judgments. The English wording shown in the annotation interface is reproduced below; the deployed interface also included Japanese translations.

Table 11: Human pairwise rationale evaluation. Each cell is *first model better / second model better / no difference*; all paired models predict the same Overall verdict.

Models	Overall	Communication	Grounding	Consistency	Dim. hygiene
SFT vs. SFT+RL	3/**11**/14	2/3/23	1/**11**/16	1/**6**/21	2/5/21
SFT vs. Gemini 3.1 Pro	**10**/7/9	3/6/17	6/6/14	1/1/24	**9**/1/16
SFT vs. audio-adapted SFT	**10**/6/13	1/3/25	**6**/3/20	2/1/26	**5**/4/20
OPD (Qwen3 teacher) vs. OPD (SFT teacher)	**15**/5/5	**5**/4/16	**13**/2/10	**5**/4/16	**6**/5/14

Instructions for Human Evaluation of Rationales

- Listen to the reference and both candidates before reading either rationale.
- Judge only the quality of the explanation, not which candidate you would have chosen.
- Do not prefer a rationale because its final verdict agrees with your own opinion.
- A rationale that states a difference you cannot hear is worse than one that stays silent.

Comparison questions. For each question, choose *Rationale X is better*, *Rationale Y is better*, *Both are good / no meaningful difference*, or *Both are poor*.

1. Which rationale more clearly explains the important differences between the candidates, and gives more useful directions for improvement?
2. Which rationale supports its claims with more accurate, specific, audibly verifiable evidence (cited words or phrases)?
3. Which rationale's dimension-level observations more logically support its dimension-level judgments and its Overall verdict?
4. Which rationale assigns observations to the correct evaluation dimensions more consistently, and avoids conflating different kinds of issues?

Overall rationale preference. Overall, which rationale is the better diagnostic explanation? Choose *Rationale X*, *Rationale Y*, or *Equally good*.
Required comment. Briefly say why. One or two sentences.

E.2 Pairwise Results

Table 11 reports comparisons after mapping the anonymous presentation order back to model identity. The four model pairings shown account for 108 of the 265 complete judgments; the remaining 157 cover auxiliary baseline and ablation pairings not used for the claims below. "No difference" pools the two responses in which raters found both rationales comparably good or comparably poor. The SFT–RL comparison supports the main-text finding: most judgments find the rationales similar, but perceived differences favor RL, especially for audible-evidence grounding. For evidence grounding, 11 of the 12 directional judgments favor SFT+RL over SFT (91.7%; two-sided exact sign test $p = 0.006$); the remaining 16 judgments find no meaningful difference. The Qwen3-teacher OPD rationales receive an even stronger grounding preference over SFT-teacher OPD, despite often accompanying incorrect dimension-level or Overall verdicts. This is why we do not interpret textual detail or persuasiveness as evidence of diagnostic correctness.

Agreement on repeated comparisons is 27% for overall rationale preference and ranges from 41% to 64% across the four criteria. We therefore use this study to identify large, coherent patterns and illustrative failure modes, not to claim a complete ranking of explanation quality. Figure 5 shows one abbreviated comparison. We also tested a text-only GPT-5.4 grader, but its localization score was only weakly predictive of human grounding preferences (AUC $= 0.53$), while its localized-cue count was no better than chance (AUC $= 0.50$). Because a text-only grader cannot verify what is audible, we do not report its model rankings or use them as evidence of grounding; all grounding conclusions instead rely on listeners who inspected the audio.