AUDIOCALIPER: A CONTROLLED PROBE OF AUDIO-LLMS AS SPEECH QUALITY JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio large language models (audio-LLMs) now support reasoning beyond simple classification. They are increasingly used to train other models. They curate speech data, score text-to-speech outputs, gate deployments, and post-train generators as reward models. However, whether they can be trusted in these roles is rarely tested. In this paper, we present AUDIOCALIPER, an open-source, reusable framework for analyzing how an audio-LLM behaves as a quality evaluator. For each clip, we vary one factor at a time while keeping the rest fixed, either the acoustic degradation, the presence of speech, a quality claim in the text prompt or the answer position. This setup allows us to attribute changes in the output to the manipulated factor. We apply AUDIOCALIPER to LibriSpeech and VCTK using Qwen2-Audio, Qwen2.5-Omni, SQ-LLM, SpeechJudge-GRM, SALMONN-SQA and QualiSpeech. Our results reveal each judge's sensitivity to the audio input and text prompt. Fine-tuned judges reach an acoustic sensitivity of 0.87, while zero-shot judges stay near zero. However, fine-tuning does not eliminate other failures. Three out of six judges, two of them fine-tuned, cannot separate speech from silence, and two score silence higher than real speech. All six judges remain sensitive to quality claims in text prompts. Describing a clip as high quality rather than low quality moves a fine-tuned judge's score by 31% of its rating scale and a zero-shot judge's by 86%. In pairwise comparison, five of six judges answer by the position of the clip, although three of them order the same clips correctly when each is scored on its own. Our findings call for evaluating a judge on the decisions it will support, whether scoring, ranking or rewarding, before its scores are used to train or evaluate other models.

1 INTRODUCTION

Speech quality has traditionally been measured by human mean opinion scores (MOS) and by objective metrics that each target one aspect of it. Intrusive measures such as PESQ and STOI ITU-T (2001); Taal et al. (2011) compare a clip against a clean reference, while non-intrusive metrics such as DNSMOS, NISQA and UTMOS Reddy et al. (2021); Mittag et al. (2021); Saeki et al. (2022) score the degraded clip alone. Because no single metric covers every aspect of quality, toolkits collect dozens of metrics behind one interface Shi et al. (2025b) and recent models predict many of them at once with a single network Shi et al. (2025c;a); Tao et al. (2026). The field is now turning to audio large language models (audio-LLMs) as judges. Trained across many tasks, one model can be prompted for any of these dimensions, describe and compare what it hears and be tuned into a general and interpretable evaluator Wang et al. (2026a); Chen et al. (2025); Manakul et al. (2026). The promise is comprehensiveness rather than superiority on any one metric. The same generality lets such a model train speech systems as well as evaluate them, by curating training data or serving as a reward model when generators are post-trained Zhang et al. (2026). Every metric above receives the audio and nothing else, while an audio-LLM judge also reads a text prompt and, in comparison, sees which clip comes first. Its score can therefore rest on those inputs rather than on the audio itself. Such failures are not attacks. They arise under ordinary use and inside a training loop they compound, since a generator optimized against a biased reward learns the bias Gao et al. (2023).

Also, the evidence offered for adopting these judges is almost always their correlation with human ratings. That evidence is collected where it is cheapest, on clips whose audio is informative and whose text is truthful and there a judge that hears quality and one that infers it from the prompt return the same scores. Correlation certifies that a judge *agrees*, but it cannot say what the agreement rests on. The trustworthiness literature on audio-LLMs, now large enough to survey Luo et al. (2026), does not close this gap. It

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.

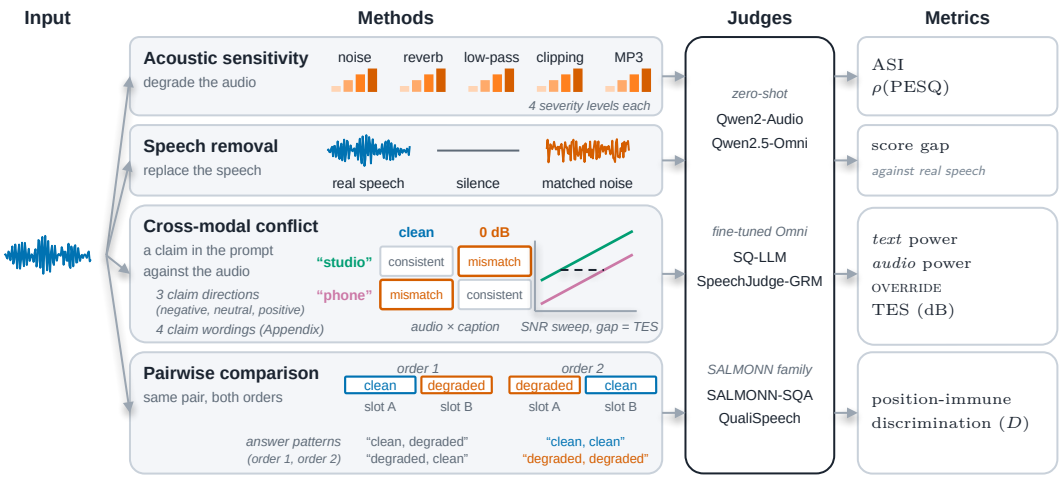


Figure 1: Overview of AUDIOCALIPER. Audio clips pass through four settings, one per row, each changing one factor of the input. All six judges score every condition and each setting is read with its own metrics, shown on the right.

concentrates on adversarial use, such as cross-modal jailbreaks, acoustic backdoors and privacy leakage and has only begun to ask how reliable these models are *as evaluators*. Correlation shows that a judge agrees with human but which part of its input produced the agreement is unclear. A concurrent study of frontier speech-to-speech judges shows what this can hide, since they agree with people on content while their agreement on acoustic dimensions is just a coinflip result David AI (2026). Whether that pattern belongs to one model or to how these models are trained is a question no single judge can answer.

Therefore, we quantify the extent to which audio-LLM quality scores are governed by acoustic input versus textual and positional cues in the prompt. A quality judge should be governed by the recording it receives, its *audio input*. Yet it also receives a *text input*, the prompt, which may supply a transcript, make a claim about the recording and fix the order in which the answer options appear. The question is how much of the score comes from each. Following shortcut-learning analyses in anti-spoofing (Shim et al., 2023), we introduce AUDIOCALIPER which derives every condition from the same utterances and changes one input factor at a time while leaving the rest unchanged, so that any change in the score is attributable to the part we changed. Among the properties an audio-LLM can judge, speech quality is the one whose audio can be degraded in known steps and checked against an objective metric, which is what makes that attribution possible. It is also the setting in which a judge becomes a reward model or a release gate, so a failure there is carried into the systems it trains. We therefore study speech quality alone, and the judges fine-tuned to predict it. The manipulations are defined on the input rather than on the task, so the same protocol can be applied wherever a judge reads audio and text together.

Specifically, we test 200 utterances with six released judges under greedy decoding, with PESQ, STOI and DNSMOS on the identical audio as an objective reference. The conditions vary the type and severity of degradation, replace the speech with silence or matched noise, cross quality claims in the text and present every pair in both orders. We repeat the main settings on VCTK (Yamagishi et al., 2019) to test whether the results hold on a different corpus. Figure 1 summarizes the pipeline.

Our contributions are as follows.

- **Designing a controlled diagnostic protocol.** We introduce a controlled protocol that attributes a judge's score to the part of the input that produced it, without human labels. Because the manipulations are defined on the input rather than on the task, the protocol is designed to transfer to other audio-LLM tasks, such as classification or reasoning over audio. We release the code as an open benchmark, so that a new judge can be put through the same tests before it is trusted.

- **Probing vulnerability and reliability under degradation, silence and position.** We degrade the audio at graded severity, replace the speech with silence and present every pair in both answer orders. These probes show how closely a score tracks audio quality, whether a judge answers from the audio

Table 1: The six released judges we examine, spanning three architecture families and the training axis. Every judge scores a single clip, and the input pairing column gives how two clips are presented when it compares them. For the judges marked *concat*, the pair is presented as clip A followed by 0.5 s of silence and clip B in one waveform, and *native* marks the native two-input path of GRM.

Judge	Audio encoder	LLM backbone	Training	Scale	Input pairing
Qwen2-Audio	Whisper-v3	Qwen2	zero-shot	1–5	concat
Qwen2.5-Omni	Whisper-v3	Qwen2.5	zero-shot	1–5	concat
SQ-LLM	Whisper-v3	Qwen2.5	quality FT	1–5	concat
SpeechJudge-GRM	Whisper-v3	Qwen2.5	SFT + RL, ref.-cond.	1–10	native
SALMONN-SQA	Whisper-v2 + BEATs	Vicuna	MOS FT	1–5	concat
QualiSpeech	Whisper-v2 + BEATs	Vicuna	quality FT (descriptive)	1–5	concat

at all and whether a pairwise choice follows the better clip or the position. They also show whether a score can be trusted on a single clip or only across many.

- **Assessing cross-modal conflict.** Beyond varying the claim in the prompt over the same audio, we cross a studio or phone claim with clean or degraded audio. This measures the effect of the text and the effect of the audio on the same scale and shows which one decides the score when they disagree.

2 Related Work

Speech quality assessment moved from signal-processing measures (Quackenbush et al., 1988; Hu & Loizou, 2008) to perceptual models of MOS such as PESQ, POLQA, ViSQOL, P.563 and STOI (ITU-T, 1996; 2001; Beerends et al., 2013; Hines et al., 2015; Malfait et al., 2006; Taal et al., 2011) and then to learned predictors that regress MOS from the audio or from self-supervised features (Lo et al., 2019; Reddy et al., 2021; 2022; Mittag et al., 2021; Cooper et al., 2022; Saeki et al., 2022) or estimate many metrics at once (Shi et al., 2025b;c;a; Tao et al., 2026; Wang et al., 2026c). These predictors are validated by their correlation with listeners, which is far higher for systems than for single utterances (Huang et al., 2022; 2024; Cooper & Yamagishi, 2021; Maniati et al., 2022; Huang et al., 2026a) and recent work questions the fairness of MOS itself (Ren et al., 2026; Huang et al., 2026b). Audio-LLMs extend assessment to any dimension that can be named in a prompt. After weak zero-shot results (Wang et al., 2025a), supervision produced purpose-built judges (Chen et al., 2025; Wang et al., 2026a; 2025b; Zhang et al., 2026) that agree closely with human rankings (Manakul et al., 2026; Chiang et al., 2025). The evidence is still correlation with listeners. That is enough for a predictor that receives only audio. An audio-LLM judge, however, also receives a text prompt and when it compares two clips it sees the order they arrive in. Its score can therefore follow these cues rather than the audio.

Reading such cues instead of the content is a well-known failure of text LLM judges, which favor answers by position, length and self-authorship and are optimized into their errors when used as rewards (Zheng et al., 2023; Wang et al., 2024; Shi et al., 2025d; Panickssery et al., 2024; Ye et al., 2025; Gu et al., 2024; Raina et al., 2024; Zhao et al., 2025; Norman et al., 2026; Gao et al., 2023), and these concerns now extend to vision and across modalities (Chen et al., 2024; Pu et al., 2025). Audio-LLMs have been examined mostly as assistants under attack (Luo et al., 2026; Li et al., 2026; Kang et al., 2025; Peng et al., 2026; Lin et al., 2026; Wang et al., 2026b) or on recognition tasks, where they underutilize the audio in favor of text (Kuan et al., 2024; Lee et al., 2026; Pang et al., 2026; Yan et al., 2026) and the effect appears as lost accuracy against a correct label, which a quality score does not have. The first studies of audio judges report positional and verbosity biases, calibration failures and weak agreement on acoustics (Manakul et al., 2026; Jeon et al., 2026; Ivry & Watanabe, 2026; David AI, 2026; Park et al., 2026), but they observe these biases on natural data, where the audio, the prompt and the answer order vary together. We instead intervene on one input at a time while the clip stays fixed, so each effect is attributable to the input that moved and we express the pull of the text in acoustic units so it can be weighed directly against the audio. No prior work asks whether the quality scores such judges produce are grounded in the audio, or how a score divides between the audio and the text that accompanies it.

Table 2: Metrics used to report results. Throughout, s is the score a judge returns for one clip and ρ is the Spearman rank correlation. The severity level of a degradation is $\ell \in \{0,1,2,3\}$. For cross-modality experiments, a is the audio condition and t the text condition. m and b are the slope and the intercept of a line fitted to mean score against SNR. In the pairwise comparison, $\Pr(\cdot)$ is the share of pairs falling in an answer pattern. S is the width of a judge's nominal scale, 4 for a judge scoring from 1 to 5 and 9 for GRM. The equation number gives where each metric is defined.

Statistic	Definition	Range
ASI	$\rho(s,-\ell)$ (Eq. 1)	$[-1,1]$
ρ(PESQ)	$\rho(s,\text{PESQ})$, per clip or per condition	$[-1,1]$
text power	$s(a,t_{\text{studio}}) - s(a,t_{\text{phone}})$ at a fixed audio condition	$[-S,S]$
audio power	$s(a_{\text{clean}},t) - s(a_{\text{noisy}},t)$ under a fixed text prompt	$[-S,S]$
OVERRIDE	$s(a_{\text{noisy}},t_{\text{studio}}) - s(a_{\text{clean}},t_{\text{phone}})$ (Eq. 2)	$[-S,S]$
$\bar{m}$	$(m_{\text{studio}} + m_{\text{phone}})/2$ (Eq. 3)	points per dB
TES	$(b_{\text{studio}} - b_{\text{phone}})/\bar{m}$ (Eq. 4)	dB
D	$\Pr(\text{clean twice}) - \Pr(\text{degraded twice})$ (Eq. 5)	$[-1,1]$

3 Experimental settings

Judges. Table 1 summarizes the panel, giving encoder, backbone, training regime, output scale and whether a two-clip comparison is native or must be concatenated. Appendix A describes each judge and explains why the panel spans several constructs.

Dataset. We draw 200 clean utterances from LibriSpeech (Panayotov et al., 2015) `test-clean` and derive every controlled condition from those same clips, so each experiment is a within-clip contrast and no effect can arise from comparing different speakers or different recordings. Appendix C gives clip-bootstrap confidence intervals for the main statistics and shows why 200 clips are enough to resolve every effect we report. To check that the effects are not an artifact of one corpus, we repeat three settings on VCTK (Yamagishi et al., 2019), a studio-recorded multi-speaker corpus. We draw 195 utterances uniformly at random from the corpus, the same way the LibriSpeech clips are drawn.

4 Method

Four settings change one factor of the input at a time. The first degrades the audio, the second removes the speech, the third sets a claim in the text prompt against the audio and the fourth swaps the order of two clips. Each is given below as the manipulation, the conditions it covers and the metric that reads it. Table 2 presents the statistics we report to evaluate the judges with their definitions and ranges. The metrics usually reported for speech quality score a judge against a reference or correlation with human ratings. Instead, our metrics estimate the contrast that one controlled change creates and only the agreement with PESQ compares a judge with anything external. Each is built from familiar estimators, such as rank correlations and differences of means, but combined into a measure defined for this study. We introduce each one with the setting it estimates. Ranks remove a judge's scale and offset. A within-clip difference removes the offset only, so the score-point columns stay in each judge's own units, while TES is in decibels and is comparable across the panel.

4.1 Acoustic Sensitivity

Manipulation. A judge that listens should lower its score as the audio degrades. We keep the text prompt fixed and degrade only the waveform at increasing severity.

Conditions. Five synthetic degradation types at four severity levels each. Additive white noise at SNR $\in \{\text{clean}, 20, 10, 0\}$ dB, clipping of $0/5/15/30\%$ of samples, low-pass filtering at $8/4/2/1$ kHz, reverberation from measured room impulse responses at estimated $T_{60} \in \{0, 0.3, 0.6, 1.0\}$ s, and MP3 coding at $\{\text{none}, 64, 32, 16\}$ kbps. A sixth condition adds recorded point-source noise from RIRS_NOISES (Ko et al., 2017) at the same SNRs, which separates an effect of synthetic noise from an effect of degradation in general. Synthetic degradation is what supplies a known severity ordering and a matched objective metric on the same utterance, which no recorded corpus provides, so the recorded condition serves as the check on it rather than as the primary axis.

Metric. The Acoustic-Sensitivity Index, computed separately for each degradation type,

$$\mathrm{ASI} = \rho(s, -\ell) \in [-1, 1], \tag{1}$$

where $\ell \in \{0,1,2,3\}$ is the severity level and 0 is the unmodified clip. A value near zero means the score is unrelated to how degraded the clip is, and a negative value means the ordering is inverted. ASI takes the designed severity as its reference and needs no labels. We check it against a measured reference with $\rho(\mathrm{PESQ})$ at two levels of aggregation. Per clip it runs over every combination of clip, degradation type and severity. Per condition it first averages the judge's score and PESQ over the 200 clips of each of the 20 conditions, then correlates the 20 pairs of means. A judge can agree at the condition level while disagreeing clip by clip, so the gap between the two says whether its score can be trusted on one recording or only on an average.

4.2 Speech Removal

Manipulation. Removing the speech tests whether the score depends on the audio at all. We replace the speech while fixing the text prompt. Removing the signal entirely is the strongest acoustic manipulation available, so it bounds from above how much the audio can influence the score.

Conditions. The real recording, digital silence of equal length, and white noise matched to the speech in level.

Metric. The score gap against real speech, read as the difference in mean score between the recording and each replacement, in the judge's own score points. Objective metrics on the same three inputs give a reference for what the replacements sound like. If silence scores like real speech, the score cannot be a readout of what was heard.

4.3 Cross-Modal Conflict

Manipulation. We measure how far a claim about quality in the text prompt moves the score while the audio is held fixed. We then measure which input governs the score when the claim and the audio disagree. The claim is varied independently of the clip.

Conditions. Three designs on the same clips. The first scores each clip unchanged under three framings of the prompt, which describe the recording negatively, neutrally with no claim, or positively. The second crosses two audio conditions, the clean clip and the same clip at 0 dB SNR, with two captions, "professional studio recording" and "low-quality phone recording". The third grades the audio more finely, with white noise at nine SNR levels from -20 to 20 dB in steps of 5 dB and the clean clip as a tenth. All ten audio conditions are scored under all three framings, which traces one score against SNR curve per framing.

Metric. One statistic per design. The three framings give the mean score under each framing on identical clean clips, where the neutral framing carries no claim and is the reference that shows whether an effect comes from the positive or the negative side. The 2×2 design gives two marginals and one contrast, writing a for the audio condition and t for the text condition. The *text* power $s(a,t_{\text{studio}}) - s(a,t_{\text{phone}})$ is the effect of the claim at a fixed audio condition, reported on clean and on 0 dB audio. The *audio* power $s(a_{\text{clean}},t) - s(a_{\text{noisy}},t)$ is the effect of the degradation under a fixed claim, reported under each caption. A judge that follows the audio has small text powers and large audio powers. The OVERRIDE puts the two against each other,

$$\text{OVERRIDE} = s(a_{\text{noisy}},t_{\text{studio}}) - s(a_{\text{clean}},t_{\text{phone}}), \tag{2}$$

so a positive value means the degraded clip with the positive claim outscores the clean clip with the negative one. It reads as a conflict only when both powers are positive, so whenever either is negative we report the powers rather than the sign of the OVERRIDE. The SNR sweep expresses the same conflict in decibels, so judges on different scales can be compared. We fit one straight line of mean score against SNR for each text condition, $\bar{s} = m_t \mathrm{SNR} + b_t$, and with the mean slope

$$\bar{m} = (m_{\text{studio}} + m_{\text{phone}})/2, \tag{3}$$

the text-equivalent SNR is

$$\mathrm{TES} = \frac{b_{\text{studio}} - b_{\text{phone}}}{\bar{m}} \quad [\mathrm{dB}], \tag{4}$$

the approximate horizontal distance between the two lines. A larger TES means the claim compensates for more degradation. The two lines are parallel only when their slopes agree, so this is an approximate equivalent shift rather than a constant distance. Appendix E gives the rule that selects the fitted range, together with the fit diagnostics.

Table 3: Acoustic-Sensitivity Index (ASI) for every judge and degradation type at four severity levels each. A judge whose score follows the audio is positive, and a larger value means a more sensitive judge. RIR noise is added at the same SNRs as the white noise. The lower block gives two intrusive metrics and one non-intrusive metric on the identical degraded clips, which bound what is achievable here. In bold are the indices within 0.12 of zero, where the score is unrelated to how degraded the clip is.

Judge	Noise (white)	Noise (RIR)	Clipping	Low-pass	Reverb	MP3 codec
Qwen2-Audio	**0.03**	**0.05**	**0.01**	**0.04**	**0.06**	**−0.04**
Qwen2.5-Omni	**−0.07**	**0.06**	**0.00**	**0.08**	**−0.11**	**0.12**
SQ-LLM	0.77	0.52	0.43	0.28	0.23	0.25
SpeechJudge-GRM	0.40	0.25	**0.12**	**0.10**	**0.05**	**0.06**
SALMONN-SQA	0.87	0.71	0.83	0.80	0.60	0.45
QualiSpeech	0.80	0.54	0.50	0.65	0.39	**−0.03**
PESQ	0.96	0.91	0.96	0.89	0.92	0.97
STOI	0.95	0.88	0.96	0.97	0.95	0.97
DNSMOS P.808	0.95	0.68	0.82	0.88	0.89	0.55

4.4 Position Bias in Pairwise Comparison

Manipulation. Pairwise comparison admits a further failure, in which a judge prefers one answer slot regardless of the audio placed in it. We build each pair from a single utterance, putting the original recording in one slot and a degraded version of the same recording in the other. A judge that hears the difference picks the clean version in both orders, whereas a single presentation cannot separate that from a fixed preference for one slot.

Conditions. Three thousand pairs, from the five degradation types at three severity levels on each of the 200 clips. Every pair is judged twice, with the clean version first in one presentation and second in the other.

Metric. Each answer selects the first or the second position, so the two answers of a pair fall into exactly four patterns. *Always first* and *always second* pick the same position twice, and therefore a different recording in each order. *Clean twice* and *degraded twice* switch position between the two orders, and therefore pick the same recording twice, either the clean one or the degraded one. Only the last two carry a preference that survives the swap. The four shares sum to one, and writing $\Pr(\cdot)$ for the share of pairs falling in a pattern, the position-immune discrimination is

$$D = \Pr(\text{clean twice}) - \Pr(\text{degraded twice}), \tag{5}$$

so that $D \in [-1,1]$. A judge that always prefers the clean clip reaches $D = 1$, and one that decides by position scores $D = 0$, which reflects the absence of any preference that survives the swap rather than an inability to hear the difference. Let A be the share of ordered trials in which the judge names the clean recording. When every pair receives two parseable answers, $D = 2A - 1$, so D reports that accuracy on a signed scale. The four shares add the slot rate, which A does not expose.

5 Results

5.1 Acoustic Sensitivity and Agreement with PESQ

Table 3 shows that whether a judge follows the audio is decided by training rather than by architecture. Every fine-tuned judge is positive on most degradation types. No judge reaches the objective metrics, and every judge responds far more to some degradations than to others. SALMONN-SQA is the most sensitive and comes closest to the objective metrics, while GRM responds clearly only to additive noise. Recorded noise gives the same picture, since every fine-tuned judge stays positive and both zero-shot judges stay near zero. Their insensitivity therefore does not depend on the noise being synthetic.

Reverberation and MP3 coding move the fine-tuned judges least, and QualiSpeech does not respond to MP3 coding at all. The axes are not matched in magnitude, so the weak response to MP3 coding partly follows from it being the mildest manipulation, with a PESQ drop of 2.41 against 3.60 for white noise. Reverberation admits no such explanation. It costs almost as much PESQ and more STOI than any other type, yet

Table 4: Agreement of the quality judges with PESQ. The per-clip column is the Spearman correlation over all 4,000 combinations of clip, degradation type and level. The condition column is the same correlation after averaging each of the 20 conditions over its 200 clips. The ASI column is the mean over the five synthetic types of Table 3. In bold is the judge whose agreement depends most on aggregation.

Judge	ASI (no labels) ↑	per clip ρ(PESQ) ↑	per condition ρ(PESQ) ↑
SQ-LLM	0.39	0.363	0.663
SpeechJudge-GRM	0.15	**0.164**	**0.694**
SALMONN-SQA	0.71	0.667	0.838
QualiSpeech	0.46	0.492	0.728

Table 5: Speech removal experimental results. The columns give the mean score for real speech, for digital silence of equal length, and for white noise matched to the speech in level. A score in bold marks a replacement that is rated at least as high as real speech. The lower block gives two intrusive metrics and one reference-free metric on the identical inputs, where PESQ is undefined on silence because it finds no utterance to align.

Judge	Speech	Silence	Noise
Qwen2-Audio	3.70	**3.95**	3.51
Qwen2.5-Omni	3.54	1.96	1.00
SQ-LLM	3.58	**3.78**	1.04
SpeechJudge-GRM	5.94	2.53	1.35
SALMONN-SQA	4.01	3.00	2.00
QualiSpeech	4.05	**4.00**	1.26
PESQ	4.64	n/a	1.07
STOI	1.00	0.00	0.35
DNSMOS P.808	3.82	2.15	2.10

every fine-tuned judge responds to it far less. A single sensitivity index therefore hides which degradations a judge cannot hear. Appendix D gives the mean score behind every index, at each severity level of each type.

Table 4 shows that this sensitivity does not make a single score trustworthy. Agreement with PESQ tracks each judge's ASI clip by clip, so the label-free index already predicts how well a judge matches an objective metric. Averaging over the clips of each condition raises agreement for every judge and most for the least sensitive ones, because their clip scores scatter widely around condition means that still follow PESQ. For GRM, agreement rises from 0.16 per clip to 0.69 per condition. These judges can therefore compare conditions or systems over many clips, while a single clip's score is too weak to support uses such as a post-training reward or deployment gating.

5.2 Speech Removal

Table 5 shows that following degradation is no guarantee that a judge registers the presence of speech. A judge that depends on the audio should score both replacements well below real speech, yet three of the six rate silence about as high as speech or higher. Two of those three are fine-tuned, so this failure does not follow the training axis. SQ-LLM scores silence above real speech, yet it still pushes white noise to the bottom of the scale. Training on graded distortions therefore teaches a judge to detect artifacts and silence contains none. Silence is the harder test and whether a judge passes it depends on what it was trained on rather than on whether it was fine-tuned, since the zero-shot Qwen2.5-Omni passes while two fine-tuned judges fail.

5.3 Cross-Modal Conflict

Table 6 shows that a claim in the text moves every judge, and that the claim dictates the score for some of them. No judge is immune. The panel splits into three groups, those the text decides, those where the two effects nearly cancel and those the audio still governs. The effect is one-sided for every judge, and which side it takes differs across the panel. The three framings show where each judge's movement comes from. QualiSpeech and SALMONN-SQA already score clean clips near the top of their scale, so a positive

Table 6: Cross-modal conflict between the text and the audio. The first block gives the mean score on identical clean clips under the three framings, with the framing that moves a judge further from the neutral prompt in bold. The remaining columns are defined in Table 2, and arrows give the direction a quality judge should move in, while the mean scores carry no arrow since neither a high nor a low score is right on its own. TES is reported only when the implied shift is no larger than the 50 dB span of the SNR sweep. It is far larger for the two zero-shot judges, so the equivalent-shift reading does not apply to them, and Appendix E gives the fits.

Judge	Mean score (clean audio)			*Text* power ↓		*Audio* power ↑		OVERRIDE↓	SNR shift	
	Negative	Neutral	Positive	Clean	0 dB	Studio	Phone		$\bar{m}$ (per dB) ↑	TES (dB) ↓
Qwen2-Audio	3.83	3.80	**4.22**	0.38	0.21	0.11	−0.06	+0.27	−0.004	outside range
Qwen2.5-Omni	**1.42**	3.46	4.84	3.42	3.45	0.10	0.12	+3.32	+0.005	outside range
SQ-LLM	**2.93**	3.68	4.16	1.23	1.48	1.16	1.40	+0.08	+0.045	32.5
SpeechJudge-GRM	5.21	5.14	**5.78**	0.57	0.93	1.71	2.06	−1.14	+0.073	11.4
SALMONN-SQA	**3.62**	4.10	3.98	0.36	0.42	1.48	1.54	−1.12	+0.054	2.8
QualiSpeech	**2.76**	4.60	4.49	1.73	0.54	2.52	1.33	−0.79	+0.061	10.7

Table 7: Position bias in pairwise comparison. The four patterns and D are defined in Section 4. The final column scores each clip of the same pairs on its own and gives the share whose higher score belongs to the clean recording, over the pairs where the two scores differ, and Appendix H gives the full comparison. In bold are the dominant pattern of each judge, every discrimination that survives the swap, and the judges that order the clips one at a time yet answer by slot in comparison.

Judge	Input pairing	always first ↓	always second ↓	clean twice ↑	degraded twice ↓	D↑	pointwise acc. ↑
Qwen2-Audio	concat	0.002	**0.994**	0.002	0.002	0.00	0.58
Qwen2.5-Omni	concat	**0.992**	0.000	0.006	0.002	0.00	0.48
SQ-LLM	concat	**0.919**	0.000	0.080	0.001	**0.08**	**0.82**
SpeechJudge-GRM	pair	0.036	**0.663**	0.285	0.015	**0.27**	0.71
SALMONN-SQA	concat	**1.000**	0.000	0.000	0.000	0.00	**0.98**
QualiSpeech	concat	**1.000**	0.000	0.000	0.000	0.00	**0.89**

claim leaves them slightly lower rather than raising them. A negative claim pulls both down. All of their movement comes from that side.

The remaining columns weigh the text against the audio. The text decides the scores of the zero-shot judges, the two effects nearly cancel for SQ-LLM and the audio outweighs the text for GRM, SALMONN-SQA and QualiSpeech. Those three also have the smallest text-equivalent SNR (TES), so they keep this precedence across the whole range of SNR and a false studio claim is worth 32.5 dB to SQ-LLM but only 2.8 dB to SALMONN-SQA. For SQ-LLM and GRM the *text* power grows from clean audio to 0 dB, so these judges lean on the text precisely when the audio is poor. The text can also mask what a judge hears, as the phone claim halves the *audio* power of QualiSpeech. Sensitivity to the audio and dependence on the text vary independently, since the two zero-shot judges share an *audio* power near zero while the text moves Qwen2.5-Omni furthest in the panel and Qwen2-Audio least. Appendix E plots the score curves these values are fitted to and Appendix F repeats the design under three further wordings and finds the effect in every one. Across the four fine-tuned judges, sensitivity to the audio rises without a matching fall in sensitivity to the text.

5.4 Position Bias in Pairwise Comparison

Table 7 shows that, used pairwise, five of six judges answer by slot in almost every pair, and four have a position-immune discrimination of exactly zero. The concatenated interface admits two readings of this. Either the judge cannot order the two recordings, or it can and the comparison loses that ordering. Scoring each clip of the same pairs on its own separates the two, in the last column of Table 7 and in Appendix H. The zero-shot judges order the recordings at chance in both formats, so their failure belongs to the judge. SQ-LLM, SALMONN-SQA and QualiSpeech order the recordings well above chance when each is scored alone, yet answer by slot when the two are concatenated, so their failure belongs to the interface. GRM is the only judge with a native two-input path, and among the judges that can order the recordings it is the one whose two readings agree most closely. Where a judge prefers the clean clip in both orders, the preference appears mainly under white noise and grows with severity (Appendix G).

Table 8: Replication on VCTK. ASI is measured on white noise and the override gives the LibriSpeech value before the VCTK one. In bold are the overrides in which the text outweighs the audio.

Judge	ASI ↑	*Text* power ↓		*Audio* power ↑		OVERRIDE ↓
		Clean	Noisy	Studio	Phone	(LS → VCTK)
Qwen2-Audio	−0.05	+0.93	+0.38	+0.55	+0.01	**+0.27 → +0.37**
Qwen2.5-Omni	+0.14	+3.92	+3.16	+0.61	−0.15	**+3.32 → +3.31**
SQ-LLM	+0.75	+1.19	+1.80	+0.73	+1.34	**+0.08 → +0.46**
SpeechJudge-GRM	+0.29	+0.18	−0.02	+1.13	+0.93	−1.14 → −0.95
SALMONN-SQA	+0.91	+0.16	+0.48	+1.48	+1.79	−1.12 → −1.32
QualiSpeech	+0.84	+1.71	+0.51	+2.77	+1.57	−0.79 → −1.06

5.5 Replication on VCTK

To verify that these biases are not unique to LibriSpeech, we repeated the key experiments on 195 utterances from the VCTK corpus. As detailed in Appendix B and Table 8, the sensitivity ordering of the fine-tuned judges and the sign of every override reproduce on VCTK, while effect sizes shift in the direction each judge already leaned.

6 Conclusion and discussion

AudioCaliper measures how controlled changes to the audio and the text prompt affect an audio-LLM judge's score, by degrading or removing the audio, adding a claim about quality to the prompt and swapping the order of two clips, across six released judges. Two findings run through the settings. The first is that the audio's hold on the score weakens as its evidence weakens. SQ-LLM and GRM lean further on the claim at 0 dB than on clean audio, and outside white noise a preference for the clean clip appears mainly under severe degradation. The second is that sensitivity to acoustic degradation does not guarantee robustness to a textual claim or to the presentation format. Judges that track degradation well can still score silence at or above real speech, and three of them lose in a concatenated comparison an ordering they produce when each clip is scored alone. These judges must therefore be tested where the audio and the text disagree, which is not a rare condition. A reward model sees the target text beside a clip that may not contain it, and an evaluation prompt often names the system a clip came from. In each case the text asserts something the audio may not support, which is what our settings create deliberately.

For practitioners, the evidence supports a narrow set of uses. Averaged over dozens of clips, the fine-tuned judges can rank systems. Any use that reads one score, such as a reward for post-training or a gate before deployment, inherits every failure above. For pairwise decisions, a judge without a native two-input path should score each clip on its own and compare the scores, which recovers the ordering for SQ-LLM, SALMONN-SQA and QualiSpeech, and any pairwise judging should be run in both orders with D reported. No single instruction removes the influence of the text, since telling a judge to ignore the claim lowers its *text* power without making it listen, and describe-then-rate moves only the two zero-shot judges toward the audio (Appendix I). Only a reference the judge was trained to use moves the balance back. Withholding GRM's transcript hands the decision to the text, and writing a transcript into the prompt of a judge never trained with one does not help (Appendix J).

These conclusions are limited to the evaluated models and conditions. Scoring every clip under every condition trades clip count for conditions per clip, so the sample is small by intent, and Appendix C shows that it resolves every effect we report. Synthetic degradation is what supplies a known severity ordering and a matched objective metric, and the recorded noise and recorded room impulse responses of Section 5.1 are the check that the findings do not depend on it. Three limits remain. All clips are read English speech from two corpora. The degradations principally stress signal quality, so the naturalness-tuned GRM is examined partly outside its target construct. The panel is limited to released open-weight judges, which fixes every checkpoint and keeps the protocol reproducible, and sampled decoding may shift the size of the effects.

AI USE STATEMENT

Generative AI helped edit prose the authors drafted, produce the figures and recompute reported statistics from the committed result files as a check. No number or figure entered the paper without an author verifying it against those files. The authors take full responsibility for the content.

REPRODUCIBILITY STATEMENT

Every judge reply behind the paper is released with the code, so each table recomputes from those files without a GPU. The six checkpoints are used as published under greedy decoding, the clip samplers are seeded and no audio is redistributed, since the two corpora are public and the repository builds every clip manifest from them. The code and the result files are available at `AUDIOCALIPER`.

REFERENCES

J. G. Beerends, C. Schmidmer, et al. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement Part I—temporal alignment. *J. Audio Eng. Soc.*, 61(6):366–384, 2013.

D. Card, P. Henderson, et al. With little power comes great responsibility. In *Proc. EMNLP*, 2020.

C. Chen, Y. Hu, et al. Audio large language models can be descriptive speech quality evaluators. In *Proc. ICLR*, 2025.

D. Chen, R. Chen, et al. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Proc. ICML*, 2024.

S. Chen, Y. Wu, et al. BEATs: Audio pre-training with acoustic tokenizers. In *Proc. ICML*, 2023.

C.-H. Chiang, X. Wang, et al. Audio-aware large language models as judges for speaking styles. In *Proc. EMNLP*, 2025.

Y. Chu, J. Xu, et al. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

E. Cooper and J. Yamagishi. How do voices from past speech synthesis challenges compare today? In *Proc. ISCA Speech Synthesis Workshop (SSW)*, 2021.

E. Cooper, W.-C. Huang, et al. Generalization ability of MOS prediction networks. In *Proc. ICASSP*, 2022.

David AI. Can large audio language models reliably judge speech-to-speech models? David AI Research Blog, July 2026. [Online]. Available: `https://research.withdavid.ai/blog/lalm-as-judge-vs-hitl` (accessed Sep. 8, 2026).

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *Proc. ICML*, 2023.

J. Gu, X. Jiang, et al. A survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*, 2024.

A. Hines, J. Skoglund, et al. ViSQOL: An objective speech quality model. *EURASIP J. Audio, Speech, Music Process.*, 2015(13), 2015.

Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(1):229–238, 2008.

W.-C. Huang, E. Cooper, et al. The VoiceMOS challenge 2022. In *Proc. Interspeech*, 2022.

W.-C. Huang, S.-W. Fu, et al. The VoiceMOS challenge 2024: Beyond speech quality prediction. In *Proc. IEEE SLT*, 2024.

W.-C. Huang, E. Cooper, and T. Toda. MOS-Bench: Benchmarking generalization abilities of subjective speech quality assessment models. *IEEE Trans. Audio, Speech, Lang. Process.*, 2026a.

W.-C. Huang, N. Sanders, and E. Cooper. CodecMOS-Accent: A MOS benchmark of resynthesized and TTS speech from neural codecs across English accents. *arXiv preprint arXiv:2603.14328*, 2026b.

ITU-T. Methods for subjective determination of transmission quality. Technical Report Recommendation P.800, International Telecommunication Union, 1996.

ITU-T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical Report Recommendation P.862, International Telecommunication Union, 2001.

A. Ivry and S. Watanabe. LALM-as-a-Judge: Benchmarking large audio-language models for safety evaluation in multi-turn spoken dialogues. In *Proc. ICML*, 2026.

J. Jeon, S. Jwa, et al. ParaPairAudioBench: Paralinguistic pairwise audio benchmark for LALM-as-a-Judge. In *Proc. Interspeech*, 2026.

M. Kang, C. Xu, and B. Li. AdvWave: Stealthy adversarial jailbreak attack against large audio-language models. In *Proc. ICLR*, 2025.

T. Ko, V. Peddinti, et al. A study on data augmentation of reverberant speech for robust speech recognition. In *Proc. ICASSP*, 2017.

C.-Y. Kuan, W.-P. Huang, and H. y. Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. In *Proc. Interspeech*, 2024.

J. Lee, J. Pyo, et al. SpeakerSleuth: Can large audio-language models judge speaker consistency across multi-turn dialogues? In *Proc. ACL*, 2026.

K. Li, C. Shen, et al. AudioTrust: Benchmarking the multifaceted trustworthiness of audio large language models. In *Proc. ICLR*, 2026.

L. Lin, M. Yu, et al. Hidden in the noise: Unveiling backdoors in audio LLM alignment through latent acoustic pattern triggers. In *Proc. AAAI*, 2026.

C.-C. Lo, S.-W. Fu, et al. MOSNet: Deep learning-based objective assessment for voice conversion. In *Proc. Interspeech*, 2019.

K. Luo, Z. Zhou, et al. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*, 2026.

L. Malfait, J. Berger, and M. Kastner. P.563—the ITU-T standard for single-ended speech quality assessment. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):1924–1934, 2006.

P. Manakul, W. H. Gan, et al. AudioJudge: Understanding what works in large audio model based speech evaluation. In *Proc. EACL*, 2026.

G. Maniati, A. Vioni, et al. SOMOS: The Samsung open MOS dataset for the evaluation of neural text-to-speech synthesis. In *Proc. Interspeech*, 2022.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

G. Mittag, B. Naderi, et al. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech*, 2021.

J. D. Norman, M. U. Rivera, and D. A. Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-Judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

V. Panayotov, G. Chen, et al. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, 2015.

J. Pang, A. Chaubey, and M. Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with VoxParadox. In *Proc. ICML*, 2026.

A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. In *Proc. NeurIPS*, 2024.

J. Park, D. M. Chan, et al. Auditing protocol-level shortcuts in large audio language model judges for speech evaluation. *arXiv preprint arXiv:2607.13477*, 2026.

Z. Peng, Y. Liu, et al. JALMBench: Benchmarking jailbreak vulnerabilities in audio language models. In *Proc. ICLR*, 2026.

S. Pu, Y. Wang, et al. Judge anything: MLLM as a judge across any modality. In *Proc. ACM SIGKDD*, 2025.

S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective measures of speech quality*. Prentice Hall signal processing series. Prentice Hall, 1988.

A. Radford, J. W. Kim, et al. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

V. Raina, A. Liusie, and M. Gales. Is LLM-as-a-Judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proc. EMNLP*, 2024.

C. K. A. Reddy, V. Gopal, and R. Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*, 2021.

C. K. A. Reddy, V. Gopal, and R. Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*, 2022.

W. Ren, Y.-C. Lin, et al. MOS-Bias: From hidden gender bias to gender-aware speech quality assessment. *arXiv preprint arXiv:2603.10723*, 2026.

T. Saeki, D. Xin, et al. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proc. Interspeech*, 2022.

Z. Shao, P. Wang, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

J. Shi, Y. Cheng, et al. ARECHO: Autoregressive evaluation via chain-based hypothesis optimization for speech multi-metric estimation. In *Proc. NeurIPS*, 2025a.

J. Shi, H. Shim, et al. VERSA: A versatile evaluation toolkit for speech, audio, and music. In *Proc. NAACL (System Demonstrations)*, 2025b.

J. Shi, H.-J. Shim, and S. Watanabe. Uni-VERSA: Versatile speech assessment with a unified network. In *Proc. Interspeech*, 2025c.

L. Shi, C. Ma, et al. Judging the judges: A systematic study of position bias in LLM-as-a-Judge. In *Proc. IJCNLP-AACL*, 2025d.

H.-J. Shim, R. G. Hautamäki, et al. How to construct perfect and worse-than-coin-flip spoofing countermeasures. In *Proc. Interspeech*, 2023.

C. H. Taal, R. C. Hendriks, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, 2011.

C. Tang, W. Yu, et al. SALMONN: Towards generic hearing abilities for large language models. In *Proc. ICLR*, 2024.

Z. Tao, J. Shi, et al. ANCHOR: Autoregressive non-intrusive chunk-ordered refinement for joint multi-resolution speech quality modeling. In *Proc. Interspeech*, 2026.

H. Wang, J. Zhao, et al. SpeechLLM-as-Judges: Towards general and interpretable speech quality evaluation. In *Proc. ACL*, 2026a.

J. Wang, K. Luo, et al. HearSay benchmark: Do audio LLMs leak what they hear? In *Proc. ACL*, 2026b.

P. Wang, L. Li, et al. Large language models are not fair evaluators. In *Proc. ACL*, 2024.

S. Wang, W. Yu, et al. Enabling auditory large language models for automatic speech quality evaluation. In *Proc. ICASSP*, 2025a.

S. Wang, W. Yu, et al. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proc. ACL*, 2025b.

W. Wang, W. Zhang, et al. UrgentMOS: Unified multi-metric and preference learning for robust speech quality assessment. *arXiv preprint arXiv:2601.18438*, 2026c.

J. Xu, Z. Guo, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

J. Yamagishi, C. Veaux, and K. MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive*, 2019.

X. Yan, B. Cao, et al. Beyond text-dominance: Understanding modality preference of omni-modal large language models. *arXiv preprint arXiv:2604.16902*, 2026.

J. Ye, Y. Wang, et al. Justice or prejudice? quantifying biases in LLM-as-a-Judge. In *Proc. ICLR*, 2025.

X. Zhang, C. Wang, et al. SpeechJudge: Towards human-level judgment for speech naturalness. In *Proc. ICLR*, 2026.

Y. Zhao, H. Liu, et al. One token to fool LLM-as-a-Judge. *arXiv preprint arXiv:2507.08794*, 2025.

L. Zheng, W. L. Chiang, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proc. NeurIPS Datasets and Benchmarks Track*, 2023.

APPENDIX

This appendix supports the main text with procedures and results that did not fit within the page limit. Each section states which part of the main text it expands.

- Appendix A describes each judge and explains why the panel spans several constructs, for Section 3.
- Appendix C gives confidence intervals and the sample size they justify, for Section 3.
- Appendix D gives the mean score behind every acoustic-sensitivity index, for Section 5.1.
- Appendix E shows the score curves behind TES, for Section 5.3.
- Appendix F repeats the 2×2 design under four wordings of the same claim, for Section 5.3.
- Appendix G breaks position bias down by degradation type and severity, for Section 5.4.
- Appendix H scores each clip of a pair on its own, for Section 5.4.
- Appendix B gives the full VCTK replication, for Section 5.5.
- Appendix I tests two prompt-level defenses, for Section 6.
- Appendix J tests whether a reference transcript anchors a judge to the audio, for Section 6.

A THE JUDGE PANEL

We examine a panel of **six released judges**, spanning three architecture families and the training axis, meaning whether and how a model was trained to grade quality. The *zero-shot general audio-LLMs* are **(i) Qwen2-Audio-7B-Instruct** (Chu et al., 2024), which pairs a Whisper-large-v3 (Radford et al., 2023) encoder with a Qwen2 LLM and **(ii) Qwen2.5-Omni-7B** (Xu et al., 2025), both used *zero-shot* by prompting. Qwen2.5-Omni is the *base* of the two fine-tuned Omni judges, so examining it zero-shot isolates the effect of fine-tuning from architecture. The *fine-tuned Qwen2.5-Omni judges* are **(iii) SQ-LLM** (Wang et al., 2026a), fine-tuned pointwise on speech-*quality* dimensional ratings and **(iv) SpeechJudge-GRM** (Zhang et al., 2026), a *pairwise reward model* for TTS *naturalness*, trained by supervised fine-tuning on chain-of-thought rationales followed by GRPO (Shao et al., 2024) reinforcement learning, of which we use the released SFT+RL checkpoint. The *SALMONN family* (Tang et al., 2024) follows a different design, with dual Whisper-large-v2 (Radford et al., 2023) and BEATs (Chen et al., 2023) encoders and a Vicuna LLM. It contributes **(v) SALMONN-SQA** (Wang et al., 2025a), fine-tuned for signal-*quality* MOS on NISQA, BVCC and SOMOS and **(vi) QualiSpeech** (Wang et al., 2025b), a descriptive quality model that answers in natural language.

The judges target *different* constructs, covering signal quality, TTS naturalness and quality description, whereas our perturbations principally stress signal quality. That spread is deliberate, because the effects we report concern *how a judge weights audio against text* rather than its training target. It does mean the naturalness-tuned judge is graded partly off its home construct, which plausibly softens its acoustic sensitivity. The SALMONN family also brings a different *encoder*, since its BEATs branch is trained on acoustic events unlike the ASR-tuned Whisper front-end of the Qwen judges, so it is architecturally better positioned to hear non-speech degradation. We use every judge as released and adapt none of them, so the biases we report belong to deployable, off-the-shelf judges rather than to our training choices.

B VCTK REPLICATION DETAILS

On VCTK we repeat three settings, acoustic sensitivity under white noise, the three prompt framings and the 2×2 crossing. White noise is the degradation every judge responds to most, so it gives the replication the clearest signal to reproduce. Table 8 shows that these are properties of the judges rather than of the corpus. Sensitivity reproduces with the same ordering. Which input wins reproduces for every judge, since none changes the sign of its override. What shifts is the size of the effects, and it shifts in the direction each judge already leaned. The text gains ground where it already had some, on Qwen2-Audio and SQ-LLM. The audio gains where it already won, on SALMONN-SQA and QualiSpeech. GRM is the one judge that weakens, from 0.40 to 0.29. SALMONN-SQA again behaves as a quality judge should on both counts, with the highest sensitivity in the panel and the smallest *text* power.

C Confidence Intervals and Sample Size

Table 9 reports clip bootstrap intervals from 1,000 resamples of clips with replacement. Resampling clips rather than scores keeps every condition derived from one utterance together (Card et al., 2020; Miller, 2024). We call an effect resolved when its 95% interval excludes zero. At n=200 every effect the paper rests on is resolved and the intervals are roughly an order of magnitude narrower than the effects, as with the text power of Qwen2.5-Omni at $+3.42$ against a half-width of 0.10. Most effects already clear zero with 10 to 30 clips. The full 2,620-utterance split would shrink every interval by only $\sqrt{2620/200} \approx 3.6$, taking the widest ASI half-width from 0.06 to 0.02, which no conclusion here turns on. Two quantities do not exclude zero, the noise sensitivity of Qwen2-Audio and the OVERRIDE of SQ-LLM, and each is the finding itself rather than an unresolved sample. The noise sensitivity of Qwen2.5-Omni is resolved and negative, at -0.07 with an interval of $[-0.12, -0.01]$, so that judge is mildly inverted rather than insensitive and its score rises slightly as the noise grows.

Table 9: Clip-bootstrap 95% confidence intervals at n=200, from 1,000 resamples. An effect is resolved when its interval excludes zero, and every effect the paper rests on is resolved. Two intervals straddle zero, the ASI of Qwen2-Audio because it does not follow the degradation and the OVERRIDE of SQ-LLM because its two effects cancel, and both are narrow enough to bound the effect near zero rather than leave it undetermined. The *text* power is measured on clean audio. SpeechJudge-GRM uses a 1 to 10 scale and the others use 1 to 5, so the two score-point columns are in each judge's own units.

Judge	ASI (noise)	*Text* power	OVERRIDE
Qwen2-Audio	$[-0.02, +0.10]$	$[+0.32, +0.47]$	$[+0.20, +0.34]$
Qwen2.5-Omni	$[-0.12, -0.01]$	$[+3.32, +3.52]$	$[+3.21, +3.42]$
SQ-LLM	$[+0.75, +0.80]$	$[+1.16, +1.30]$	$[-0.01, +0.17]$
SpeechJudge-GRM	$[+0.34, +0.47]$	$[+0.35, +0.76]$	$[-1.39, -0.89]$
SALMONN-SQA	$[+0.86, +0.89]$	$[+0.29, +0.43]$	$[-1.19, -1.05]$
QualiSpeech	$[+0.77, +0.83]$	$[+1.64, +1.80]$	$[-0.90, -0.69]$

D Acoustic Sensitivity by Severity Level

This section expands Section 5.1 with the mean scores behind each index. Figure 2 and Table 10 show that the two zero-shot judges produce almost the same score at every level of every type, moving within 0.2 points across all twenty conditions, so their curves never descend. The fine-tuned judges fall with severity and white noise moves them most, with SQ-LLM dropping from 3.58 on clean speech to 1.38 at 0 dB. The other four types move them far less and the decline is usually concentrated at the most severe level, while SALMONN-SQA is the only judge whose score falls steadily on every type, which is why it holds the highest index everywhere in Table 3.

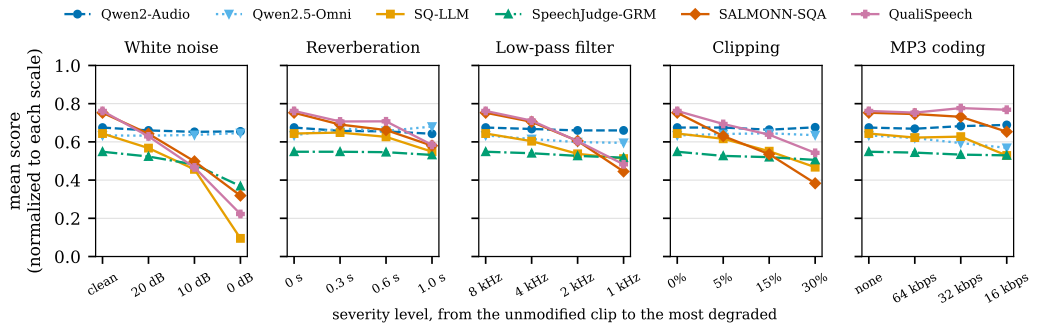


Figure 2: Mean score of each judge at the four severity levels of each degradation type. The leftmost point of each condition is the unmodified clip.

Table 10: Mean score of each judge at every severity level of every degradation type. Level 0 is the unmodified clip, and levels 1 to 3 are increasingly severe. The shaded row of each block gives the value of each level. Score cells are shaded by how far the score has fallen from that judge's own level 0, as a share of its own scale. A darker orange means a larger drop. Shading this way keeps the judges comparable, although GRM rates on 1 to 10 and the others on 1 to 5. The last column repeats the acoustic-sensitivity index of Table 3. It is shaded blue in proportion to its value, or orange when it is negative.

Judge	Level 0	Level 1	Level 2	Level 3	ASI
White noise	clean	20 dB	10 dB	0 dB	
Qwen2-Audio	3.70	3.64	3.61	3.62	+0.03
Qwen2.5-Omni	3.54	3.53	3.54	3.58	−0.07
SQ-LLM	3.58	3.27	2.83	1.38	+0.77
SpeechJudge-GRM	5.93	5.71	5.33	4.32	+0.40
SALMONN-SQA	4.01	3.57	2.99	2.27	+0.87
QualiSpeech	4.05	3.52	2.87	1.89	+0.80
Reverberation	0 s	0.3 s	0.6 s	1.0 s	
Qwen2-Audio	3.70	3.63	3.62	3.57	+0.07
Qwen2.5-Omni	3.54	3.68	3.63	3.72	−0.11
SQ-LLM	3.58	3.60	3.50	3.19	+0.23
SpeechJudge-GRM	5.93	5.93	5.92	5.79	+0.05
SALMONN-SQA	4.01	3.76	3.65	3.32	+0.60
QualiSpeech	4.05	3.83	3.83	3.34	+0.39
Low-pass filter	8 kHz	4 kHz	2 kHz	1 kHz	
Qwen2-Audio	3.70	3.67	3.64	3.64	+0.04
Qwen2.5-Omni	3.54	3.46	3.40	3.38	+0.08
SQ-LLM	3.58	3.42	3.15	3.05	+0.28
SpeechJudge-GRM	5.93	5.87	5.74	5.66	+0.10
SALMONN-SQA	4.01	3.82	3.42	2.78	+0.80
QualiSpeech	4.05	3.85	3.42	2.93	+0.65
Clipping	0%	5%	15%	30%	
Qwen2-Audio	3.70	3.70	3.65	3.71	+0.01
Qwen2.5-Omni	3.54	3.56	3.58	3.54	−0.00
SQ-LLM	3.58	3.46	3.21	2.87	+0.43
SpeechJudge-GRM	5.93	5.74	5.68	5.55	+0.12
SALMONN-SQA	4.01	3.52	3.14	2.53	+0.83
QualiSpeech	4.05	3.77	3.55	3.17	+0.50
MP3 coding	none	64 kbps	32 kbps	16 kbps	
Qwen2-Audio	3.70	3.67	3.73	3.75	−0.04
Qwen2.5-Omni	3.54	3.48	3.38	3.27	+0.12
SQ-LLM	3.58	3.49	3.51	3.12	+0.25
SpeechJudge-GRM	5.93	5.89	5.80	5.76	+0.06
SALMONN-SQA	4.01	3.98	3.92	3.61	+0.46
QualiSpeech	4.05	4.01	4.11	4.07	−0.03

E TEXT-EQUIVALENT SNR (TES) CURVES

Lines are fitted to condition means over the SNR levels that are unsaturated and monotonic, on the support shared by the studio and phone conditions. Levels are dropped from the clean end downward as soon as the mean score rises, since below some SNR the audio carries little information and the judge returns the claim instead of rating the clip, which reverses the curve. Table 11 gives the surviving support, both slopes, the fit quality and a clip bootstrap interval for every judge. For the two zero-shot judges the fitted slope is small and the implied shift is far larger than the span the sweep covers. Both slopes are resolved away from zero, so the shift is an extrapolation rather than an unbounded value.

Figure 3 shows why the equivalent shift is measurable for the fine-tuned judges and not for the zero-shot ones. The zero-shot curves stay nearly level across every SNR, so the gap between the two claims never closes within the range measured. The fine-tuned judges rise toward clean speech, and the claim shifts the whole curve sideways. For SQ-LLM and GRM the rise appears only under the studio claim, so these judges return the claim once the audio carries little information, whereas all three of SALMONN-SQA's

Table 11: Fit diagnostics behind TES. Intervals are clip bootstraps over 1,000 resamples. TES is reported where the implied shift is no larger than the 50 dB span of the sweep, which the two zero-shot judges exceed at roughly −86 dB and +741 dB.

Judge	Fit range (dB)	Levels	m_{studio}	m_{phone}	$\bar{m}$ [95%]	R^2 std./ph.	TES [95%]
Qwen2-Audio	−20 to 30	10	−0.0023	−0.0047	−0.0035 [−0.0046,−0.0026]	0.22 / 0.80	beyond span
Qwen2.5-Omni	−5 to 20	6	+0.0000	+0.0093	+0.0046 [+0.0023,+0.0068]	0.00 / 0.97	beyond span
SQ-LLM	−5 to 30	7	+0.0436	+0.0461	+0.0448 [+0.0430,+0.0468]	0.94 / 0.99	32.5 [30.3,34.9]
SpeechJudge-GRM	−10 to 30	8	+0.0762	+0.0689	+0.0725 [+0.0677,+0.0773]	0.91 / 0.97	11.4 [10.1,12.9]
SALMONN-SQA	−15 to 30	9	+0.0640	+0.0437	+0.0539 [+0.0525,+0.0551]	0.96 / 0.97	2.8 [2.4,3.3]
QualiSpeech	0 to 30	6	+0.0830	+0.0399	+0.0614 [+0.0587,+0.0643]	0.94 / 0.96	10.7 [9.4,12.2]

curves rise together. QualiSpeech shows the mirror image of a positive claim, since its studio curve lies on its no-caption curve while its phone curve sits well below.

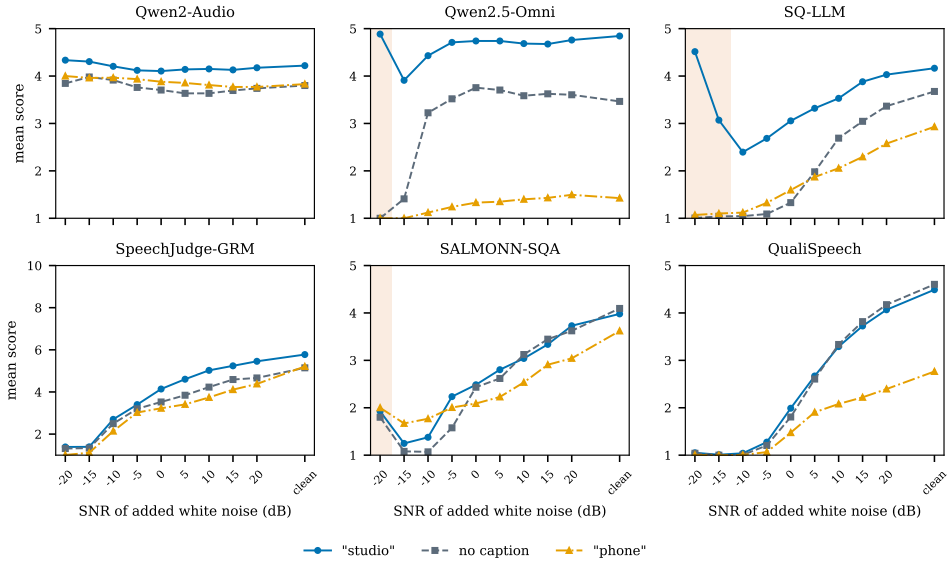


Figure 3: Mean score against the SNR of the added white noise for every judge. Clean speech, which has no finite SNR, sits at the right end of each axis. Shaded parts are left out when the lines behind Table 6 are estimated, either because the curves reverse below them or because one caption curve has reached the end of the judge's scale.

F Caption Wording

To verify whether the wording of the claim changes the result, Table 12 gives four wordings of the same quality claim and Table 13 repeats the 2×2 design under each. As a result, the effect does not belong to the studio and phone wording used in the main text. How far it moves them depends on what the claim asserts rather than on the judge alone, and a claim that states a numerical rating is the strongest of the four. A stated rating is the wording judges follow most closely. Three of them mostly repeat the number the prompt names, and Qwen2-Audio returns it for every clip. It also displaces part of what a judge would otherwise take from the audio, cutting the *audio* power of three fine-tuned judges, most sharply for SQ-LLM from 1.28 to 0.38. The judge least moved by the numerical rating is QualiSpeech, while the scenic wording leaves GRM and SALMONN-SQA almost unmoved.

Table 12: The four wordings of the claim. Each cell is the sentence placed before the rating instruction, which is "Rate its overall quality from 1 to 5. Answer with only a number." in every case. The neutral framing adds no sentence at all and asks "Rate the overall quality of this speech from 1 to 5. Answer with only a number." Provenance is the wording used everywhere else in the paper. Within a wording the two poles assert the same thing, that quality is high or low, so only the form of the claim changes.

Wording	Positive claim	Negative claim
Provenance	"The following is a professional studio recording."	"The following is a low-quality phone recording."
Quality	"The following is a high-quality recording."	"The following is a low-quality recording."
MOS	"Listeners rated this recording 4.5 out of 5."	"Listeners rated this recording 2 out of 5."
Scenic	"The following was recorded in an acoustically treated studio."	"The following was recorded on a noisy street."

Table 13: Caption wording control on the 2×2 design, with the four wordings given in full in Table 12. The left block gives the effect of the claim and the right block the effect of the same 0 dB degradation under that claim. Each value averages over the other factor, so text power is the claim's effect across both audio conditions and audio power is the degradation's effect across both claims.

	text power				*audio* power			
Judge	Prov.	Qual.	MOS	Scenic	Prov.	Qual.	MOS	Scenic
Qwen2-Audio	$+0.29$	$+0.74$	$+3.00$	$+0.92$	$+0.02$	-0.00	$+0.00$	-0.06
Qwen2.5-Omni	$+3.43$	$+3.92$	$+2.00$	$+3.71$	$+0.11$	-0.03	$+0.00$	$+0.11$
SQ-LLM	$+1.36$	$+1.47$	$+2.59$	$+1.67$	$+1.28$	$+1.47$	$+0.38$	$+1.31$
SpeechJudge-GRM	$+0.75$	$+0.76$	$+2.20$	$+0.03$	$+1.88$	$+1.74$	$+0.95$	$+1.90$
SALMONN-SQA	$+0.39$	$+0.30$	$+1.49$	$+0.06$	$+1.51$	$+1.32$	$+0.57$	$+1.47$
QualiSpeech	$+1.13$	$+1.41$	$+0.44$	$+0.39$	$+1.92$	$+1.83$	$+2.10$	$+2.09$

G Position Bias by Degradation Level

Table 14 asks how large the acoustic difference must be before the audio overrides the slot, by breaking the discrimination down by degradation type and severity. The other four judges are left out because each of those four picks the same slot in at least 96% of pairs, whatever the degradation type or severity, so their discrimination never rises above 0.02 anywhere, not even where the two clips differ most. Our result shows that a preference for the clean clip appears mainly under white noise and mainly at its strongest levels, so the slot decides most trials in which the two clips sound alike. For GRM, white noise is the only degradation on which the preference holds at every level. It reaches 0.74 at the medium level. The other four types stay at most 0.12 for mild degradation. They grow to between 0.25 and 0.38 only when the degradation is severe. SQ-LLM repeats that pattern on a smaller scale, with a discrimination near zero except under the strongest white noise reaching 0.76.

Table 14: Position-immune discrimination D of SpeechJudge-GRM and SQ-LLM for each degradation type at each severity level. Mild, medium and severe are levels 1, 2 and 3 of Table 10. $D{=}1$ means the clean clip is always preferred and $D{=}0$ means no preference survives the swap.

	SpeechJudge-GRM			SQ-LLM		
Degradation	Mild	Medium	Severe	Mild	Medium	Severe
noise (white)	0.51	0.74	0.70	0.01	0.12	0.76
clipping	0.07	0.19	0.38	0.00	0.00	0.08
low-pass	0.06	0.16	0.27	0.00	0.02	0.12
reverb	0.12	0.16	0.34	0.00	0.01	0.03
MP3 codec	0.01	0.09	0.25	0.00	0.00	0.04

Table 15: Pointwise control for the pairwise setting. Each clip is scored on its own, and a pair counts as decided when the two scores differ. Pointwise accuracy is the share of decided pairs whose higher score belongs to the clean recording. Pairwise accuracy is the share of ordered trials that name the clean recording, taken from Table 7. The gap separates a judge that cannot order the two recordings from a judge whose comparison interface loses an ordering it can produce. In bold are the judges whose ordering survives one clip at a time but not in comparison.

Judge	Input pairing	Pairwise acc. ↑	Decided	Pointwise acc. ↑	Gap
Qwen2-Audio	concat	0.50	0.17	0.58	+0.08
Qwen2.5-Omni	concat	0.50	0.21	0.48	−0.02
SQ-LLM	concat	0.54	0.48	**0.82**	**+0.28**
SpeechJudge-GRM	pair	0.64	0.29	0.71	+0.07
SALMONN-SQA	concat	0.50	0.69	**0.98**	**+0.48**
QualiSpeech	concat	0.50	0.53	**0.89**	**+0.39**

H Pointwise Control for Pairwise Comparison

Five of the six judges receive the two clips concatenated into one waveform, so a near-constant answer could mean the judge never located the boundary between them rather than that it answers by slot. Table 15 separates the two readings. It scores each clip of the same 3,000 pairs on its own, under the single-clip prompt of Section 5.1, and asks whether the clean recording receives the higher number.

The answer divides the panel. SALMONN-SQA answers by slot in every pair, yet one clip at a time it puts the clean recording above the degraded one in 98% of the pairs it decides. QualiSpeech and SQ-LLM behave the same way. The ordering is available to these judges, and the concatenated interface loses it. The two zero-shot judges sit at chance on this test as well, so their pairwise failure is not an interface effect. GRM is the only judge that takes the two clips through its own two-input path, and it is also the judge whose pairwise accuracy comes closest to its pointwise accuracy. Position bias is therefore a property of the comparison interface for judges that can hear the difference, and a property of the judge for those that cannot.

This control has two limits. A pair is undecided when the two scores are equal, which is frequent because the scales are coarse, so the decided column varies widely across the panel. The pointwise route also asks two separate questions rather than one comparative question, so it bounds what a judge can order rather than reproducing the comparison.

I Prompt-Level Defenses

This section asks whether the influence of the text can be removed by changing the instruction alone, without retraining a judge or altering its input. Each judge is run through the framing and OVERRIDE settings three times, with the same claims as in the main experiments and one of three instructions. GRM is left out of this test. Its prompt is a fixed reference-conditioned template rather than a free instruction, so replacing the instruction would change its interface rather than only its wording. Appendix J takes the equivalent step for GRM, by manipulating the one field that template does expose.

Baseline. "Rate the overall quality of this speech from 1 to 5. Answer with only a number." This carries no defense, and the two defenses below are measured against it.

Anti-suggestion. "Ignore any claims about how or where this was recorded; judge only the audio you actually hear." followed by the baseline instruction.

Describe-then-rate. "First, in one sentence, describe any noise, distortion, or artifacts you actually hear. Then rate the overall quality of this speech from 1 to 5, writing the number at the very end."

Table 16 asks whether an instruction alone can stop a judge from following the claim. It reports the text power on clean audio, the OVERRIDE and the *audio* power, the last as a guardrail, since an instruction that made a judge ignore its input entirely would also drive the first two to zero. Each one lowers the text power of some judges and raises it for others, so no instruction can be recommended without knowing the judge in advance. The one instruction that lowers a bias while keeping the audio effect intact does so only for the two judges that were barely listening to begin with. Anti-suggestion cuts the text power of Qwen2.5-Omni

Table 16: Prompt-level defenses in the framing and OVERRIDE settings. Each cell gives the baseline value before the value under that defense, and Appendix I gives every instruction in full. A defense should lower the *text* power and the OVERRIDE while keeping the *audio* power high, which acts as a guardrail, since an instruction that made a judge ignore its input entirely would also drive the first two to zero. In bold are the cells that move the other way. (†QualiSpeech states no rating in 38% of its describe-then-rate replies, at a rate that follows the condition, so what remains is not comparable across conditions.)

Judge	anti-suggestion			describe-then-rate		
	text ↓	OVERRIDE↓	*audio* ↑	*text* ↓	OVERRIDE↓	*audio* ↑
Qwen2-Audio	$+0.75 \rightarrow +0.58$	$+0.44 \rightarrow +0.20$	$+0.07 \rightarrow +0.11$	$\mathbf{+0.75 \rightarrow +0.83}$	$+0.44 \rightarrow +0.21$	$+0.07 \rightarrow +0.28$
Qwen2.5-Omni	$+3.76 \rightarrow +1.03$	$+3.73 \rightarrow +1.19$	$\mathbf{+0.08 \rightarrow -0.07}$	$+3.76 \rightarrow +1.42$	$+3.73 \rightarrow +0.05$	$+0.08 \rightarrow +1.12$
SQ-LLM	$+1.02 \rightarrow +0.97$	$\mathbf{-0.06 \rightarrow -0.02}$	$\mathbf{+1.29 \rightarrow +0.98}$	$\mathbf{+1.02 \rightarrow +1.94}$	$\mathbf{-0.06 \rightarrow -0.03}$	$+1.29 \rightarrow +1.63$
SALMONN-SQA	$\mathbf{+0.39 \rightarrow +0.44}$	$-1.12 \rightarrow -1.25$	$+1.44 \rightarrow +1.56$	$\mathbf{+0.39 \rightarrow +1.25}$	$\mathbf{-1.12 \rightarrow -1.03}$	$+1.44 \rightarrow +1.72$
QualiSpeech	$+1.82 \rightarrow +1.22$	$-0.83 \rightarrow -1.18$	$+1.99 \rightarrow +1.99$	n/a†		

from 3.76 to 1.03 while its *audio* power stays at -0.07, so that judge follows the claim less without listening more and the instruction barely moves SQ-LLM or SALMONN-SQA. Describe-then-rate lowers the bias and raises the guardrail for Qwen2.5-Omni and Qwen2-Audio, yet the same instruction nearly doubles the text power of SQ-LLM and triples that of SALMONN-SQA. QualiSpeech is not reported under that instruction. It is a descriptive model, so it often answers with the description alone and states no number, leaving no score in 38% of replies. That share depends on the condition, reaching 50% under the studio claim on clean audio against 29% under the phone claim, so what remains is not a fair sample across claims. These runs also phrase the rating instruction slightly differently from Section 5.3, so a judge's baseline row does not match Table 6. Appendix F shows the same sensitivity to wording.

J REFERENCE GROUNDING

Three judges let the audio outweigh the text and GRM is the only one of them built on Qwen2.5-Omni. It is also the only judge given a reference transcript and this section asks whether the two facts are connected. Table 17 asks whether the advantage comes from GRM's reference transcript, by taking that transcript away. From Table 18, we demonstrate writing a transcript into the prompt of a judge never trained with one does not help and it moves two of the five further from the audio.

Table 17 demonstrates whether the advantage comes from the transcript, by taking it away. GRM's prompt carries a "Target text" field, which holds the true transcript in every main experiment, and the table replaces it with "(transcript unavailable; judge the speech on its own audio merits)" while leaving the audio and the claim untouched. Withholding the transcript roughly halves the *text* power, from 0.75 to 0.37, but removes most of the *audio* power, from 1.89 to 0.32. The OVERRIDE therefore changes sign, from -1.14 to $+0.05$, and the text takes over the decision. Table 18 tests whether the same advantage transfers to judges never trained with a reference, by writing the transcript into their prompts. None of the five improves. The transcript lowers the *text* power for three judges and raises it for the other two, while the *audio* power falls for four. No judge combines a lower *text* power with a higher *audio* power, so none moves toward the audio. The transcript also pulls the OVERRIDE toward zero for the three judges whose audio was already winning, which weakens the audio rather than anchoring the judge to it. Therefore, a reference helps only a judge trained to use one.

Table 17: Reference grounding for SpeechJudge-GRM. Both columns run the same 2×2 design. With the reference, the target-text field of the prompt holds the true transcript of the clip, as in every main experiment. Without the reference, that field holds a fixed placeholder sentence for every clip. The upper block gives the mean score in each of the four cells. A positive OVERRIDE means the caption outweighs the audio.

Quantity	With reference	Without reference
clean audio, phone caption	5.21	1.39
clean audio, studio caption	5.78	1.86
noisy audio, phone caption	3.15	1.17
noisy audio, studio caption	4.07	1.44
text power	0.75	0.37
audio power	1.89	0.32
OVERRIDE	-1.14	$+0.05$

Table 18: The transcript written into the prompt of judges never trained with a reference. Each cell gives the baseline value before the transcript value. The baseline uses the instruction of Table 16, and the transcript condition places "The intended words are" and the clip's transcript before that instruction. A transcript should lower the *text* power and the OVERRIDE while keeping the *audio* power high. In bold are the cells that move the other way. All five judges use a 1 to 5 scale, and every reply in both conditions was parseable.

Judge	*Text* power ↓	OVERRIDE ↓	*Audio* power ↑
Qwen2-Audio	$\mathbf{+0.75 \rightarrow +1.15}$	$\mathbf{+0.44 \rightarrow +0.53}$	$+0.07 \rightarrow +0.29$
Qwen2.5-Omni	$+3.76 \rightarrow +3.43$	$+3.73 \rightarrow +3.36$	$\mathbf{+0.08 \rightarrow +0.00}$
SQ-LLM	$+1.02 \rightarrow +0.73$	$\mathbf{-0.06 \rightarrow +0.04}$	$\mathbf{+1.29 \rightarrow +1.16}$
SALMONN-SQA	$\mathbf{+0.39 \rightarrow +0.48}$	$\mathbf{-1.12 \rightarrow -0.94}$	$\mathbf{+1.44 \rightarrow +1.35}$
QualiSpeech	$+1.82 \rightarrow +1.39$	$\mathbf{-0.83 \rightarrow -0.54}$	$\mathbf{+1.99 \rightarrow +1.69}$