THE PERCEPTION-OVERRIDE REFLEX: WHEN SPEECH LANGUAGE MODELS FOLLOW PERCEPTION OVER INSTRUCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech language models have become increasingly capable of recognizing paralinguistic information such as gender, emotion, and accent from audio. However, we find that their performance degrades sharply when an affirmative instruction is replaced by its negated counterpart, with models frequently returning the option explicitly excluded by the instruction when that option is supported by the audio. We term this failure mode the **perception-override reflex**. To characterize it, we introduce **OVERRIDE**, a controlled benchmark that evaluates 4,204 audio clips under matched affirmative, negated, and text-control variants, allowing us to control for affirmative-task errors and text-side instruction execution. Across seven speech LLMs, negated instructions induce substantial accuracy drops even after restricting evaluation to items answered correctly under the corresponding affirmative instruction. We then ask what predicts variation in the reflex. Across 56 phrasings, text-side executability consistently predicts susceptibility under audio across six models, but does not fully account for the observed behavior: substantial reflex remains for highly executable instructions, and counterfactual formulations exhibit additional difficulty beyond that captured by text-side executability. These findings motivate on-policy self-distillation from the frozen model conditioned on the attribute in text and a reliably executable canonical phrasing. On Qwen2.5-Omni, this raises aggregate negated-instruction accuracy from 63.7% to 83.6% without degrading affirmative audio performance. Together, our results identify a reproducible failure in combining auditory evidence with negated instruction following, and show that text-side instruction competence provides both a predictor of vulnerability and a useful signal for mitigation.

1 INTRODUCTION

Play a clearly male voice to Qwen2.5-Omni and ask: *"The speaker's gender is NOT which—'male' or 'female'?"* The model answers *male*. This is not simply an affirmative-task error: when asked which gender the speaker is, the model answers *male* 98% of the time. Nor is the negated wording itself necessarily difficult: when the same attribute is stated explicitly in text, followed by the same negated question, the model answers *female* nearly perfectly. Yet when the attribute must be obtained from audio, the model often returns the option that the negated instruction asks it to exclude. We call this systematic failure mode the **perception-override reflex**.

To test whether this behavior is systematic across tasks, phrasings, and model families, we construct OVERRIDE, a controlled evaluation suite containing 4,204 two-option items across 40 perception and reasoning tiers derived from MMSU (WANG et al., 2026). Each item is evaluated under an affirmative instruction ($\mathrm{I_a}$), three fixed negated forms ($\mathrm{I_b}$–$\mathrm{I_d}$), and two held-out phrasings drawn from a bank of ten; every formulation has a matched text control in which the relevant attribute is stated explicitly rather than recovered from audio. Across seven speech LLMs (Qwen2.5-Omni, Qwen3-Omni, MiMo-Audio, Phi-4-MM, Qwen3.5-Omni, and Gemini 2.5 Pro and Flash), negated instructions cause substantial degradation even when evaluation is restricted to items answered correctly under the corresponding affirmative instruction (Table 1). Moreover, the resulting errors are structured: models disproportionately return the audio-supported option that the negated instruction asks them to exclude.

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.

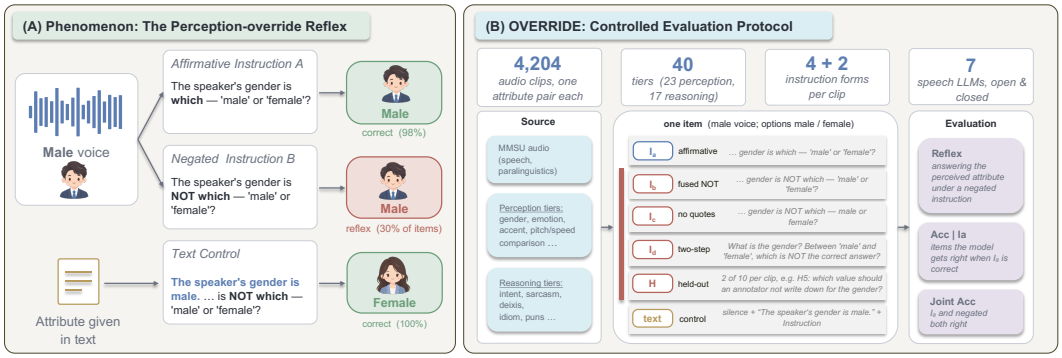

Figure 1: **(a)** The perception-override reflex: the same male voice is answered correctly under the affirmative instruction and when the attribute is stated in text, yet under the negated instruction with audio the model returns the option that the instruction asks it to exclude. **(b)** OVERRIDE evaluates 4,204 clips under four fixed instruction forms and two of ten held-out negated phrasings, each paired with a matched text-control version. The paired design controls for affirmative-task errors and text-side instruction execution.

The paired design lets us distinguish this behavior from several simpler explanations. First, restricting evaluation to affirmative-correct items removes failures that already occur under the corresponding non-negated task. Second, matched text controls are substantially more accurate than the audio-conditioned negated queries for the models that exhibit the reflex, and further restricting to items answered correctly under both the affirmative instruction and the text control leaves the effect largely intact. Thus, the reflex is not readily explained by affirmative-task errors or by an inability to execute the same instruction in text.

We next ask what predicts variation in the reflex (§3). Across 56 negated phrasings and six models, susceptibility under audio consistently tracks the model's ability to execute the same phrasing in text, with Pearson correlations ranging from -0.44 to -0.81. This relationship is not a complete account: substantial reflex remains even for highly executable instructions, and counterfactual formulations incur additional difficulty beyond that captured by text-side executability. Controlled minimal pairs further provide little evidence that the broader phrasing effect is explained by the tested syntactic choices alone. Together, these results identify text-side executability as a useful predictor of vulnerability while also revealing structured residual variation that it does not capture.

This asymmetry directly suggests a mitigation (§4). We use on-policy self-distillation in which the frozen model itself serves as the teacher, but under a privileged condition: the relevant attribute is supplied in text and the teacher is queried with a reliably executable canonical phrasing. The audio-conditioned model is trained on its own trajectories toward this teacher distribution. On Qwen2.5-Omni, the resulting model improves aggregate negated-instruction accuracy from 63.7% to 83.6%, without degrading affirmative audio performance. Teacher ablations further show that this choice is consequential: teachers that execute the privileged instruction reliably produce substantially better students, whereas poorly executable teacher formulations can degrade performance. The intervention therefore uses, rather than merely observes, the text-side asymmetry identified by our analysis.

Contributions. Our contributions are threefold: (i) we identify the perception-override reflex, a reproducible audio-conditioned failure of speech LLMs under negated instructions, and introduce OVERRIDE, a controlled evaluation protocol that separates it from affirmative-task errors and text-side instruction-execution failures across seven models; (ii) we show across 56 phrasings and six models that text-side executability consistently predicts susceptibility under audio, while substantial reflex remains for highly executable instructions and counterfactual formulations exhibit additional difficulty; and (iii) we develop an analysis-guided on-policy self-distillation intervention that uses the frozen model's more reliable text-conditioned behavior as privileged supervision, substantially improving negated-instruction following while preserving affirmative audio performance.

2 THE PERCEPTION-OVERRIDE REFLEX AND THE OVERRIDE EVALUATION PROTOCOL

Definition. Each item consists of an audio clip x, a ground-truth attribute value a^*, and a distractor value a'. Under an affirmative instruction, the correct answer is a^*; under its negated counterpart, the correct answer is a'. We define a *perception-override reflex* as a negated-instruction response equal to a^*, i.e., the model returns what is present in the audio rather than the answer required by the instruction. Unless otherwise noted, the *reflex rate* is computed over items that the model answers correctly under the corresponding affirmative instruction.

Dataset construction. OVERRIDE (Fig.1b) is derived from MMSU(WANG et al., 2026), which contains 47 speech tasks spanning perception and reasoning. We retain the 40 tasks that admit our controlled two-option transformation, excluding seven tasks for which this transformation is not well defined. Following the MMSU taxonomy, the resulting set contains 23 perception tasks (e.g., gender, emotion, and pitch comparison) and 17 reasoning tasks (e.g., intent, sarcasm, and deixis). For each example, we retain the correct option and one distractor, yielding 4,204 two-option items. This gives a uniform setting across tiers, including four inherently binary tasks, and, crucially, a unique negated target; with four options, excluding the audio-supported answer leaves three valid alternatives and raises chance negated accuracy to 75%. We therefore use two options as the primary setting; four-option and probability-based analyses in Appendix A confirm that the audio-conditioned preference persists beyond binary choice.

Instruction variants and controls. Each item is evaluated under six instruction forms: four fixed templates ($\mathrm{I_a}$–$\mathrm{I_d}$) and two additional held-out phrasings sampled from a bank of ten templates (H1–H10). For a gender-recognition example, $\mathrm{I_a}$is the affirmative form, *The speaker's gender is which—male or female?* The three fixed negated variants progressively alter the instruction: $\mathrm{I_b}$introduces an explicit *NOT*, $\mathrm{I_c}$removes quotation marks around the answer options, and $\mathrm{I_d}$uses a two-step formulation, *What is the speaker's gender? Between male and female, which is NOT the correct answer?*

The held-out bank covers substantially different linguistic frames, including fact-checking, proof-reading, betting, and counterfactual re-recording. Each item is assigned two of these ten phrasings, so that every held-out phrasing is evaluated on approximately 840 items. None of H1–H10 or $\mathrm{I_b}$–$\mathrm{I_d}$appears in the training data used later.

For every instruction form, we also construct a matched *text control* by replacing the audio with silence and explicitly stating the relevant attribute in text, e.g., *The speaker's gender is male.* This isolates instruction execution from audio perception.

Evaluation protocol. Raw negated-instruction accuracy can be misleading because it conflates instruction following with perceptual errors. We therefore report negated accuracy (Acc$|\mathrm{I_a}$) and reflex rate primarily on each model's own $\mathrm{I_a}$-correct subset, so that the analysis focuses on items the model can answer correctly under the affirmative instruction. As a complementary unfiltered metric, we also report *joint accuracy*, defined as the fraction of all items answered correctly under both $\mathrm{I_a}$ and the corresponding negated instruction.

The reflex is widespread across models and phrasings. We evaluate seven speech LLMs, including Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), MiMo-Audio (Team et al., 2025), Phi-4-MM (Microsoft et al., 2025), Qwen3.5-Omni (Team, 2026), and Gemini 2.5 Pro and Flash (Comanici et al., 2025). Table 1 shows a consistent degradation under negated instructions. On items answered correctly under $\mathrm{I_a}$, negated accuracy ranges from 33.9% to 86.5% on the fixed forms and from 36.1% to 83.2% on the held-out phrasings, whereas the corresponding text controls reach 77.7–94.7% for all except Phi-4-MM. This gap suggests that the failure is not explained by affirmative perceptual errors alone or by an inability to execute the same instruction in text.

The errors are strongly structured: they show a consistent tendency to return the audio-supported option that the negated instruction asks the model to exclude. Under the two-option evaluation used here, the reflex rate closely tracks a substantial portion of the accuracy loss. For Qwen2.5-Omni, for example, accuracy on $\mathrm{I_b}$–$\mathrm{I_d}$falls to 66.6%, a 33.4-point loss relative to its $\mathrm{I_a}$-correct subset, while

Table 1: **Performance of seven speech LLMs on OVERRIDE.** I_a denotes affirmative accuracy. *Acc*$|I_a$ and *Reflex* are computed on each model's I_a-correct subset, while *Joint* is the unfiltered fraction of items answered correctly under both I_a and the corresponding negated instruction. The first block averages the three fixed negated forms I_b–I_d, the second averages the ten held-out phrasings H1–H10, and the final block aggregates all negated queries; its *Text ctrl* column is the accuracy on the same queries when the audio is replaced by silence and the attribute is stated in text. Arrows indicate better performance.

		I_b–I_d			H1–H10			All negated			
Model	I_a↑	Acc$\|I_a$↑	Reflex↓	Joint↑	Acc$\|I_a$↑	Reflex↓	Joint↑	Acc$\|I_a$↑	Reflex↓	Joint↑	Text ctrl↑
Open-weight models											
Qwen2.5-Omni-7B	71.3	66.6	30.3	47.5	59.4	32.4	42.3	63.7	31.1	45.4	91.5
Qwen3-Omni-30B-A3B	77.0	63.5	34.7	48.9	64.0	32.1	49.3	63.7	33.7	49.1	86.3
MiMo-Audio-7B	70.7	65.5	25.9	46.4	54.0	32.0	38.2	60.9	28.3	43.1	77.7
Phi-4-MM	61.4	33.9	60.7	20.8	36.1	47.4	22.1	34.7	55.4	21.3	37.5
Proprietary models											
Qwen3.5-Omni	85.2	86.5	11.8	73.7	76.0	18.1	64.8	82.3	14.3	70.2	93.3
Gemini 2.5 Pro	83.3	81.9	17.1	68.2	83.2	11.6	69.3	82.5	14.9	68.7	92.9
Gemini 2.5 Flash	78.7	73.8	20.8	58.1	76.8	15.5	60.4	75.1	18.6	59.1	94.7

the reflex rate is 30.3%. A similar pattern appears on the held-out phrasings, suggesting that the effect is not restricted to the fixed benchmark templates. Proprietary models exhibit lower reflex rates overall, but the pattern remains present. Phi-4-MM is a boundary case: with a text-control accuracy of only 37.5%, its negated errors cannot be cleanly separated from failures to execute the instruction itself.

Takeaway 1. Across model families and negated phrasings, speech LLMs show an audio-conditioned accuracy drop together with a consistent tendency to return the audio-supported option that the negated instruction asks them to exclude; the same instructions are substantially easier when the relevant attribute is provided in text.

3 WHAT SHAPES THE REFLEX?

Section 2 established a consistent audio-conditioned failure under negated instructions. We now ask what factors predict variation in the reflex. We first test several simple explanations for the observed failures (§3.1). We then examine variation across instruction phrasings and show that text-side executability consistently predicts susceptibility to the reflex across models (§3.2). Finally, we characterize the limits of this relationship, identifying systematic residual variation that text-side executability alone does not capture (§3.3). Unless otherwise noted, analyses are restricted to items answered correctly under the corresponding affirmative instruction I_a.

3.1 SIMPLE EXPLANATIONS ARE INSUFFICIENT

Affirmative-task errors. All reflex rates are computed on each model's I_a-correct subset. The observed degradation under negation therefore persists even after excluding items that the model fails on under the corresponding affirmative instruction.

Text-side instruction failure. Qwen2.5-Omni achieves 91.5% accuracy on the text controls, and the other models that exhibit the reflex reach 78–95%, with Phi-4-MM as the clear exception. More stringently, restricting evaluation to items answered correctly under both I_a and the matched text control changes the I_b reflex rate only modestly (30.2% to 29.1% for Qwen2.5-Omni, and by less than six points for the other applicable models). Thus, for most models, the reflex persists on items for which the same instruction can be executed correctly when the relevant attribute is supplied in text. Phi-4-MM, whose text-control accuracy is only 37.5%, does not satisfy this condition and is therefore treated separately.

Negation token inattention. We find no evidence for the simplest account in which reflexive responses merely fail to attend to the negation token. On I_b, attention from the answer position to *NOT* is nearly identical for correct and reflexive responses (0.0137 vs. 0.0142 on 2,953 I_a-correct items), with similar profiles across layers and for the remaining instruction tokens (Appendix B).

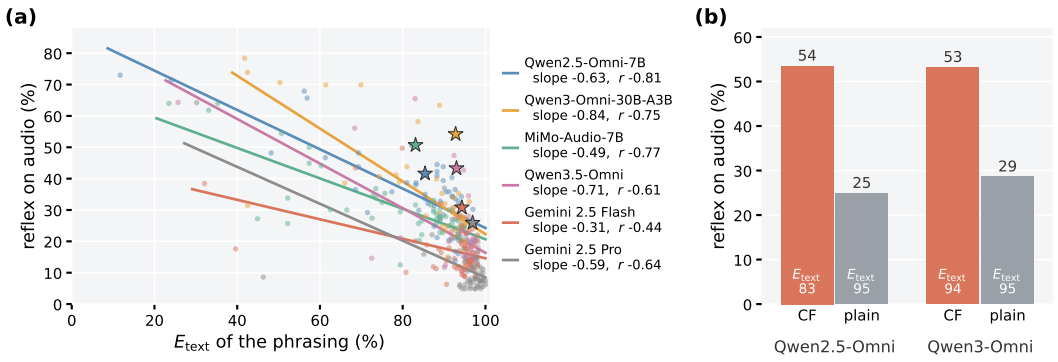

Figure 2: **Text-side executability and residual phrasing effects. (a)** Audio reflex rate versus text-control accuracy E_{text} for 56 phrasings in each model. Lines show least-squares fits; stars denote the counterfactual held-out phrasing H2. **(b)** Reflex rates for five counterfactual (CF) and five plain-negation phrasings evaluated on identical items; numbers inside bars report mean E_{text}.

This does not establish that the negation is used equivalently in the two cases, but argues against the token simply being overlooked.

3.2 Phrasing sensitivity tracks text-side executability

Section 3.1 shows that the reflex persists even when the same instruction can be executed correctly in text. We next ask whether variation in text-side executability nevertheless predicts which phrasings are most susceptible to the reflex. For a phrasing p, we define $E_{\text{text}}(p)$ as its accuracy in the matched text control and $R_{\text{audio}}(p)$ as its reflex rate under audio, both on $\mathrm{I_a}$-correct items.

To obtain sufficient variation in phrasing, we evaluate 56 negated instructions: the three fixed forms $\mathrm{I_b}$–$\mathrm{I_d}$, ten held-out phrasings H1–H10, 25 free-form paraphrases, and 18 phrasings forming nine controlled minimal pairs (Appendix C). The three fixed forms are evaluated on the full benchmark; the remaining 53 phrasings are evaluated on the same fixed subset of 200 clips, five per tier, for every model.

Text-side executability consistently predicts phrasing sensitivity. Figure 2a shows a negative relationship between E_{text} and R_{audio} across all six models included in this analysis.[1] Least-squares slopes range from -0.31 to -0.84, with Pearson correlations from -0.44 to -0.81. For Qwen2.5-Omni, the phrasing with the lowest text-control accuracy ($E_{\text{text}} = 12\%$) produces a 73% reflex rate under audio. More generally, phrasings that are less reliable in text tend to yield higher reflex rates under audio.

This relationship is strongest when text-side difficulty varies substantially. For the three proprietary models, E_{text} exceeds 90% on most phrasings, so the fitted trend is driven mainly by the smaller subset of phrasings on which text execution degrades; correspondingly, Spearman correlations are weaker ($\rho = -0.20$ to -0.30) than for the three open-weight models ($\rho = -0.68$ to -0.77). The fitted relationship nevertheless has the same direction across all six models.

The tested syntactic distinctions do not explain the effect. One possible explanation is that the reflex is specifically triggered when negation is attached to the act of hearing, as in *the gender you hear*. We test this with eight minimal pairs differing only in this attachment and a ninth pair contrasting explicit *NOT* with lexical *wrong*. Across the six models, the mean within-pair effect of perceptual attachment ranges from -1.3 to $+1.8$ points, with no significant difference in any model (Wilcoxon $p \geq 0.12$). The *NOT–wrong* contrast likewise shows no consistent direction. These controlled comparisons provide little evidence that either tested syntactic choice accounts for the broader phrasing effect.

[1] Phi-4-MM is omitted because its text-control accuracy is only 37.5% (Table 1), making phrasing-level text executability difficult to interpret.

3.3 TEXT-SIDE EXECUTABILITY IS NOT A COMPLETE ACCOUNT

The relationship in §3.2 is substantial but incomplete. Even among phrasings that the models execute reliably in text ($E_{\text{text}} \geq 90\%$), non-trivial reflex rates remain: $26.9 \pm 7.6\%$ for Qwen2.5-Omni, $25.6 \pm 7.2\%$ for Qwen3-Omni, $24.1 \pm 3.5\%$ for MiMo-Audio, $19.2 \pm 8.0\%$ for Qwen3.5-Omni, $15.6 \pm 5.3\%$ for Gemini 2.5 Flash, and $9.6 \pm 5.7\%$ for Gemini 2.5 Pro. Thus, high text-side executability reduces susceptibility but does not eliminate the audio-conditioned failure.

Counterfactual framing contributes an additional penalty. One source of residual variation is counterfactual framing. H2, which asks about a counterfactually re-recorded version of the clip, lies above the fitted relationship in all six models despite relatively high text-control accuracy. We therefore compare five counterfactual phrasings with five plain negations that request the same output on the same 160 items, scoring each model on the items it answers correctly under $\mathrm{I_a}$ (107 for Qwen2.5-Omni, 133 for Qwen3-Omni; Fig. 2b). Counterfactual prompts produce higher reflex rates by 28.6 points for Qwen2.5-Omni and 24.7 points for Qwen3-Omni (paired Wilcoxon $p < 10^{-9}$ in both cases). For Qwen3-Omni, the two groups differ by less than one point in E_{text}; for Qwen2.5-Omni, the 12-point text-control gap accounts for only part of the observed reflex increase under the fitted relationship. Counterfactual framing is therefore associated with additional difficulty beyond what is captured by text-side executability alone. We do not attempt to identify its mechanism here.

Additional analyses. We also examine whether the reflex varies systematically across MMSU task categories. Perception tiers show somewhat higher reflex rates than reasoning tiers across all six models, but the difference is small for five of them; we therefore treat this as a secondary observation and report the full tier-level analysis in Appendix D. As a complementary internal diagnostic, we examine intermediate answer readouts and find that the audio–text discrepancy becomes visible primarily in later layers. Because such projections do not localize the underlying computation, we report this analysis only in Appendix E.

Takeaway 2. Across models, text-side executability consistently predicts which negated phrasings are most susceptible to the reflex, but it does not fully determine their behavior: substantial reflex remains for highly executable instructions, and counterfactual framing introduces additional difficulty. This motivates using the model's more reliable text-conditioned behavior as a training signal in the next section.

4 MITIGATION: TEXT-PRIVILEGED SELF-DISTILLATION WITH A CANONICAL TEACHER

The analysis above suggests a natural mitigation strategy. Since the same model follows negated instructions more reliably when the relevant attribute is supplied in text, and since phrasing executability strongly predicts vulnerability under audio, we use the model's own reliable text-conditioned behavior as privileged supervision for its audio-conditioned behavior. We implement this idea with on-policy self-distillation and a canonical teacher phrasing.

Method. We adopt on-policy self-distillation (Agarwal et al., 2024; Zhao et al., 2026), with the frozen base model π_{θ_0} serving as its own teacher. For each training item, the student first generates a response y from its current policy. Let $c_s = (x, q)$ denote the audio-conditioned student context and $c_t = (x_{\text{sil}}, t(a^*), q_c)$ the privileged teacher context, where $t(a^*)$ states the ground-truth attribute in text. Along the resulting student-generated trajectory, we minimize the token-level reverse KL

$$\mathcal{L}_{\text{OPSD}} = \frac{1}{|y|} \sum_{i=1}^{|y|} D_{\text{KL}}(\pi_\theta(\cdot \mid c_s, y_{<i}) \,\|\, \pi_{\theta_0}(\cdot \mid c_t, y_{<i})) \,.$$

We additionally mix affirmative and general retention examples, trained with the same objective against the frozen base model, to preserve its existing capabilities. Training uses 25 negated phrasings disjoint from both the fixed evaluation forms $\mathrm{I_b}$–$\mathrm{I_d}$ and the held-out phrasings H1–H10, and the

Table 2: Mitigation results on both model families. Acc|I_a, Reflex and Joint as in Table 1; the last block pools all negated queries (three benchmark forms and two held-out phrasings per item) and adds the text control. Arrows indicate better performance. Qwen2.5-Omni OPSD-text results are averaged over three seeds (s.d. ≤ 0.5 points across reported metrics).

		I_b–I_d			held-out H1–H10			all negated queries			
	I_a↑	Acc|I_a↑	Reflex↓	Joint↑	Acc|I_a↑	Reflex↓	Joint↑	Acc|I_a↑	Reflex↓	Joint↑	Text ctrl↑
Qwen2.5-Omni-7B											
Base	71.3	66.6	30.3	47.5	59.4	32.4	42.3	63.7	31.1	45.4	91.5
Verbalize-then-answer prompt	71.3	57.4	41.8	40.9	55.6	38.7	39.6	56.7	40.6	40.4	86.4
SFT	70.1	77.7	21.1	54.4	72.2	25.3	50.6	75.5	22.8	52.9	92.3
OPSD-audio	72.9	79.7	18.6	58.1	69.7	26.3	50.7	75.7	21.7	55.1	94.4
OPSD-text w/o canonical phrasing	**73.3**	81.5	16.4	59.7	73.4	22.6	53.8	78.3	18.9	57.3	95.6
OPSD-text (Ours)	72.8	**86.6**	**12.3**	**63.1**	**79.0**	**17.8**	**57.6**	**83.6**	**14.5**	**60.9**	**97.2**
Qwen3-Omni-30B-A3B											
Base	77.0	63.5	34.7	48.9	64.0	32.1	49.3	63.7	33.7	49.1	86.3
Verbalize-then-answer prompt	77.0	59.2	40.4	45.6	69.9	28.1	53.8	63.5	35.5	48.9	83.1
SFT	65.6	67.6	32.1	44.3	67.7	29.9	44.4	67.6	31.2	44.3	77.3
OPSD-audio	**77.9**	74.7	25.0	58.2	75.1	21.6	58.5	74.9	23.6	58.3	92.6
OPSD-text (Ours)	77.6	**82.6**	**17.0**	**64.0**	**77.7**	**19.7**	**60.2**	**80.6**	**18.1**	**62.5**	**97.0**

audio encoder remains frozen; full training details are given in Appendix F. The distillation objective itself is standard; our design choice is the teacher condition: it combines privileged text access to the attribute with a phrasing that the frozen model can execute reliably, directly following the asymmetry identified in §3.

Baselines and variants. **SFT**: supervised fine-tuning on the same training examples using the gold option as the target. **Verbalize-then-answer**: no training; the base model is prompted to produce exactly two text lines, first stating the attribute it hears and then answering the negated question. This is a fixed-format verbalization baseline rather than free-form chain-of-thought reasoning. **OPSD-audio**: the teacher receives the original audio together with a textual statement of the ground-truth attribute and is queried using the student's phrasing. **OPSD-text w/o canonical phrasing**: the teacher receives silence plus a textual statement of the ground-truth attribute and is queried using the student's phrasing. **OPSD-text**: the full method, which uses the same text-informed teacher but replaces the student's phrasing with a canonical phrasing selected for high text-side executability.

4.1 Main results

Prompting alone does not resolve the reflex. A verbalize-then-answer prompt, which asks the model to state the perceived attribute before answering the negated question, does not improve overall performance (Table 2). Pooled negated accuracy decreases from 63.7% to 56.7% on Qwen2.5-Omni and remains essentially unchanged on Qwen3-Omni (63.7% to 63.5%). Explicitly eliciting the perceived attribute is therefore insufficient to reproduce the gains obtained through training.

Text-privileged self-distillation substantially reduces the reflex. Table 2 reports the main mitigation results. On Qwen2.5-Omni, OPSD-text raises pooled negated-instruction accuracy from 63.7% to 83.6%, reduces the reflex rate from 31.1% to 14.5%, and increases unfiltered joint accuracy from 45.4% to 60.9%. The improvement transfers to held-out phrasings (59.4% to 79.0%) and does not trade off affirmative performance (71.3% to 72.8%). The same qualitative pattern holds for Qwen3-Omni: pooled negated accuracy improves from 63.7% to 80.6%, reflex rate falls from 33.7% to 18.1%, and joint accuracy rises from 49.1% to 62.5%. In both model families, the text control also improves after training, indicating that the intervention does not achieve its gains by weakening text-side instruction following.

The teacher condition matters. Teacher ablations support both components of the design: on Qwen2.5-Omni, pooled accuracy rises from 75.7% with an audio-conditioned teacher to 78.3% with text conditioning and 83.6% with canonical phrasing; Qwen3-Omni likewise improves from 74.9% to 80.6% with OPSD-text.

Table 3: **Effect of teacher-phrasing executability on Qwen2.5-Omni.** Only the canonical phrasing presented to the teacher is varied. E_{text} denotes the base model's text-control accuracy on that phrasing (Appendix C); evaluation metrics follow Table 2. Results are from a single seed.

Teacher phrasing	E_{text}	I_a	Acc|I_a (I_b–I_d)	Reflex	Acc|I_a (held-out)	Text ctrl
T1 *"does the audio not support"* (ours)	93	72.9	86.2	12.4	79.6	97.9
T9 *"name the incorrect option"*	96	73.1	**88.0**	**11.0**	**81.0**	**98.4**
T18 *"skip whichever you actually hear"*	69	72.8	71.4	26.8	63.4	88.0
T3 *"I'm after what you did NOT hear"*	12	72.7	42.1	57.3	31.9	66.2
Base (no training)	–	71.3	66.6	30.3	59.4	94.7

Distillation outperforms direct supervised fine-tuning. SFT improves over the base model but remains below OPSD-text on both Qwen2.5-Omni (75.5% vs. 83.6%) and Qwen3-Omni (67.6% vs. 80.6%). OPSD-text also preserves affirmative performance, whereas SFT reduces it.

Canonical distillation removes most phrasing sensitivity. Section 3.2 showed that the base model's reflex rate strongly depends on text-side phrasing difficulty. After OPSD-text, this relationship nearly disappears: across the same 56 phrasings, the fitted slope changes from negative in the base model to $+0.01$ ($r = 0.04$), and reflex rates concentrate around $18.0 \pm 4.3\%$. By contrast, OPSD-audio and OPSD-text without canonicalization retain substantial phrasing dependence (slopes -0.79 and -0.64). This is consistent with canonicalizing the teacher condition substantially reducing the phrasing sensitivity identified in our analysis.

4.2 Teacher executability predicts student quality

Table 3 varies only the phrasing the teacher reads. Two phrasings the base model executes almost perfectly in text (93 and 96) give similar student performance (86.2 and 88.0). A phrasing it executes 69% of the time yields a student only 5 points above the base model, whereas one it executes 12% of the time yields a student 24 points *below* the base model, with text-control accuracy falling from 94.7 to 66.2. On the phrasing bank, all four students are nearly flat against base-model difficulty (slopes $+0.01$, $+0.03$, -0.06, -0.15) but at markedly different levels (18.0, 16.3, 29.9, and 65.0). Teacher-side executability therefore strongly predicts student quality, consistent with §3.2.

4.3 Capability retention

Finally, we evaluate whether training preserves capabilities outside the negated-instruction objective. On the original four-choice MMSU evaluation and LibriSpeech test-clean, all distilled variants remain close to the base model, with no evident degradation on either evaluation (Table 4). SFT behaves differently: it reduces MMSU reasoning accuracy on Qwen2.5-Omni (74.5%→71.9%) and shows a lower LibriSpeech WER, but the latter should not be interpreted as a general ASR gain because the SFT retention set includes LibriSpeech dev-clean transcripts.

5 Related Work

Speech LLM evaluation. Recent benchmarks evaluate increasingly broad capabilities of audio-language models, including general audio understanding and instruction following in AIR-Bench (Yang et al., 2024), diverse speech tasks in Dynamic-SUPERB (Huang et al., 2024), voice-assistant robustness in VoiceBench (Chen et al., 2026), and fine-grained spoken-language perception and reasoning in MMSU (WANG et al., 2026). MMSU is particularly relevant to our setting: it evaluates 47 perception and reasoning tasks, of which we retain the 40 that admit controlled two-option transformations. These benchmarks primarily measure whether a model can answer a question from audio under a single instruction formulation. In contrast, OVERRIDE evaluates the same audio content under matched affirmative, negated, and text-informed formulations, allowing us to distinguish failures already present under the affirmative task from those that emerge under negated instruction following.

Table 4: **Capability retention on Qwen2.5-Omni.** Accuracy on the original four-choice MMSU evaluation (4,981 items) and word error rate on LibriSpeech test-clean (Panayotov et al., 2015).

Method	MMSU all	Perception	Reasoning	WER (%)
Base	58.5	43.4	74.5	2.45
OPSD-audio	59.4	43.8	75.9	2.47
OPSD-text (no canon.)	59.4	43.8	**76.0**	2.53
OPSD-text	**59.7**	44.9	75.4	2.53
SFT	58.4	**45.5**	71.9	1.65

Negation and cross-modal conflicts. Negation remains challenging for both conventional language models and LLMs (Hossain et al., 2022; Truong et al., 2023; Kim et al., 2025). Kim et al. (Kim et al., 2025) study *negation blindness* in text-only LLMs, where models that answer a positive query correctly may nevertheless repeat the same answer after the query is negated. Our setting instead examines how negated instruction following interacts with the modality through which the relevant evidence is provided: the same negated instructions are generally more reliably executed when the target attribute is supplied in text than when it must be recovered from audio. NegBench (Alhamoud et al., 2025) similarly studies negation in multimodal models, showing substantial failures on negated image/video retrieval and multiple-choice queries. OVERRIDE instead focuses on matched within-item contrasts among affirmative, negated, and text-informed speech queries, and characterizes a systematic tendency to return the audio-supported option that the negated instruction asks the model to exclude. This behavior is also related to modality–language conflicts studied in vision: VQA models can exploit language priors rather than visual evidence (Goyal et al., 2017), while large vision-language models may hallucinate objects unsupported by the image (Li et al., 2023). The perception-override reflex concerns a complementary interaction: auditory evidence remains influential even when the instruction requires the audio-supported option to be rejected.

On-policy distillation. Our mitigation builds on recent on-policy distillation and self-distillation methods, which train a student on its own generated trajectories while using a stronger or privileged teacher signal (Agarwal et al., 2024; Zhao et al., 2026). We use the same general machinery in a different setting: the teacher is the frozen model itself under a text-conditioned formulation that our analysis identifies as more reliably executable than the corresponding audio-conditioned instruction.

6 Limitations and Conclusion

Limitations. Our main evaluation uses two-option questions so that each negated instruction has a unique target answer and the same protocol can be applied across heterogeneous MMSU tasks. Although four-option and probability-based analyses show that the audio-conditioned preference persists beyond this binary setting, broader evaluation with naturally multi-option and open-ended negated questions remains important. Our analyses are behavioral and diagnostic rather than mechanistic: attention and layer-wise readouts constrain several simple explanations but do not identify the causal origin of the reflex. Finally, the mitigation experiments are limited to two trainable Qwen model families; extending the intervention to additional architectures and training recipes is left for future work.

Conclusion. We identify a systematic failure in speech LLMs when auditory evidence must be combined with negated instructions: even on items they answer correctly under the corresponding affirmative question, models often return the audio-supported option that the instruction asks them to exclude. Matched text controls show that this behavior cannot be explained by affirmative-task errors or text-side instruction failure alone, while phrasing-level analyses reveal that text-side executability is a strong but incomplete predictor of susceptibility under audio. Using this asymmetry as supervision, text-privileged on-policy self-distillation substantially improves negated-instruction following across two model families while preserving affirmative performance. Understanding the causal origin of the remaining audio-conditioned failures remains an open problem.

AI USE STATEMENT

Generative AI tools were used for code assistance, data-analysis scripts, and manuscript editing. All reported results were independently verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This work evaluates existing speech language models on public benchmark data and does not involve human-subject experiments or the collection of new personal data. We do not identify additional ethical concerns beyond those associated with the underlying models and datasets.

REPRODUCIBILITY STATEMENT

The benchmark construction, evaluation protocol, instruction variants, model configurations, and training details are described in the main paper and appendices. The full phrasing bank, per-tier analyses, and optimization settings are provided in the appendix to support reproduction of the reported results.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622, 2025. doi: 10.1109/CVPR52734.2025.02757.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 04 2026. ISSN 2307-387X. doi: 10.1162/TACL.a.628. URL https://doi.org/10.1162/TACL.a.628.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. An analysis of negation in natural language understanding corpora. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 716–723, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.81. URL https://aclanthology.org/2022.acl-short.81/.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-Yi Lee. Dynamic-superb: Towards a dynamic,

collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12136–12140, 2024. doi: 10.1109/ICASSP48485.2024.10448257.

Jinsung Kim, Seonmin Koo, and Heuiseok Lim. Semantic inversion, identical replies: Revisiting negation blindness in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21434–21471, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1088. URL https://aclanthology.org/2025.emnlp-main.1088/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, Heyu Chen, Jiawei Li, Lei Li, Menghang Zhu, Peidian Li, Qiying Wang, Sirui Deng, Weimin Xiong, Wenshan Huang, Wenyu Yang, Yilin Jiang, Yixin Yang, Yuanyuan Tian, Yue Ma, Yue Yu, Zihan Zhang, Zihao Yue, Bangjun Xiao, Bingquan Xia, Bofei Gao, Bowen Ye, Can Cai, Chang Liu, Chenhong He, Chunan Li, Dawei Zhu, Duo Zhang, Fengyuan Shi, Guoan Wang, Hailin Zhang, Hanglong Lv, Hanyu Li, Hao Tian, Heng Qu, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianguang Zuo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Linghao Zhang, Meng Chen, Nuo Chen, Peng Zhang, Qianli Chen, Qiantong Wang, Rang Li, Shaohui Liu, Shengfan Wang, Shicheng Li, Shihua Yu, Shijie Cao, Shimao Chen, Shuhao Gu, Weikun Wang, Wenhan Ma, Xiangwei Deng, Xing Yong, Xing Zhang, Xu Wang, Yifan Song, Yihao Zhao, Yingbo Zhao, Yizhao Gao, Yu Cheng, Yu Tu, Yudong Wang, Zhaojun Huang, Zhengju Tang, Zhenru Lin, Zhichao Song, Zhipeng Xu, Zhixian Zheng, and Zihan Jiang. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://arxiv.org/abs/2512.23808.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In Alexis Palmer and Jose Camacho-collados (eds.), *Proceedings of the 12th Joint Conference on Lexical and*

*Computational Semantics (*SEM 2023)*, pp. 101–114, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.10. URL https://aclanthology.org/2023.starsem-1.10/.

Dingdong WANG, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/34e278fbbd7d6d7d788c98065988e1a9-Paper-Conference.pdf.

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.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

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, 2026. URL https://arxiv.org/abs/2601.18734.

A BEYOND TWO OPTIONS

Four-option evaluation. For 27 tiers where four valid answer options can be constructed, we add two same-tier distractors (40 items per tier). Under uniform random choice, the excluded audio-supported option is selected with probability 25%, compared with 50% in the two-option setting. Table 5 reports its selection rate on $\mathrm{I_a}$-correct items. The absolute rate decreases from 31.5% to 16.9% with four options, while the matched text control remains near zero; OPSD-text further reduces the rate in both settings. We report this comparison for Qwen2.5-Omni because Qwen3-Omni reaches only 54.5% text-control accuracy in the four-option setting, which would confound this comparison.

Table 5: Four-option variant, Qwen2.5-Omni, 27 tiers. *Excluded-option rate*: share of $\mathrm{I_a}$-correct items on which the model returns the option the negated instruction excludes; chance is 25% with four options and 50% with two. The two-option column is computed on the same items.

	$\mathrm{I_a}$ accuracy			excluded-option rate		
	audio	text	n	four options (audio / text)	two options (audio)	ratio to chance (4 / 2)
Base	50.4	95.7	544	16.9 / 1.7	31.5	0.68 / 0.63
OPSD-text	51.3	96.1	554	4.3 / 1.1	8.4	0.17 / 0.17

Probability assigned to the excluded option. We additionally compare the output probability assigned to the excluded audio-supported option with that assigned to the negated target, avoiding dependence on discrete choice. On matched items, the excluded option receives substantially more probability under audio than in the text control (Table 6); under $\mathrm{I_b}$, the corresponding values are 0.40 vs. 0.08 for Qwen2.5-Omni and 0.26 vs. 0.04 for Qwen3-Omni. The gap remains even on items answered correctly by the audio-conditioned run. For T3, where the text control itself fails, the modality gap largely disappears.

Table 6: Mean probability of the excluded option at the answer position (two-way normalized against the correct option), audio input vs. matched text control on identical items.

		all items		items the audio run answers correctly	
Model	instruction	audio	text	audio	text
Qwen2.5-Omni	$\mathrm{I_b}$	0.40	0.08	0.30	0.07
Qwen2.5-Omni	T9 (easy)	0.28	0.09	0.22	0.08
Qwen2.5-Omni	T3 (hard)	0.71	0.78	0.69	0.74
Qwen3-Omni	$\mathrm{I_b}$	0.26	0.04	0.13	0.03
Qwen3-Omni	T9 (easy)	0.19	0.01	0.11	0.01

B ATTENTION TO THE NEGATION TOKEN

For each item, we take the attention weights from the position that generates the first answer token, average them across heads, and then sum them over each input span and divide by the number of tokens in that span, so that spans of different lengths are comparable. As shown in Fig. 3, attention to the negation token is nearly indistinguishable between correct and reflexive responses across layers (all-layer means 0.0137 and 0.0142, respectively). The same is true for the remaining instruction tokens. Attention to audio tokens at the answer position is substantially smaller on a per-token basis (below 10^{-3}).

These results provide no evidence that reflexive responses arise because the negation token is simply unattended. We emphasize, however, that attention weights are only a coarse diagnostic: similar attention does not imply that the attended information is represented or used in the same way. We therefore treat this analysis only as a control for the simplest inattention hypothesis; Appendix E gives a complementary layer-wise readout.

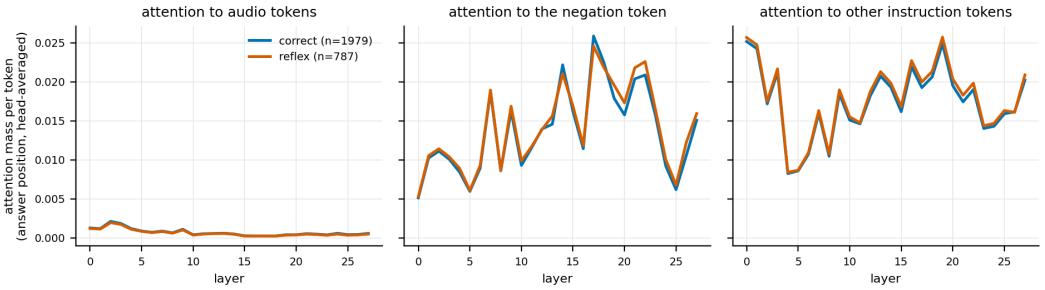

Figure 3: Per-token attention at the answer position to the audio tokens, the negation token and the remaining instruction tokens, by layer (Qwen2.5-Omni, $\mathrm{I_b}$ with audio, 2,953 $\mathrm{I_a}$-correct items; shared y-axis). Blue: responses that answer correctly ($n = 1{,}979$); orange: responses that return the perceived attribute ($n = 787$).

C THE PHRASING BANK

Table 7 lists every negated instruction used in §3.2: the three benchmark forms $\mathrm{I_b}$–$\mathrm{I_d}$, the ten held-out phrasings H1–H10, the 25 free-form paraphrases T01–T25 (later used as training phrasings in §4), and the nine minimal pairs MP1–MP9. In each minimal pair the two versions differ only in whether the negation is attached to the act of hearing (P, "the *attr* you hear") or to the label (L); MP9 contrasts explicit *NOT* (N) with lexical *wrong* (W). *attr* is replaced by the tier's attribute name and a, b by the two options. For each phrasing we give, for the two open base models used throughout, the text-control accuracy E_{text} and the audio reflex rate R_{audio} on $\mathrm{I_a}$-correct items, on the item sets described in §3.2.

Table 7: The 56 negated instructions and their per-phrasing statistics. Columns: E_{text} / R_{audio} for Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B.

		Qwen2.5-Omni		Qwen3-Omni	
ID	Instruction	E_{text}	R_{audio}	E_{text}	R_{audio}
Benchmark forms (full set)					
Ib	The *attr* is NOT which — 'a' or 'b'? Answer with just that option.	97	30	79	39
Ic	The *attr* is NOT which — a or b? Answer with just that option.	92	40	74	41
Id	What is the *attr*? Between 'a' and 'b', which is NOT the correct answer? Answer with just that option.	95	21	94	24
Held-out phrasings					
H1	Suppose exactly one of a and b is wrong about the *attr* — which one is it? Give only that option.	69	44	86	42
H2	If someone re-recorded this clip and deliberately changed the *attr*, which of a and b would the new version have? Reply with only that option.	85	42	93	54
H3	Act as a fact-checker for the *attr* in this clip. Two claims were filed, a and b. Which claim do you mark false? Reply with only that value.	72	29	89	29
H4	You are proofreading audio descriptions. For the *attr* in this clip, spot the wrong value in this pair: a, b. Give only that value.	90	39	90	24
H5	A junior annotator asks which value they should not write down for the *attr* in this clip. Their options are a and b. What do you tell them? Reply with only that value.	93	21	93	22
H6	Which of a and b would you never use to describe the *attr* here? Reply with only that option.	90	36	95	28
H7	If I were describing the *attr* in this clip, which of a and b should I avoid saying? Give only that option.	89	27	93	24
H8	You have to bet against one of these two values for the *attr*: a or b. Which do you bet against? Reply with only that value.	77	36	96	24
H9	Something is off about one of these two descriptions of the *attr*: a, b. Which one is off? Give only that option.	93	27	88	40

ID	Instruction	E_{text}	R_{audio}	E_{text}	R_{audio}
H10	Which label for the *attr* would fail a spot check on this clip, a or b? Reply with only that label.	92	34	95	26
Free-form paraphrases (bank; training phrasings in §4)					
T1	Which of a and b does the audio not support as the *attr*? Reply with only that option.	93	31	94	28
T2	Tell me the *attr* you don't hear: a or b. Just that option, please.	57	66	50	71
T3	I'm after what you did NOT hear. For the *attr*, would that be a or b? Reply with a single option.	12	73	42	74
T4	One of these two does not describe the *attr* in this recording: a, b. Which one? Give only that option, nothing else.	93	14	95	18
T5	Listen again, then tell me which option does not fit the *attr* — a or b. Answer with that option only.	91	15	95	28
T6	Don't tell me the *attr* you hear — tell me the other one: a or b. Reply with that option alone.	24	72	42	78
T7	Which label does not hold true for the *attr* here, a or b? Reply with only that label.	88	21	94	20
T8	For the *attr*, one of a and b isn't right. Which one isn't? Give me that option only.	86	32	86	28
T9	Name the incorrect option for the *attr*: a or b. Reply with only that option.	96	20	97	18
T10	Identify the option that is untrue of the *attr* here (a vs. b). Give only that option.	88	25	91	25
T11	One of these two labels is wrong for the *attr*. Which is it: a or b? Give only that label.	85	34	95	26
T12	Two annotators labeled the *attr* in this clip: one put a, the other put b. One of them is wrong — which value is wrong? Reply with only that value.	91	20	91	15
T13	My notes list two values for the *attr* in this clip, a and b, and one of them is an error. Which value should I erase? Reply with only that value.	96	23	97	19
T14	Rule out one of a and b as the *attr*. Which one do you rule out? Reply with only that option.	90	28	90	25
T15	Cross off whichever of a and b can't be the *attr*, then name it. Reply with only that choice.	69	36	42	31
T16	Before this annotation ships, one of the two proposed values for the *attr* has to be thrown out: a, b. Which one goes? Reply with only that value.	84	30	94	23
T17	Eliminate the option that doesn't match the *attr*: a or b. Name the one you eliminate. Reply with only that option.	93	16	97	15
T18	For the *attr*, skip whichever of a and b you actually hear and give me the other one. One option only, no explanation.	69	48	61	70
T19	Ignore the *attr* this clip actually has and pick the other option: a or b. Reply with one option and nothing more.	56	68	70	70
T20	Cross off the value that matches the *attr* you hear, a or b, and report the one left over. Reply with only that value.	48	48	89	63
T21	Instead of the *attr* you can hear, give me the one you can't: a or b. Just that option, please.	88	41	84	58
T22	Which of a and b contradicts the *attr* in this audio? Reply with only that option.	88	44	94	35
T23	Sanity check on the *attr*: one of a and b clashes with what you hear. Which one clashes? Give only that option.	61	47	66	39
T24	Name the option that runs contrary to the *attr* in this clip: a or b. Reply with only that option.	96	26	96	33
T25	If you had to reject one of a and b as the *attr*, which would you reject? Reply with only that option.	81	39	92	31
Minimal pairs					
MP1L	Which of a and b is wrong for the *attr*? Reply with only that option.	89	33	93	29
MP1P	Which of a and b is wrong for the *attr* you hear? Reply with only that option.	91	36	93	28
MP2L	Which of a and b does not match the *attr*? Reply with only that option.	93	33	94	28

ID	Instruction	E_{text}	R_{audio}	E_{text}	R_{audio}
MP2P	Which of a and b does not match the *attr* you hear? Reply with only that option.	92	34	95	28
MP3L	One of a and b is incorrect for the *attr*. Which one? Reply with only that option.	83	38	78	41
MP3P	One of a and b is not the *attr* you hear. Which one? Reply with only that option.	83	39	81	44
MP4L	Name the option that is untrue of the *attr*: a or b. Reply with only that option.	93	20	98	20
MP4P	Name the option that is not what you hear as the *attr*: a or b. Reply with only that option.	93	17	98	12
MP5L	Which of a and b should be ruled out as the *attr*? Reply with only that option.	94	34	94	28
MP5P	Which of a and b do you not hear as the *attr*? Reply with only that option.	85	35	90	35
MP6L	Which of a and b contradicts the *attr*? Reply with only that option.	90	44	92	34
MP6P	Which of a and b contradicts the *attr* you hear? Reply with only that option.	88	42	90	32
MP7L	For the *attr*, which is the false option: a or b? Reply with only that option.	95	25	95	20
MP7P	For the *attr*, which is the option you did not hear: a or b? Reply with only that option.	92	27	95	20
MP8L	Between a and b, which is not the *attr*? Reply with only that option.	93	36	94	27
MP8P	Between a and b, which is not the *attr* you hear? Reply with only that option.	93	34	92	29
MP9N	Which of a and b is NOT the *attr*? Reply with only that option.	92	30	94	28
MP9W	Which of a and b is wrong as the *attr*? Reply with only that option.	88	31	92	31

D REFLEX BY TIER AND TASK CATEGORY

Table 8 gives the reflex rate of each base model on each of the 40 tiers, restricted to tiers on which the model's affirmative accuracy is at least 75% and, within them, to items answered correctly both under $\mathrm{I_a}$ and in the text control of the same negated instruction ($\mathrm{I_b}$–$\mathrm{I_d}$ pooled; a dash marks a tier below the threshold or with fewer than 30 remaining rows). Tiers follow MMSU's category assignment (23 perception, 17 reasoning). The last rows give the category means and a tier-level two-sided Mann–Whitney test: perception tiers reflex more than reasoning tiers in all six models, significantly so only for Qwen2.5-Omni. Across models, the lowest rates are on idiom, polysemy, homophone and syntactic-structure reasoning and the highest on background scene, prolonged-sound perception and intonation-based reasoning.

E LAYER-WISE ANSWER-READOUT ANALYSIS

We complement the behavioral analyses in Section 3 with a layer-wise diagnostic of when the difference between audio-conditioned and text-informed runs becomes visible in the model's answer representation. We emphasize that this analysis is descriptive rather than mechanistic: intermediate projections through the final unembedding reveal when answer-relevant information becomes linearly exposed to the output head, but do not identify where the underlying computation itself first differs.

For each item, we run the same negated instruction under two matched input conditions: the original audio input and the corresponding text control in which the relevant attribute is stated explicitly in text. At every transformer layer ℓ, we take the residual state at the answer position, apply the model's final normalization and unembedding, and compute the log-probability margin

$$M_\ell = \log p_\ell(a_{\text{neg}}) - \log p_\ell(a_{\text{perc}}),$$

Table 8: Reflex rate (%) per tier on the restricted subset, base models, MMSU category assignment.

Tier	Qwen2.5-Omni	Qwen3-Omni	MiMo	Qwen3.5-Omni	Gem. Flash	Gem. Pro
Perception tiers						
accent identification	–	–	–	7	–	19
age prediction	14	18	–	8	14	16
consonant vowel perception	–	30	–	7	11	19
dialogue turn counting	–	–	–	–	19	16
disfluency detection	–	–	–	11	20	13
emotion recognition	24	10	–	14	–	–
gender prediction	11	15	7	0	11	16
intonation perception	–	–	–	–	–	–
near homophone perception	35	–	8	1	23	6
non verbal sound detection	26	1	3	0	11	4
pause perception	–	–	–	11	36	–
pitch comparison	–	–	–	–	–	–
plosive sound identification	–	–	–	–	–	–
prolonged sound perception	42	28	29	7	17	16
speaker identity recognition	–	–	–	–	–	–
speech act classification	34	7	11	1	10	11
speech duration estimation	–	44	–	20	–	11
speech stress perception	–	–	–	6	–	–
speed comparison	–	–	–	–	–	–
syllable perception	–	42	–	6	13	7
total speaker counting	–	–	–	14	14	3
vocal range comparison	–	–	–	8	–	9
volume comparison	–	–	–	36	–	–
Reasoning tiers						
background scene recognition	32	24	–	14	36	29
code switch question answering	16	20	17	10	23	15
deixis resolution	16	32	14	4	6	7
dialogue reasoning	11	45	3	4	4	0
homophone based reasoning	2	1	1	0	34	6
idiom reasoning	1	1	0	0	0	0
intent detection	6	15	12	4	28	16
intonation based reasoning	–	29	19	20	25	22
logical reasoning	11	7	15	2	4	4
pause based reasoning	10	9	8	5	14	10
polysemy reasoning	5	6	8	1	3	1
prolonged sound reasoning	9	21	3	0	5	3
puns interpretation	15	26	–	1	2	16
sarcasm detection	–	–	–	–	–	–
stress based reasoning	–	13	–	4	11	11
syntactic structure matching	15	13	10	1	2	0
synthetic speech detection	–	–	–	8	–	–
Mean, perception tiers	26.5 (7)	21.8 (9)	11.5 (5)	9.2 (17)	16.6 (12)	11.9 (14)
Mean, reasoning tiers	11.4 (13)	17.4 (15)	9.2 (12)	4.8 (16)	13.1 (15)	9.3 (15)
Mann–Whitney p	0.011	0.474	1.000	0.052	0.213	0.183

where a_{neg} is the answer required by the negated instruction and a_{perc} is the attribute supported by the audio. To remove item-specific answer preferences, we subtract the corresponding margin from the affirmative run of the same item. We then measure the modality difference

$$D_\ell = M_\ell^{\text{text}} - M_\ell^{\text{audio}}.$$

Figure 4 plots D_ℓ for Qwen2.5-Omni and Qwen3-Omni across representative tiers. Through most of the network, the difference between the text-informed and audio-conditioned runs remains small in this output-oriented readout. The gap becomes pronounced only in later layers, approximately when the answer itself becomes decodable from the intermediate residual state. On Qwen2.5-Omni, this transition occurs around layers 22–23 of 28; on Qwen3-Omni, it occurs around layers 39–40 of 48.

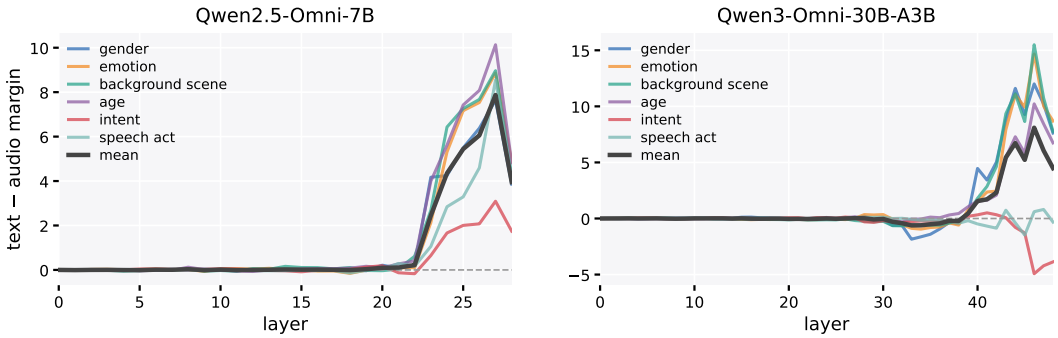

Figure 4: **Layer-wise emergence of the audio–text gap in the answer readout.** For each item, the same negated instruction $\mathrm{I_b}$ is evaluated with audio and with the matched text control. At each layer, we project the answer-position residual state through the final normalization and unembedding, compute the log-probability margin between the negated target and the attribute supported by the audio, subtract the same item's affirmative-run margin, and plot the resulting text-minus-audio difference D_ℓ. Thin lines denote tiers and the grey line their mean. The modality gap becomes visible primarily in later layers, near the point at which answer identity itself becomes decodable.

The qualitative pattern is similar when items are grouped by phrasing difficulty or by the eventual response of the audio-conditioned run: substantial separation appears only after answer-related logits begin to emerge. We therefore interpret the late divergence as indicating when the modality-dependent behavioral difference becomes visible to the model's output head, rather than as evidence that the underlying representations are identical before that point.

Interpretation and limitation. This analysis should not be read as localizing the source of the reflex. A near-zero value of D_ℓ at an intermediate layer only indicates that the difference is not yet exposed along the final answer directions captured by the model's unembedding matrix. The audio- and text-conditioned states may already differ in other representational subspaces. Accordingly, we use this experiment only as a diagnostic of the emergence of the behavioral gap in an output-oriented readout.

F Training Details

Data. The training set has 33,722 examples in three groups: 11,400 *override* examples (a negated instruction whose gold answer is the distractor, rendered with one of the 25 training phrasings in Appendix C, none of which appears in the benchmark), 11,619 *anchor* examples (the affirmative question on the same clips, to keep perception in place) and 10,703 *retention* examples (general audio instructions, including transcription of LibriSpeech training-split utterances, to keep other abilities). SFT trains on the same examples with the gold answer as target.

Teacher input. For override items, the OPSD-text teacher receives silence in place of the audio, the prefix "You are told: the {attribute} is {true value}." and the canonical phrasing *"Which of a and b does the audio not support as the {attribute}? Reply with only that option."*; the variant without canonical phrasing uses the same prefix followed by the student's own instruction. The OPSD-audio teacher receives the audio, the prefix "What is the {attribute}? The answer is '{true value}'." and the student's instruction. Affirmative items use the audio-teacher form with the affirmative question; retention items use the student's exact input. The student always sees the audio and its own instruction. The 25 training phrasings are listed in Appendix C.

Optimization. Qwen2.5-Omni: full fine-tuning of the thinker with the audio encoder frozen; rollouts are greedy (at most 32 new tokens) and the reverse KL is computed at temperature 1 over the full vocabulary; AdamW with learning rate 10^{-6}, micro-batch 1 with 8 accumulation steps on 4 GPUs (effective batch 32), one epoch. SFT uses the same schedule with a cross-entropy loss on the gold answer. Qwen3-Omni-30B-A3B: LoRA (Hu et al., 2021) with rank 64 and $\alpha = 128$ on the

query, key, value and output projections of every language-model layer of the thinker; learning rate 10^{-4}, micro-batch 1 with 8 accumulation steps on 8 GPUs (effective batch 64), same data and objective; the teacher is the same weights with the adapter disabled. The distilled variants were trained for one epoch. SFT was trained for half an epoch on 4 GPUs at learning rates 10^{-4}.