How Sound Meets Language: Causal Mechanisms of Audio-Language Integration in Large Audio Language Models

Anonymous authors
Paper under double-blind review

Abstract

How do large audio language models (LALMs) integrate task-relevant auditory evidence into language computation? We introduce AudioEvidence, a controlled testbed with temporal evidence annotations, matched non-evidence controls, and counterfactual sound replacements, to trace this process across nine LALMs. At the backbone input, the causal effects of auditory evidence remain temporally localized. Within the backbone, eight models share a depth-wise organization: auditory evidence states exert stronger effects early, direct evidence access is most consequential at intermediate layers, and text-state effects strengthen with depth. Joint interventions establish that blocking intermediate evidence access weakens the influence of subsequent text states on the answer. The critical access stage also supports additional tasks and longer audio inputs. Head contributions vary in concentration across models, with MiMo-Audio exhibiting sparse organization that parallels retrieval-head findings in text LLMs. High evidence-attention scores alone do not establish causal importance. Finally, amplifying selected attention heads in MiMo-Audio improves audio hallucination detection and reduces task deviations in speech recognition. These findings provide mechanistic priors for more efficient audio scaling and modality integration.

1 Introduction

Large audio language models (LALMs) extend large language models to acoustic inputs, combining auditory perception with language reasoning (Tang et al., 2024; Xu et al., 2025b). Unlike discrete text symbols, audio is continuous and temporally fine-grained; depending on how audio is encoded and compressed, similar semantic content can require substantially more tokens than text (Hsu et al., 2026; Mitsui et al., 2024). Audio understanding therefore requires both cross-modal interpretation and evidence selection over potentially long contexts (Lin et al., 2025). The challenge is not only to encode sound, but also to identify task-relevant auditory evidence and make it available to language-side computation. Our broader goal is to establish mechanistic priors for more efficient audio scaling and better integration of audio with language intelligence.

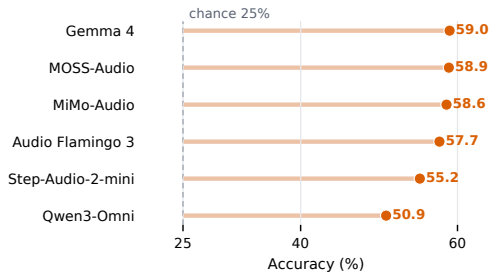


Figure 1: Several LALMs retain substantial accuracy on MMAU-test-mini when the original audio is replaced by equal-duration **SILENCE**.

Prior studies have identified limitations in LALMs' auditory grounding (Kuan & Lee, 2025), including substantial accuracy when audio is replaced by silence (He et al., 2026). We observe a similar pattern in more recent LALMs on MMAU (Sakshi et al., 2025): several models retain substantial accuracy under equal-duration silence (Figure 1). This reinforces the distinction between *answering correctly* and *using auditory evidence*, motivating a closer examination of how specific sounds contribute to model predictions. Controlled evaluations of speech emotion understanding also reveal

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.

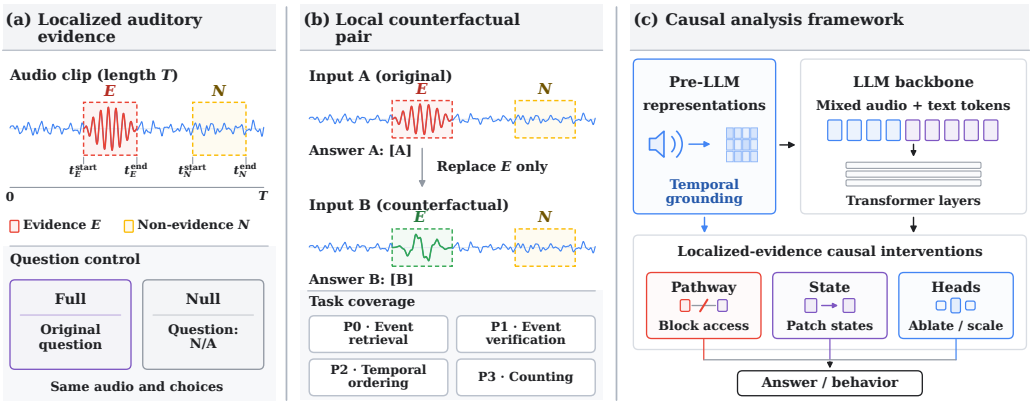


Figure 2: **AUDIOEVIDENCE and our causal analysis framework.** Localized counterfactual evidence enables causal analysis of pre-LLM grounding and evidence access, state, and component effects within the LLM backbone.

a preference for lexical content over acoustic cues (Chen et al., 2026a), highlighting the need to distinguish which aspects of an audio input support an answer.

Mechanistic studies have examined this problem through causal tracing of audio–text fusion, representation analyses of auditory and task-relevant semantics, and attention interventions on audio-sensitive components and cross-modal pathways (Chen et al., 2026b; Fan et al., 2026; Yang et al., 2025; Glazer et al., 2026; Suharitdamrong et al., 2026). Yet many analyses operate at the whole-audio or modality level, making it difficult to separate the contribution of answer-determining evidence from other distortion in the input. Many studies also examine only one or a few models, leaving unclear which mechanisms recur and which vary across LALMs. We therefore ask: Where and how is auditory evidence integrated into language computation? Across different architectures and scaling recipes, which mechanistic properties are shared, and which differ?

A natural hypothesis comes from text and multimodal LMs, where specialized *retrieval heads* select relevant information and route it into downstream computation. In text LMs, a small subset of such heads can make disproportionate causal contributions (Wu et al., 2025b); analogous heads have also been identified in vision–language models (Li et al., 2026), with recruitment varying across generation steps (Lin et al., 2026). This motivates our initial hypothesis: *LALMs may exhibit a similarly sparse retrieval-head organization in which evidence-attending heads carry disproportionate causal importance.*

To study these problems, we introduce **AUDIOEVIDENCE**, a controlled multitask mechanistic testbed with explicit temporal annotations of answer-relevant sounds and matched non-evidence controls. In its core retrieval task, counterfactual pairs change the correct answer by replacing only the annotated evidence event, while preserving the question, answer choices, event timing, and waveform outside the evidence interval. This design enables controlled causal analysis of how localized auditory evidence affects language computation (Figure 2). We study nine LALMs spanning different architectures and scaling regimes, separating pre-LLM representations from computation within the LLM backbone.

The causal effects of auditory evidence remain temporally localized at the backbone input. When we move the same sound to different temporal positions, the regions supporting answer recovery move with it. This extends prior temporal-alignment observations in two LALMs (Ma et al., 2026) and establishes a starting point for tracing evidence use through the backbone.

MiMo-Audio exhibits concentrated causal effects in a small subset of evidence-attending heads. We test the retrieval-head hypothesis by ranking heads according to their question-conditioned increase in evidence attention and measuring the effects of their ablation. Ablating the top 5% of heads under the T_{All} ranking in MiMo-Audio reduces MMAU accuracy from 77.8% to 47.7%, while equal-budget global random ablation retains 74.8%. The distribution of its high-scoring heads also resembles that observed in text and multimodal LMs. In other models, the

accuracy gap between targeted and random ablation is smaller and varies with the ranking scope and ablation budget. We therefore investigate whether audio–language integration follows a shared mechanism.

Eight of the nine LALMs studied share an intermediate stage of audio–language integration that shapes subsequent language computation. Counterfactual state replacement reveals a shift in causal influence across depth: auditory evidence states have stronger effects in earlier layers, while text states become more influential later in computation. We then examine the pathways underlying this shift by blocking attention access to auditory evidence. These interventions independently identify critical access intervals, typically at intermediate depths, that overlap the transition in state effects. Combining the two interventions, we find that blocking evidence access within these intervals weakens the ability of later text states to transfer the donor's answer preference to a fixed recipient. Together, these results establish that evidence access within the identified intervals causally contributes to the subsequent influence of text states on the answer. Across models, the intervals differ in location and width, and head-level contributions within them differ in concentration.

Targeted head interventions reduce hallucination-related errors and task deviations. On MiMo-Audio, amplifying heads selected by our evidence-attention ranking procedure improves audio hallucination detection and reduces WER on LibriSpeech (Panayotov et al., 2015) from 3.441% to 2.866%. The WER gain comes primarily from correcting task deviations, especially generating translations instead of the requested transcriptions.

Our contributions are fourfold:

- **A controlled testbed for auditory evidence use.** We introduce AUDIOEVIDENCE, with localized evidence annotations, counterfactual pairs, and matched non-evidence controls that support causal analysis beyond whole-audio perturbations.
- **A shared pattern of audio–language integration.** We show that the causal effects of auditory evidence remain temporally localized at the backbone input and that intermediate-depth evidence access shapes subsequent text-state effects across eight LALMs.
- **Model-dependent head organization.** We establish that attention-based rankings have variable causal relevance across models, with MiMo-Audio showing pronounced concentration of ablation effects in a small subset of heads.
- **Targeted error reduction.** We use evidence-guided head interventions to improve audio hallucination detection and reduce task deviations in speech recognition.

2 RELATED WORK

Mechanistic interpretability of LALMs. Causal tracing and activation patching localize internal states and circuits that affect language-model predictions (Meng et al., 2022; Wang et al., 2023), with methodological studies examining how intervention design and evaluation metrics shape the conclusions (Zhang & Nanda, 2024). In LALMs, causal tracing has characterized layer- and token-dependent audio–text fusion (Chen et al., 2026b). AudioLens tracks auditory attributes across layers and tests their dependence on direct access to auditory positions (Yang et al., 2025). Middle-layer workspace analysis combines semantic readout, waveform substitution, and activation patching to study audio-driven task representations (Fan et al., 2026). Attention-knockout studies trace cross-modal information flow across input settings and model scales (Zhang et al., 2025; Suharitdamrong et al., 2026). Building on these approaches, we use localized counterfactual evidence and matched controls to connect evidence access with the formation of downstream answer-related text-state effects across diverse LALMs.

Attention heads and evidence retrieval. Head-ablation studies show that attention heads differ substantially in their contributions to task performance (Michel et al., 2019). In text LMs, a small subset of *retrieval heads* can carry disproportionate causal responsibility for long-context retrieval (Wu et al., 2025b). MMRetHeads extends attention-based identification and causal validation to textual and visual evidence (Li et al., 2026), while dynamic retrieval analysis finds that head recruitment varies across generation steps (Lin et al., 2026). In audio models, CALM learns class-dependent weights over attention-head features for few-shot classification (Mehta et al., 2026).

We examine whether evidence-attending heads make concentrated causal contributions to auditory evidence access and compare this organization across models.

Inference-time interventions. MATA modifies attention scores from the final sequence position to audio tokens in intermediate layers to address audio–text attention imbalance (Wang et al., 2025). Specialist-guided steering identifies heads whose audio attention predicts correctness and uses their locations to construct an audio–silence activation direction (Glazer et al., 2026). Under conflicting audio and text inputs, back-patching transfers later audio representations to earlier layers to reduce text dominance (Cho et al., 2026). Our control experiments scale the outputs of heads selected by question-conditioned attention on evidence, with configurations calibrated before downstream evaluation. The accompanying causal analyses test how evidence access through selected heads contributes to predictions.

3 AUDIOEVIDENCE AND EXPERIMENTAL SETUP

We construct AUDIOEVIDENCE to test how specific sound events causally affect model predictions (Figure 2). We collect events from AudioSet Strong (Hershey et al., 2021) and AudioEdit (Tian et al., 2026), screen them using metadata, acoustic checks, and auxiliary Gemini (Comanici et al., 2025) and CLAP (Wu et al., 2023) assessments, and assemble them into audio examples with controlled evidence and distractors.

AUDIOEVIDENCE contains four auditory tasks: P0 *Retrieve*, P1 *Verify*, P2 *Relate temporal order*, and P3 *Aggregate* event counts. We primarily use non-speech sounds to study auditory evidence without the additional linguistic content of speech.

Gemma 4's reported audio evaluations focus on speech (Gemma Team et al., 2026), leaving its non-speech capabilities less characterized. To study its encoder-free architecture and broaden the scope of AUDIOEVIDENCE, we introduce a speech extension evaluated across all nine models (Appendix H).

For each example, temporal annotations identify the answer-relevant events and are mapped to the model's native audio positions E. We construct a non-overlapping matched control N from the same audio, with $|E| = |N|$. The **Full** condition contains the original question, whereas **Null** replaces it with `N/A` while preserving the audio and answer choices, allowing us to measure question-conditioned evidence attention.

For causal analysis, we construct counterfactual pairs A and B that preserve the question, choices, timing, and waveform outside the evidence interval, but replace the target event so that the correct answer changes. The testbed includes 10-s and 30-s versions. Core mechanistic experiments use 10-s P0 examples; selected analyses extend to 30-s inputs and additional tasks to examine sensitivity to context duration and task type. For inference-time head selection, we construct a larger discovery sample using the same sampling and screening procedures and preserving the data distribution. Dataset construction, evidence annotation, sampling, and quality-control procedures, together with sample counts and experiment coverage, are detailed in Appendix A.

We study nine LALMs: MiMo-Audio, Qwen2-Audio, Qwen2.5-Omni, Step-Audio-2-mini, Audio Flamingo 3, Kimi-Audio, MOSS-Audio, Qwen3-Omni, and Gemma 4 (Xiaomi LLM-Core Team, 2025; Chu et al., 2024; Xu et al., 2025a; Wu et al., 2025a; Ghosh et al., 2025; KimiTeam et al., 2025; Yang et al., 2026; Xu et al., 2025b; Gemma Team et al., 2026). These models cover discrete, continuous, hybrid, and encoder-free audio interfaces, different temporal compression rates, dense and MoE backbones, and varied training regimes. This selection allows us to examine shared mechanisms alongside differences across model families.

We retain each model's native frontend and prompt format. We use T_{All} to denote all input positions after the final audio token, through the end of the input sequence. Our interventions operate on three objects: **state patching** replaces hidden states between counterfactual runs; **path blocking** removes selected attention connections; and **head intervention** suppresses or scales selected attention-head outputs (Figure 7). We measure intervention effects using accuracy and answer logit margins, capturing both changes in correctness and shifts in answer preference. Appendix B provides the model comparison and further experimental details.

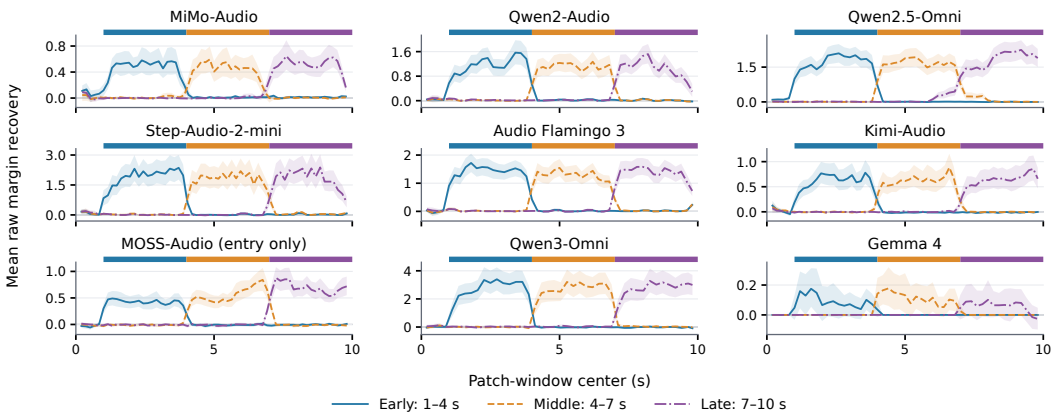


Figure 3: Pre-LLM patching curves for three event placements across nine models, using 400-ms windows. Colored bars mark evidence intervals; shading indicates 95% confidence intervals. Vertical scales vary by model.

4 Evidence Localization in Pre-LLM Representations

AudioEvidence provides temporal evidence annotations at the waveform level. To trace this evidence inside the models, we first ask whether its causal effects on the answer remain localized at the corresponding temporal positions after frontend processing, since models have different frontends to encode audio features.

We construct 120 controlled event compositions from the P1 event-presence task. Each composition contains a counterfactual pair A/B with the same question and choices, differing only in the sound within the evidence interval and its corresponding correct answer. We place the same event pair at 1–4, 4–7, and 7–10 seconds of a 10-s waveform, keeping the event waveforms, question, choices, and distractor fixed across placements. This yields 360 pairs and allows us to test whether the region supporting answer recovery follows the event's position.

Let U^A and U^B denote the pre-LLM representations of A and B. For an audio-position window W, we replace B's representations within W with the corresponding representations from A (Figure 8):

$$\widetilde{U}^{B,W}[j] = \begin{cases} U^A[j], & j \in W, \\ U^B[j], & j \notin W. \end{cases}$$

We then continue computation through the language backbone and measure the change in preference toward A's answer:

$$R(W) = m_a(\widetilde{B}^W) - m_a(B),$$

where $\widetilde{B}^W$ denotes the modified run of B, and m_a is the logit margin of A's correct answer a over the highest-scoring alternative. Positive values indicate a shift toward A's answer. We scan 400-ms windows and report complementary scans with evidence-width windows of 3 s in Appendix C to examine localization at a broader patching scale.

As shown in Figure 3, mean answer recovery is concentrated near the evidence interval. When the evidence moves, the regions with stronger recovery move with it. This localization pattern appears across the tested models with different audio interfaces. These results show that answer-related causal effects remain temporally localized at the backbone input, supporting the use of evidence timestamps to identify internal intervention positions. We next examine how the language backbone accesses this evidence and integrates it into text computation.

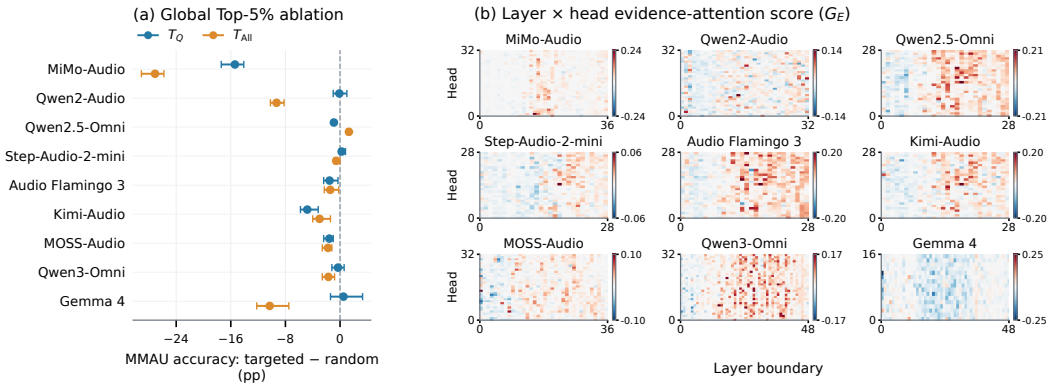


Figure 4: (a) MMAU accuracy differences between global top-5% head ablation ranked by T_Q or T_{All} and random-head ablation; points and bars show the mean and range across three random seeds. (b) Layer–head evidence-attention scores G_E with model-specific color scales.

5 From Evidence Attention to Causal Integration

5.1 Retrieval-Head Hypothesis

In text and multimodal LMs, retrieving task-relevant evidence can rely heavily on a small subset of attention heads (Wu et al., 2025b; Li et al., 2026; Wu et al., 2026). We therefore ask whether LALMs use a similar sparse retrieval-head mechanism to select answer-relevant auditory evidence.

We rank heads by question-conditioned attention to evidence. For each head (l, h), let $A^c_{i,l,h}(q, k)$ denote the attention from query position q to audio position k for example i under condition $c \in \{\text{Full}, \text{Null}\}$. For a query set S, we define

$$M^c_{i,l,h}(S, E) = \frac{1}{|S^c_i|} \sum_{q \in S^c_i} \sum_{k \in E_i} A^c_{i,l,h}(q, k),$$

and rank heads by their question-conditioned evidence-attention gain,

$$G_E(l, h; S) = \frac{1}{n} \sum_{i=1}^{n} \left[M^{\text{Full}}_{i,l,h}(S, E) - M^{\text{Null}}_{i,l,h}(S, E) \right].$$

Our primary ranking uses question tokens, $S = T_Q$, while T_{All} includes all input positions after the final audio token, through the end of the input sequence. We rank heads across all layers using 480 P0 discovery examples from AudioEvidence, then zero the outputs of the top-ranked 5% of heads on MMAU-test-mini and compare their effects with ablation of the same number of globally sampled random heads. Details of ranking, random sampling, and ablation are provided in Appendix D.1.

MiMo-Audio provides strong evidence for sparse head organization, paralleling findings on retrieval heads in text LLMs. As shown in Figure 4(b), its high-G_E heads are sparsely distributed and concentrated around intermediate layers, resembling retrieval-head patterns reported in text and multimodal language models. Ablation further confirms their contribution to prediction: removing the top 5% of globally ranked heads sharply reduces MMAU accuracy, whereas equal-budget global random ablation causes substantially smaller degradation (Figure 4(a)). Thus, in MiMo, a small set of heads selected by evidence attention makes a substantial contribution to model predictions.

Appendix D.2 provides detailed MMAU and MMAR (Ma et al., 2025) results at ablation budgets of 2.5%, 5%, 7.5%, and 10%. These results show that the additional damage caused by targeted ablation over random ablation is significantly smaller in other models than in MiMo-Audio on both benchmarks, with many head-ablation conditions showing no significant advantage over random ablation. These results show that ablation effects in MiMo-Audio are concentrated in a small subset of evidence-attending heads. In the next subsection, we ask whether a broader audio–language integration mechanism is shared across models.

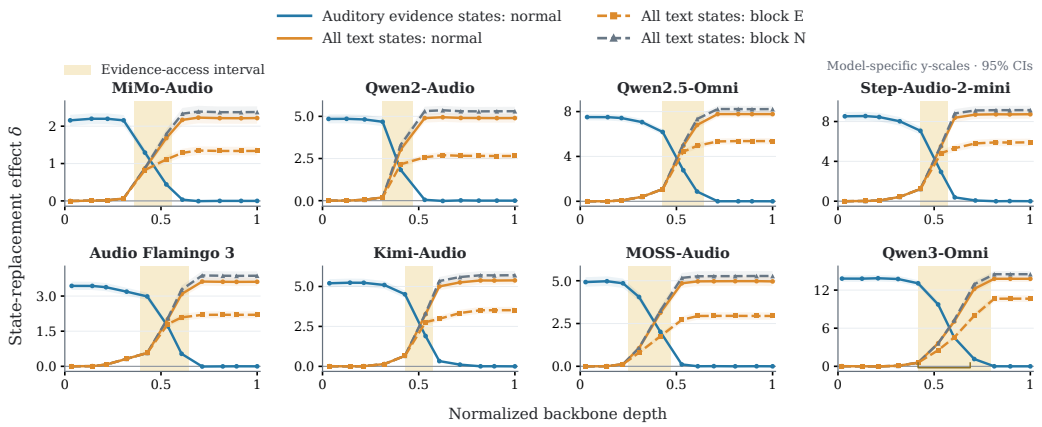


Figure 5: State-patching effects across normalized backbone depth under normal access and upstream auditory evidence (E) or non-evidence (N) blocking. Bands show pointwise 95% confidence intervals; shaded regions mark access envelopes. Vertical scales vary by model.

5.2 From Auditory Evidence to Language Computation

Since head-level organization varies across models, we next ask whether they share a broader mechanism of audio–language integration. We first trace the causal effects of auditory evidence states and text states, then locate critical access stages and test whether this access drives the state effects change.

State interventions reveal a shift in causal influence with depth. We use 480 counterfactual pairs from the 10-s P0 evaluation set. Each pair A, B shares the same question and answer choices and differs only in auditory evidence, with correct answers a, b, respectively. We place 11 probes uniformly across normalized backbone depth, mapped to integer layer indices (Figure 7(a)). At each probe, we replace block-output states at auditory evidence positions E or all text positions T_{All} in the run on B with the corresponding states from a normal run on A, leaving other positions unchanged. Let $\widetilde{B}_{S,\ell}$ denote the run after replacing positions S at layer ℓ, and define $m_{a,b}(X) = z_a(X) - z_b(X)$, where z_a, z_b are the answer-token logits at the first answer position. The patching effect is

$$\delta_S(\ell) = m_{a,b}(\widetilde{B}_{S,\ell}) - m_{a,b}(B), \qquad S \in \{E, T_{\text{All}}\}.$$

Positive values indicate a shift toward A's answer. Figure 5 plots mean effects against $(\ell + 1)/L$, where ℓ is zero-indexed and L is the number of backbone layers. Across eight models, replacing auditory evidence states has strong effects early; these effects weaken with depth, while text-state effects strengthen. Qwen3-Omni exhibits a later, broader transition. Because A and B have identical input text, text-state effects reflect differences induced by auditory evidence. **The causal influence of auditory evidence thus shifts toward text states with depth.**

Pathway blocking localizes critical access from text to auditory evidence. We control attention from T_{All} queries to E keys across all heads (Figure 7(b)). For a layer boundary b, Prefix interventions permit this access only before b, while Suffix interventions permit it only at or after b. Sweeping b, we measure the mean margin of the correct answer over the strongest incorrect option, using all-layer blocking and normal access as the 0% and 100% recovery references. Prefix tests how far early access must extend to recover answer preference; Suffix tests how long access can be delayed while preserving it. On the P0 discovery split, the transition ranges covering 20%–80% recovery define access envelopes used to select validation intervals. **The access intervals span intermediate layers in eight models** (Figure 5); full curves and selection rules are in Appendix E.2.

We then test the selected interval's causal specificity against non-evidence access and evidence access at other depths. On the P0 evaluation split, blocking auditory evidence access within these intervals reduces accuracy by 26.7–39.6 percentage points. In all eight models, this effect exceeds blocking matched N positions within the same interval or blocking E in equally wide early or late

intervals. **Text to auditory evidence access at intermediate depths therefore constitutes a critical stage in answer causal influence.** Surprisingly, these intervals overlap the depth range of the causal shift from auditory evidence states to text states.

To test **whether access within these intervals drives the shift**, we repeat text-state replacement while processing A under normal access, E blocking, or matched N blocking. At each probe, blocking applies only to validation-interval layers strictly before that probe. The resulting T_{All} states are substituted into the same normal run on B, and their effects are measured using δ. Thus, only access during the formation of the replacement states changes. Blocking E weakens later text-state effects, whereas blocking N generally preserves or increases them (Figure 5). At the prespecified primary probe—the first grid layer at or beyond the interval's endpoint—N-blocked text states have stronger effects than E-blocked states in all eight models, with paired 95% confidence intervals above zero and significance after Holm correction across eight comparisons. Together, state and pathway interventions identify **a shared audio–language integration mechanism** across eight models: a critical intermediate access stage drives the shift in answer-related causal influence from auditory evidence states to text states.

We then revisit head-level concentration within the critical access intervals, using the attention rankings from Section 5.1. In MiMo-Audio, blocking all text-to-auditory-evidence access in the top 10% of interval heads produces 82.4% and 81.8% of the margin reduction from blocking all heads in the interval, under the T_Q and T_{All} rankings, respectively. **The effects of blocking evidence access differ in their head-level concentration across models**, with pronounced concentration in MiMo-Audio. Complete budget curves and matched non-evidence controls are provided in Appendix E.4.

High evidence-attention scores alone do not establish causal importance. Attention-based head rankings have different causal relevance across models. At the layer level, strong evidence attention in Step-Audio-2-mini, Qwen2-Audio, and MOSS-Audio extends into later layers beyond their critical access intervals (Appendix E.5). Thus, attention strength and the causal importance of evidence access can differ across both heads and layers. Our interventions provide an audio-specific test of the broader distinction between attention weights and explanatory importance (Jain & Wallace, 2019; Serrano & Smith, 2019).

5.3 Generalization Across Tasks and Audio Durations

We further examine whether this depth-wise organization extends to longer audio and other tasks. On paired 10-s and 30-s P0 inputs, eight models retain the overall pattern of decreasing auditory-evidence state effects and increasing text-state effects with depth. Using the critical intervals identified on 10-s retrieval, blocking evidence access within these intervals has stronger effects than matched non-evidence blocking or equal-width early and late blocking across all eight models on 30-s retrieval, event verification, and temporal ordering. Counting shows greater variation across models. On MMAU and MMAR, blocking text-to-audio access within the same intervals also causes larger answer-margin reductions than early or late blocking. These results support the transfer of the critical audio-access stage across tasks and durations. Full results are provided in Appendix F.

6 Inference-Time Control

Finally, we use evidence-related heads as inference-time control points. We apply the same head-ranking procedure to the expanded AudioEvidence discovery sample introduced in Section 3. With this ranking fixed, we calibrate the number of top-ranked heads K and a shared output-scaling factor α on a balanced set spanning P0–P3. The selected configuration is $K = 80$ and $\alpha = 1.3$ (Figure 21). We apply this fixed policy to all downstream evaluations below.

The selected intervention increases accuracy on HalluAudio (Zhao et al., 2026) from 68.00% to 69.24% and on AHa-Bench (Cheng et al., 2025) from 56.77% to 58.24%. It also reduces WER on LibriSpeech test-clean (Panayotov et al., 2015) from 3.441% to 2.866% (Table 1). On both hallucination benchmarks, false positives decrease while false negatives increase. On LibriSpeech, the WER gain primarily reflects fewer hallucination-related errors and task deviations, especially generating Chinese translations of English speech instead of the requested English transcriptions. Outputs containing Chinese characters at baseline account for 89.1% of the net word-error reduction. Together, these results identify selected heads as control points for hallucination-related behavior

Table 1: **Inference-time control on MiMo-Audio.** Paired 95% confidence intervals are reported for the absolute changes.

Benchmark / Metric	Baseline	Top-80, α=1.3	Gain [95% CI]
HalluAudio Acc. ↑	68.00	**69.24**	$+1.24\,[+0.43, +2.10]$
AHa-Bench Acc. ↑	56.77	**58.24**	$+1.47\,[-0.23, +3.14]$
LibriSpeech WER ↓	3.441	**2.866**	$-0.574\,[-1.002, -0.224]$

and task adherence. Error breakdowns and qualitative examples are provided in Appendix G. Scaling all heads within the identified critical access interval also improves HalluAudio accuracy in Qwen2-Audio from 56.86% to 58.32%. Additional successful configurations in Qwen2-Audio, Qwen2.5-Omni, and Qwen3-Omni are reported in Appendix G.5.

7 Discussion and Limitations

Shared organization across architectures. Across eight models with different audio interfaces and backbones, evidence-state effects weaken with depth, text-state effects strengthen, and direct evidence access is especially consequential at intermediate layers. This suggests depth-wise functional specialization reminiscent of the encoding-to-generation division in sequence-to-sequence models (Sutskever et al., 2014), with potentially different head-level realizations from text and vision–language models. The shared access stage and varying head organization suggest priors for allocating audio–language computation; testing them during training could guide more efficient audio scaling and modality fusion.

Possible sources of architectural performance differences. MiMo-Audio shows the clearest concentration of head-level effects. Its compressed audio interface supplies 6.25 temporal positions per second to the backbone, and its residual vector quantization (RVQ) tokenizer discretizes the audio representation. These properties may favor retrieval mechanisms closer to those operating over text tokens. Training scale and the balance of audio–text training may also influence whether such specialization emerges; MiMo-Audio was pretrained on over 100 million hours of audio. Conversely, continuous and hybrid interfaces may distribute evidence across features before backbone entry, potentially favoring contributions spread over more heads and layers. These are competing, potentially interacting hypotheses; temporal compression, discretization, and training scale are not independently varied here. Controlled training studies could test these hypotheses and clarify how architecture and training scale shape audio–language integration. We view Gemma 4's encoder-free architecture as a promising route to simpler modality unification. We look forward to the breakthroughs that further scaling of this architecture may bring.

Scope and limitations. Our core analyses evaluate the full sequence of state, pathway, and joint interventions on 480 examples or counterfactual pairs from the 10-s P0 evaluation split. Important next steps include larger evaluations on naturalistic audio, additional tasks, and longer durations, plus extensions to open-ended tasks.

8 Conclusion

We introduce AudioEvidence, with localized evidence annotations and counterfactual controls, to trace audio–language integration across nine LALMs. Across eight models, causal interventions identify a shared organization: auditory evidence states exert stronger effects early, direct evidence access is most consequential at intermediate layers, and text-state effects strengthen with depth. Joint interventions establish that intermediate evidence access shapes the influence of subsequent text states on the answer. Head-level concentration varies across models, and high evidence-attention scores alone do not establish causal importance. In MiMo-Audio, amplifying selected heads improves audio hallucination detection and reduces task deviations in speech recognition. These findings provide mechanistic priors for where and how to allocate computation to audio–language integration, guiding future studies of more efficient audio scaling and modality fusion.

AI USE STATEMENT

Generative AI tools were used to assist with research brainstorming and hypothesis refinement, experimental design and methodology feedback, software implementation and debugging, result interpretation, literature exploration, figure preparation, manuscript structuring, and language editing. All research decisions, experimental procedures, implementations, analyses, and scientific claims were independently reviewed and validated by the authors. AI-assisted code and generated suggestions were manually inspected and verified. The authors take full responsibility for the final content of this work.

REFERENCES

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.eacl-long.274. URL https://aclanthology.org/2026.eacl-long.274/.

Wei-Chih Chen, Chien yu Huang, and Hung yi Lee. Causal Tracing of Audio-Text Fusion in Large Audio Language Models. *arXiv preprint arXiv:2603.13768*, 2026b. URL https://arxiv.org/abs/2603.13768.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. AHa-Bench: Benchmarking Audio Hallucinations in Large Audio-Language Models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-0398. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1169afa8ad13c98239f0b666b099e159-Paper-Datasets_and_Benchmarks_Track.pdf.

Hyebin Cho, Suho Yoo, Jaehyuk Jang, Changick Kim, and Joon Son Chung. Who Wins the Conflict? Mechanistic Interpretability of Text Bias in Audio LLMs. *arXiv preprint arXiv:2606.18924*, 2026. URL https://arxiv.org/abs/2606.18924.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*, 2024. doi: 10.48550/arXiv.2407.10759. URL https://arxiv.org/abs/2407.10759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, 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. URL https://arxiv.org/abs/2507.06261.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Jiajun Fan, Jingyuan Li, Prashanth Gurunath Shivakumar, Qi Luo, Jia-Hong Huang, M. Maruf, Roger Ren, Yile Gu, Rahul Pandey, Ge Liu, and Ivan Bulyko. Can We Read the Mind of an Audio LLM? A Verbalizable, Multilingual Middle-Layer Workspace. *arXiv preprint arXiv:2608.24958*, 2026. URL https://arxiv.org/abs/2608.24958.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, et al. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41819–41886. Curran Associates, Inc., 2025. doi: 10.52202/085713-1396. URL

https://proceedings.neurips.cc/paper_files/paper/2025/hash/3babb6b453cb59d87cb58a1219ef914b-Abstract-Conference.html.

Neta Glazer, Lenny Aharon, and Ethan Fetaya. Are Audio-Language Models Listening? Audio-Specialist Heads for Adaptive Audio Steering. *arXiv preprint arXiv:2603.06854*, 2026. URL https://arxiv.org/abs/2603.06854.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, Chunyat Wu, Qianhua He, Tan Lee, Xie Chen, Wei-Long Zheng, Weiqiang Wang, Mark Plumbley, Jian Liu, and Qiuqiang Kong. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21060.

Shawn Hershey, Daniel P. W. Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R. Channing Moore, and Manoj Plakal. The Benefit of Temporally-Strong Labels in Audio Event Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370, 2021. doi: 10.1109/ICASSP39728.2021.9414579. URL https://arxiv.org/abs/2105.07031.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. Anatomy of the Modality Gap: Dissecting the Internal States of End-to-End Speech LLMs. *arXiv preprint arXiv:2603.01502*, 2026. URL https://arxiv.org/abs/2603.01502.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing Cross-modal Information Hubs in Audio-Visual LLMs. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.10815.

KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-Audio Technical Report. *arXiv preprint arXiv:2504.18425*, 2025. URL https://arxiv.org/abs/2504.18425.

Chun-Yi Kuan and Hung-Yi Lee. Can Large Audio-Language Models Truly Hear? Tackling Hallucinations with Multi-Task Assessment and Stepwise Audio Reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/icassp49660.2025.10888384. URL https://doi.org/10.1109/ICASSP49660.2025.10888384.

Aaron Branson Cigres Li, Zhaowei Wang, Yu Zhao, Yiming Du, Haobo Li, Xiyu Ren, Ginny Wong, Simon See, Lishu Luo, Haodong Duan, Pasquale Minervini, and Yangqiu Song. Can Retrieval Heads See Images? Multimodal Retrieval Heads in Long-Context Vision-Language Models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2605.27243.

Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Helen Li, and Yiran Chen. SpeechPrune: Context-aware token pruning for speech information retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025. doi: 10.1109/ICME59968.2025.11209113. URL https://doi.org/10.1109/ICME59968.2025.11209113.

Yuping Lin, Zitao Li, Yue Xing, Pengfei He, Yingqian Cui, Yaliang Li, Bolin Ding, Jingren Zhou, and Jiliang Tang. Retrieval Heads are Dynamic. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15710–15729. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.715. URL https://aclanthology.org/2026.acl-long.715/.

Yujian Ma, Jinqiu Sang, Ruizhe Li, Jiaao Yu, and Ang Li. From Semantics to Readout: Mechanistic Understanding of Audio Tokens after Fine-Tuning for Temporal Audio Grounding. *arXiv preprint arXiv:2607.25355*, 2026. URL https://arxiv.org/abs/2607.25355.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Wang Yuping, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2252. URL https://papers.nips.cc/paper_files/paper/2025/hash/610a7d6507d55be70c6d13d0b663227d-Abstract-Datasets_and_Benchmarks_Track.html.

Videet Mehta, Liming Wang, Hilde Kuehne, Rogerio Feris, James R. Glass, and M. Jehanzeb Mirza. CALM: Class-Conditional Sparse Attention Vectors for Large Audio-Language Models. *arXiv preprint arXiv:2602.07077*, 2026. URL https://arxiv.org/abs/2602.07077.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better than One? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. PSLM: Parallel generation of text and speech with LLMs for low-latency spoken dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2692–2700, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.151. URL https://aclanthology.org/2024.findings-emnlp.151/.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://ieeexplore.ieee.org/document/7178964.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d36f208919582785db965fe648b9fe59-Abstract-Conference.html.

Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Wish Suharitdamrong, Muhammad Awais, Xiatian Zhu, and Sara Atito. From Senses to Decisions: The Information Flow of Auditory and Visual Perception in Multimodal LLMs. *arXiv preprint arXiv:2606.10147*, 2026. URL https://arxiv.org/abs/2606.10147.

Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpotSound: Enhancing Large Audio-Language Models with Fine-Grained Temporal Grounding. *arXiv preprint arXiv:2604.13023*, 2026. URL https://arxiv.org/abs/2604.13023.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/5a18e133cbf9f257297f410bb7eca942-Abstract.html.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lyu, Wei Xue, and Yike Guo. Audio-Omni: Extending Multi-modal Understanding to Versatile Audio Generation and Editing. *arXiv preprint arXiv:2604.10708*, 2026. URL https://arxiv.org/abs/2604.10708.

Junyu Wang, Ziyang Ma, Zhengding Luo, Tianrui Wang, Meng Ge, Xiaobao Wang, and Longbiao Wang. Pay More Attention To Audio: Mitigating Imbalance of Cross-Modal Attention in Large Audio Language Models. *arXiv preprint arXiv:2509.18816*, 2025. URL https://arxiv.org/abs/2509.18816.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Boyong Wu, Chao Yan, Chen Hu, et al. Step-Audio 2 Technical Report. *arXiv preprint arXiv:2507.16632*, 2025a. URL https://arxiv.org/abs/2507.16632.

Qi Wu, Jianfeng Qu, Ximing Li, and Zhixu Li. Expert Heads: Robust Evidence Identification for Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2eda541827dac9f24f6bdfbd6d1f6e3e-Abstract-Conference.html.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=EytBpUGB1Z.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://ieeexplore.ieee.org/document/10095969.

Xiaomi LLM-Core Team. MiMo-Audio: Audio Language Models are Few-Shot Learners. *arXiv preprint arXiv:2512.23808*, 2025. doi: 10.48550/arXiv.2512.23808. URL https://arxiv.org/abs/2512.23808.

Xinyuan Xie, Shunian Chen, Zhiheng Liu, Yuhao Zhang, Zhiqiang Lv, Liyin Liang, and Benyou Wang. EvA: An Evidence-First Audio Understanding Paradigm for LALMs. *arXiv preprint arXiv:2603.27667*, 2026. URL https://arxiv.org/abs/2603.27667.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 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. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

Chen Yang, Chufan Yu, Hanfu Chen, et al. MOSS-Audio Technical Report. *arXiv preprint arXiv:2606.01802*, 2026. URL https://arxiv.org/abs/2606.01802.

Chih-Kai Yang, Neo Ho, Yi-Jyun Lee, and Hung yi Lee. AudioLens: A Closer Look at Auditory Attribute Perception of Large Audio-Language Models. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434723. URL https://doi.org/10.1109/ASRU65441.2025.11434723.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. HalluAudio: A Comprehensive Benchmark for Hallucination Detection in Large Audio-Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38797–38816. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1797. URL https://aclanthology.org/2026.acl-long.1797/.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026. doi: 10.1609/aaai.v40i41.40820. URL https://ojs.aaai.org/index.php/AAAI/article/view/40820.

A AUDIOEVIDENCE

A.1 DATA CONSTRUCTION AND EVIDENCE ANNOTATION

AUDIOEVIDENCE uses sound events as basic units. By controlling event classes, occurrence counts, and temporal positions, we construct audio question-answering examples with explicit evidence intervals. The construction pipeline comprises event collection and screening, controlled composition, and evidence annotation.

Event clips come from the temporal annotations in AudioSet Strong (Hershey et al., 2021) and the audio extraction task in AudioEdit (Tian et al., 2026), with class labels and source information retained. We merge adjacent same-class events, check clip boundaries and event overlap, use Gemini (Comanici et al., 2025) to standardize text labels, and construct distractor pools at different similarity levels using CLAP audio features (Wu et al., 2023).

Acoustic screening operates on the segments actually used. For P0/P1, we recheck target and distractor energy and replenish affected samples. For P2/P3, we search for high-energy windows of the required duration at 100-ms steps, requiring RMS ≥ -50 dBFS, peak amplitude ≥ -40 dBFS, a zero-valued sample fraction ≤ 0.95, and RMS loss relative to the source clip ≤ 12 dB. This branch additionally requires CLAP similarity to the class center ≥ 0.40, a similarity decrease relative to the source clip ≤ 0.30, and a top-10 similarity rank for the target class.

We place target and distractor events at specified positions to create 10-s or 30-s audio, applying cropping, volume adjustment, and boundary fades as needed. Each event retains its identity, task role, and start and end sample indices. Questions, answer choices, and correct answers are determined by the event configuration.

Evidence intervals are annotated directly from questions and event records, independently of model attention or predictions. Evidence-grounded supervision is also explored in EvA (Xie et al., 2026).

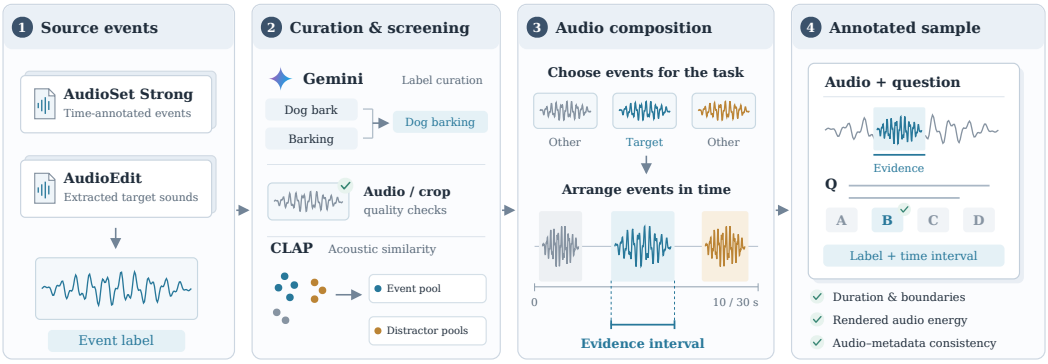


Figure 6: AUDIOEVIDENCE construction: event collection, screening, controlled composition, and evidence annotation. Waveforms and similarity clusters are schematic.

Table 2: Tasks, answer choices, and evidence scope in AUDIOEVIDENCE.

Task	Question	Answer choices	Evidence
P0 Retrieve	Which sound follows the cue?	Four sound classes	Target event after the cue
P1 Verify	Which candidate sound occurs in the audio?	Four sound classes	Target event
P2 Relate	Which of two specified sounds occurs first?	Two sound classes	Both event intervals
P3 Aggregate	How many times does the specified sound occur?	Counts 0, 1, 2, or 3	All occurrences of that sound

After saving, we read the audio back to verify its format, sample count, numerical validity, and event boundaries. Final evidence segments, excluding the cue, must have RMS $> 10^{-4}$; cue and distractor segments must have RMS $> 10^{-5}$, measured in linear full-scale amplitude. These checks apply to event intervals and preserve intentional silent gaps.

A.2 TASK DESIGN

AUDIOEVIDENCE includes four tasks that share the construction and annotation procedures but require different evidence (Table 2).

P0 uses a 0.30-s cue, with the target event beginning 0.15 s after the cue ends. The cue is annotated separately and excluded from the evidence interval. P1 uses no cue. P2/P3 use five non-overlapping event slots, with answers determined by event order and occurrence count, respectively. For multiple relevant events, evidence is the union of their intervals, excluding unrelated sounds and intervening silence. Core causal analyses use P0; the remaining tasks support cross-task extensions. Related temporal-grounding benchmarks evaluate short-event localization in longer audio (Sun et al., 2026).

A.3 COUNTERFACTUALS AND CONTROLS

Counterfactual pairs. For a P0 example A, we replace the target sound within the evidence interval with a sound from another valid answer class, producing example B with a different correct answer. Replacement selection does not depend on model outputs. The question, answer choices and their order, cue, total duration, and evidence location and duration remain fixed; waveforms outside the evidence interval are identical sample by sample.

We use the maximum-energy window of the required duration from the replacement sound, apply 20-ms fades at both ends, and adjust its amplitude toward the original evidence RMS subject to

a peak cap of 0.90. After saving, we independently verify that the waveform changes inside the evidence interval, remains identical outside it, and has the updated correct answer.

Non-evidence. We select non-evidence positions N from the same audio, preferably near a distractor event, such that they are disjoint from evidence positions E and equal in number:

$$E \cap N = \varnothing, \qquad |E| = |N|.$$

Comparing interventions at these positions tests evidence specificity.

Question-conditioning controls. **Full** retains the original question; **Null** replaces only the question with `N/A`, leaving the other inputs and correct answer unchanged. This comparison measures evidence attention with and without the original question. Null is distinct from silent audio input and answer-changing counterfactual examples.

A.4 Data Splits and Sample Counts

Data are divided into discovery and evaluation sets, used to identify head sets or layer intervals and to evaluate intervention effects, respectively.

For the core 10-s datasets, each task (P0–P3) contains 480 discovery and 480 evaluation questions. The P0 counterfactual set contains 480 discovery and 480 evaluation pairs, with the latter covering the 480 questions used in cross-task evaluation. Full/Null conditions share audio and are not counted separately. P3 offers counts 0–3, while correct counts in the current 10-s examples are 1 or 2.

Core causal analyses use 10-s P0 data. For duration comparisons, we extend 480 evaluation pairs to 30 s, preserving each example's sound content, question, choices, and correct answer while repositioning sound blocks and increasing silent gaps. State replacement uses the full A/B pairs; access blocking uses their A inputs. Matched non-evidence positions N are selected separately at each duration. Experimental coverage is detailed in the corresponding appendices.

B Models and Experimental Setup

B.1 Model Selection and Architectural Coverage

We study nine LALMs whose internal states are accessible for intervention, covering discrete audio codes, continuous audio features, hybrid representations, and direct waveform inputs. These models differ in audio compression, backbone depth and size, dense versus mixture-of-experts architectures, and training scale (Table 3). This coverage allows us to examine shared patterns of evidence use and their model-specific implementations. These cross-model observations provide mechanistic insights to guide future studies of more efficient audio–language learning and model scaling.

B.2 Representation Positions and Interventions

Representations and positions. We retain each model's native audio frontend and prompt format. Pre-LLM representations are the outputs of the audio frontend and input interface immediately before the first language-backbone block. The LLM backbone refers to the subsequent language-model blocks that process the audio and text sequence.

We map annotated evidence intervals to each model's native audio positions, denoted by E, with matched non-evidence positions denoted by N. Interval starts are rounded down and ends up. For chunked inputs without explicit chunk timestamps, duration is allocated in proportion to each chunk's number of audio positions. This mapping defines intervention locations; it does not assume that a representation contains information only from its corresponding time window.

Throughout the paper, T_{All} (*all text positions*) contains every input position after the final audio token, through the end of the input sequence. This includes the question, choices, and any intervening template positions; earlier system prompts and context are excluded. Analyses restricted to question positions are identified separately.

Table 3: Model architectures and publicly reported audio-related training scales. Frame rates count temporal audio positions entering the backbone. In the backbone column, parameter counts are followed by layers / query heads per layer. All backbones are dense except Qwen3-Omni (MoE).

Model	**Audio interface**	**Frames / s**	**Backbone params. Layers / heads**	**Training scale**
MiMo-Audio (Xiaomi LLM-Core Team, 2025)	Discrete RVQ codes; four-frame groups	6.25	7B 36 / 32	>100M hours Main-model pretraining
Qwen2-Audio (Chu et al., 2024)	Continuous; pooling and projection	25	7B 32 / 32	∼520K hours
Qwen2.5-Omni (Xu et al., 2025a)	Continuous features into Thinker	25	7B 28 / 28	Stage 2: 300B audio-related tokens + 100B video–audio-related tokens
Step-Audio-2-mini (Wu et al., 2025a)	Continuous; downsampling adapter	12.5	7B 28 / 28	Step-Audio 2 dataset: 8M hours, real + synthetic
Audio Flamingo 3 (Ghosh et al., 2025)	Continuous; pooling and adapter	25	7B 28 / 28	∼50M audio–text pairs
Kimi-Audio (KimiTeam et al., 2025)	Continuous features + discrete semantic	12.5	7B 28 / 28	∼13M hours
MOSS-Audio (Yang et al., 2026)	Continuous features + DeepStack injection	12.5	8B 36 / 32	Millions of audio hours; 1.2T mixed tokens, including text-only data
Qwen3-Omni (Xu et al., 2025b)	Continuous AuT features into Thinker	12.5	30B total 3B active 48 / 32	Encoder: 20M hours Main-model Stage 2: 0.77T audio-related tokens
Gemma 4 (Gemma Team et al., 2026)	Direct waveform-patch projection; encoder-free	25	12B 48 / 16	Audio-only volume not disclosed; speech-focused evaluation

MiMo emits discrete codes at 25 frames/s and groups four frames into each backbone position, yielding 6.25 positions/s. Training scales retain the units and training stages reported by their respective sources.

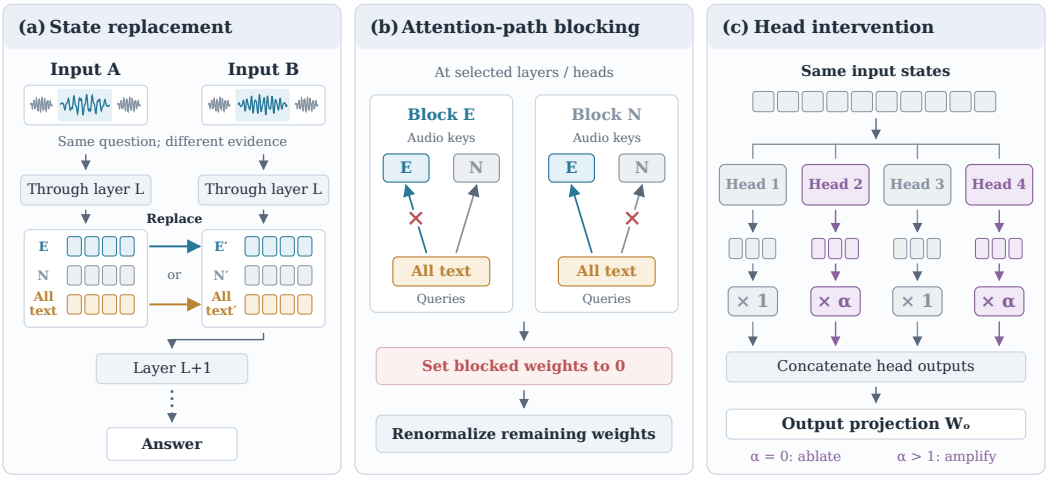


Figure 7: Intervention methods. (a) States at evidence or all text positions are replaced from run A into run B after layer L. The two arrows indicate alternative interventions; primes distinguish states in run B. (b) Attention from all text queries to evidence E or non-evidence N is blocked in selected layers and heads, and the remaining allowed weights are renormalized. (c) Selected head outputs are scaled before the output projection; $\alpha = 0$ ablates and $\alpha > 1$ amplifies them.

State replacement (Figure 7(a)). We use activation replacement to assess causal effects (Meng et al., 2022; Zhang & Nanda, 2024). For counterfactual examples A and B with the same question and choices but different evidence events and correct answers, we compute the hidden states at the output of backbone layer L in both runs. We then replace B's states at evidence positions or all text positions with the corresponding states from A, and continue computation from layer $L + 1$. All other positions retain B's states at the replacement boundary. Changes in answer preference measure the causal effects of states at different positions and depths. Pre-LLM experiments apply the same operation before the first backbone block.

Attention-path blocking (Figure 7(b)). Following attention-knockout approaches (Suharitdamrong et al., 2026), we block attention from all text positions to evidence positions E or non-evidence positions N at selected layers and heads. Specifically, the corresponding attention logits are set to $-\infty$ before softmax, giving these connections zero weight and renormalizing the remaining allowed connections. Native causal and local attention masks are preserved. Here, text→evidence denotes text positions as queries and evidence positions as keys and values, testing the dependence of text computation on direct access to auditory evidence.

Head intervention (Figure 7(c)). We intervene on individual head outputs, as in prior head-ablation and steering analyses (Michel et al., 2019; Glazer et al., 2026). Before the multi-head attention output projection, we multiply selected head outputs by a coefficient α: $\alpha = 0$ ablates them and $\alpha > 1$ amplifies them. Unselected heads are unchanged. Whereas path blocking removes connections between specified positions, this operation scales the selected heads' outputs. Head selection, affected positions, and coefficients are specified for each experiment.

Model-specific interfaces. For Kimi-Audio, pre-LLM replacement occurs after discrete and continuous features are fused. For MOSS-Audio, replacement at the backbone entrance does not modify later DeepStack injections, which retain the inputs from run B. Backbone interventions in Qwen3-Omni operate on Thinker, while Gemma 4 retains its native local and global attention structure. Layer intervals and joint-intervention designs are described in the corresponding experimental appendices.

B.3 Evaluation Metrics and Statistical Analysis

Answer scoring. AudioEvidence and the out-of-distribution benchmarks MMAU, MMAR, AHa-Bench, and HalluAudio are organized as questions with letter-labeled answer choices. We use deterministic option scoring without random sampling: at the first answer position, we compare only the logits of the valid option labels and select the highest-scoring option. Each label is verified to be a single token in its actual prompt context.

We use both accuracy and answer margins. Accuracy measures whether the final choice is correct, whereas margins capture changes in answer preference even when an intervention does not change the selected option. Let $\mathcal{C}$ be the valid option set and $z_c(X)$ the logit of option c in run X. For a correct answer y, behavioral and access-blocking analyses use the margin against the highest-scoring alternative:

$$m_y(X) = z_y(X) - \max_{c \in \mathcal{C} \setminus \{y\}} z_c(X). \tag{1}$$

For counterfactual state replacement within the backbone, let a and b be the correct answers for examples A and B, respectively. We use the margin between this fixed answer pair:

$$m_{a,b}(X) = z_a(X) - z_b(X). \tag{2}$$

Writing $\widetilde{B}$ for the modified run of B, the replacement effect is

$$\delta = m_{a,b}(\widetilde{B}) - m_{a,b}(B). \tag{3}$$

Pre-LLM localization experiments instead use the margin of a against the highest-scoring alternative, as specified in that experimental appendix. Since logit scales differ across models, raw margins are primarily used for comparisons within each model.

ASR. ASR experiments follow the audio processing, prompt format, and evaluation procedures of MiMo-Audio-Eval (Xiaomi LLM-Core Team, 2025). Transcriptions are generated with greedy decoding and evaluated using word error rate (WER).

Confidence intervals. We estimate 95% confidence intervals for mean intervention effects using paired bootstrap resampling (Efron, 1979). We draw examples or counterfactual pairs with replacement, retaining the original sample count and keeping all compared conditions for each sampled unit together. We recompute the mean effect for each resample and use the 2.5th and 97.5th percentiles of the resulting distribution as the interval endpoints. These intervals reflect uncertainty due to the evaluation sample, rather than stochastic generation.

Core state and pathway analyses use 5,000 resamples. Pre-LLM localization analyses use 2,000 resamples, jointly resampling the position variants of each event composition. Analyses involving repeated speech scenarios resample at the scenario level. Accuracy differences are reported in percentage points.

Confidence intervals on depth curves are pointwise intervals, not simultaneous bands for the full curve. Prespecified primary comparisons use Holm correction for multiple testing (Holm, 1979); the comparison families are specified in the corresponding experimental appendices.

C Evidence Localization in Pre-LLM Representations

This appendix supplements Section 4 with the event-relocation construction, sliding-window replacement protocol, and results at a wider window width.

C.1 Controlled Event Relocation

We select 120 event compositions from the P1 discovery split of AudioEvidence, balanced across the four original answer labels (30 per label). Each composition contains an original input A and a counterfactual input B. They share the question and answer choices, but contain sounds corresponding to different correct options within the evidence interval. The waveform outside this interval is identical.

We place the same A/B evidence-event pair at $[1, 4]$, $[4, 7]$, and $[7, 10]$ seconds in a 10-s waveform. The event waveforms, question, choices, and distractor are held fixed across placements. The distractor occupies $[0.2, 0.8]$ seconds and does not overlap the evidence; no cue sound is added. Each composition therefore produces three placement conditions, yielding 360 A/B pairs and 720 audio examples.

Within each placement condition, A/B state replacement measures answer-related causal effects. Across the three conditions, we test whether the recovery region follows the movement of the same event pair. All replacements are between A and B at the same placement.

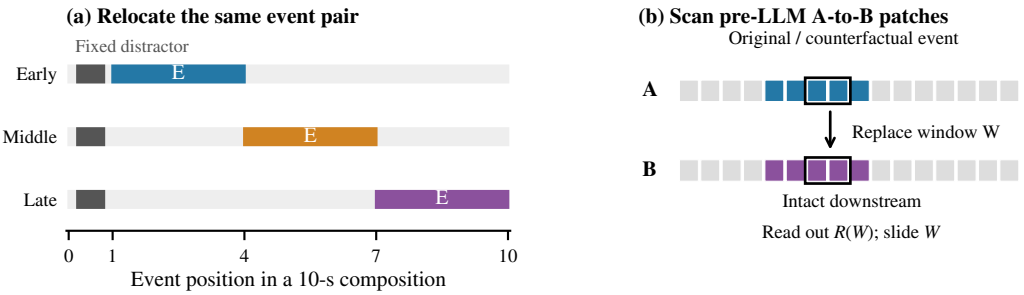


Figure 8: Controlled event relocation and pre-LLM state replacement. Left: the same A/B evidence-event pair is placed at three temporal positions while the distractor remains fixed. Right: A's states within a sliding window W replace B's corresponding states; the backbone then continues to compute $R(W)$.

Input A reuses the rendered event crop from the source P1 example. For B, a sound from the category corresponding to another valid option is cropped, faded, and amplitude-adjusted, then placed in the same interval without looping or time stretching. General construction procedures are described in Appendix A.1.

All nine models use the same compositions, and all examples are included in the analysis. This P1 relocation set is used for pre-LLM localization and is separate from the P0 counterfactual data used in the subsequent backbone experiments.

C.2 SLIDING-WINDOW STATE REPLACEMENT

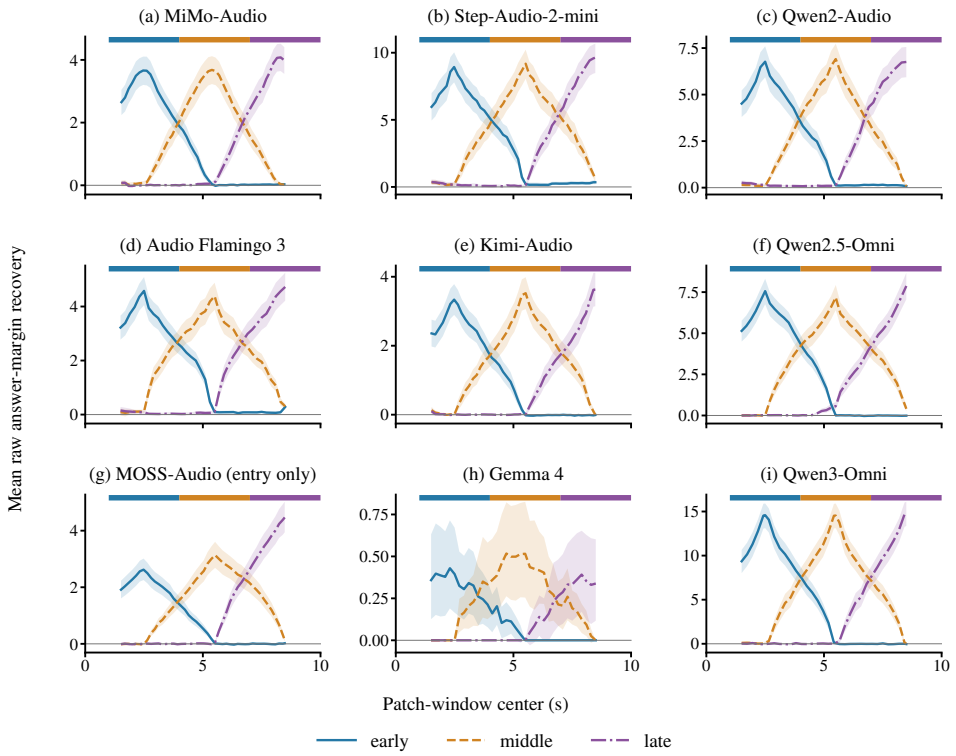


Figure 9: Pre-LLM state-replacement curves using nominal 3-s windows across nine models. Colored bars mark evidence intervals; shading shows 95% pointwise confidence intervals. Vertical scales vary by model. MOSS replaces only the primary backbone input.

We apply the state replacement described in Appendix B.2 at the interface between each model's audio frontend and language backbone. For a window W, we replace B's states with A's corresponding states inside the window, leave all other positions unchanged at the replacement step, and then run the backbone. Each window is evaluated separately, starting from the unmodified states of B.

The primary scan uses a nominal window width of 400 ms; the supplementary scan uses an approximately 3-s window matching the evidence interval. Both use a nominal stride of 200 ms. Window widths and starts are rounded to each model's native audio-position grid, and only complete windows are scanned. The evidence-width scan also includes a window aligned with the evidence onset. Actual window widths and scan-point counts may therefore differ slightly across models; all results are plotted against audio time.

We use $R(W)$ from Section 4: the change caused by replacement in the logit margin of A's correct answer over the strongest other option. Positive values indicate a shift toward A's answer. Curves show raw, unnormalized margin changes, with a separate vertical scale for each model to compare recovery across temporal positions within that model.

Each placement-specific curve averages all 120 compositions, without filtering by baseline correctness or the sign of the A/B margin gap. We obtain 95% confidence intervals from 2,000 bootstrap resamples of compositions, keeping the three placements of each composition together. Shading shows pointwise intervals; the general statistical procedure is described in Appendix B.3.

Implementation checks cover native-computation replay, empty and self replacements, and restoration of B's baseline after scanning. For MOSS-Audio, we replace only the primary audio entry; subsequent DeepStack injections retain B's inputs. Qwen3-Omni is scored using legal-option logits at the first answer position.

C.3 Window Width and Gemma 4 Results

Figure 3 in the main text shows the 400-ms scan, which measures local replacement effects across temporal positions. Figure 9 provides 3-s scans for all nine models, measuring answer changes when replacement covers a more complete event span.

With evidence-width windows, regions of stronger recovery also follow the event's temporal position. Short windows probe local sensitivity, whereas wider windows replace more positions at once and can produce larger recovery effects. The shape of the wider-window curves also depends on how much the window overlaps the evidence, reflecting the distribution of answer effects at this replacement scale.

Gemma 4 also shows recovery concentrated near the evidence interval, with the recovery region following the event's temporal position. On this non-speech task, its raw-margin changes are small, with substantial uncertainty relative to their magnitude. Option-readout accuracies are 29.4% and 20.8% on the original and counterfactual inputs, respectively, against a 25% chance level. The corresponding speech analysis is reported in Appendix H.2.

D Attention-Head Ranking and Ablation Results

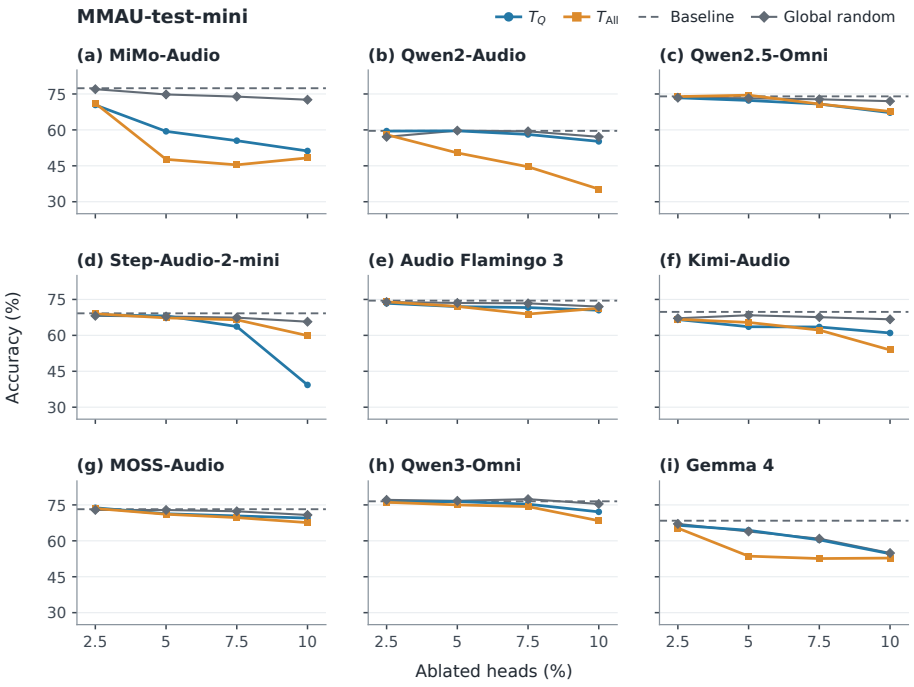


Figure 10: MMAU-test-mini accuracy across head-ablation budgets for nine models. Blue and orange curves show T_Q and T_{All} rankings, gray curves show mean global random accuracy, and dashed lines mark baseline accuracy.

D.1 Attention-Head Ranking and Ablation Settings

Ranking data and query scopes. We rank attention heads using 480 P0 discovery examples from AudioEvidence. For each example, we compute evidence attention under the Full and Null conditions: Full uses the original question, whereas Null replaces it with `N/A`, keeping the audio and answer choices unchanged. Evidence positions E are obtained by mapping temporal annotations to each model's native audio positions. T_Q contains the question tokens; T_{All} contains all input positions after the final audio token, through the end of the input sequence, including any intervening template positions. Full and Null each use their corresponding query-position sets.

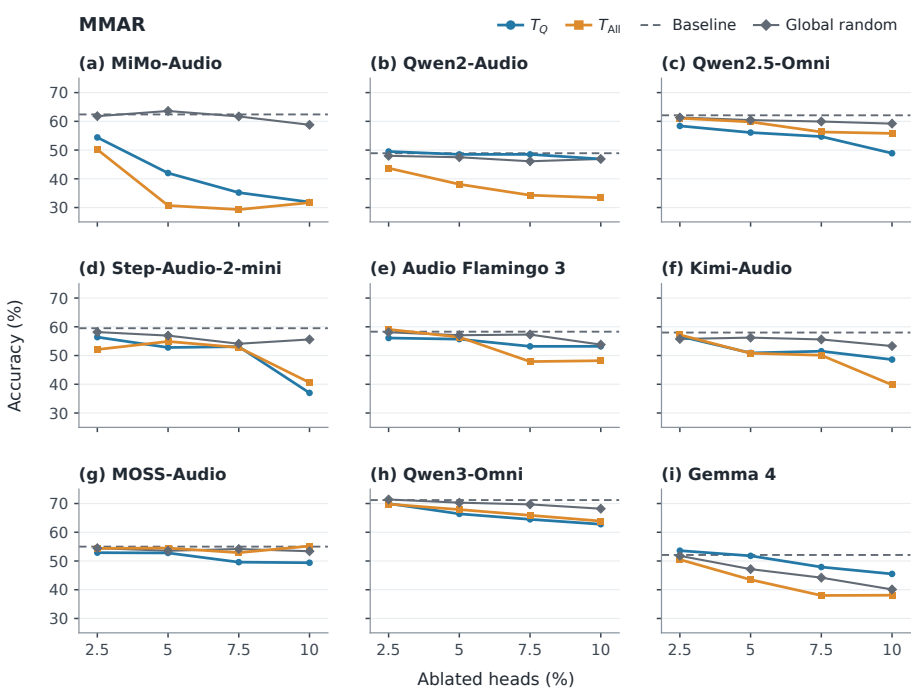


Figure 11: MMAR accuracy across head-ablation budgets for nine models. Colors and line styles follow Figure 10. Reasoning is disabled during evaluation.

Global ranking and budgets. Following the definition in Section 5.1, we sum attention over evidence positions, average over query positions, and then average the Full–Null difference over the 480 examples to obtain G_E for each head. For each query scope, we rank heads across all backbone layers in descending order of G_E. Ablation budgets are 2.5%, 5%, 7.5%, and 10% of the total number of query heads in the model, with the corresponding head count K rounded to the nearest integer. Each budget selects the top K heads from the same ranking, so smaller-budget head sets are nested within larger-budget sets. Rankings are fixed before downstream evaluation and used for both MMAU and MMAR.

Ablation and random controls. Using the head intervention defined in Appendix B, we zero the outputs of selected heads at all token positions before the attention output projection. For each model and budget, global random ablation samples the same number of heads uniformly without replacement from all query heads in the model. Sampling is independent of attention scores, and random sets may overlap with ranked sets. Targeted ablations under the two query scopes share the random control at each budget.

Evaluation and reporting. We use 1,000 questions each from MMAU-test-mini (Sakshi et al., 2025) and MMAR (Ma et al., 2025). MMAR serves as an additional out-of-distribution evaluation benchmark, with reasoning mode disabled. Questions, answer choices, audio inputs, and model prompt formats are kept the same across baseline, targeted, and random conditions. Appendix D.2 reports baseline accuracy, targeted accuracy under both rankings, and global random accuracy. Random curves show mean accuracy across the available seeds.

D.2 MMAU and MMAR Results across Ablation Budgets

We fix head rankings obtained from the AudioEvidence P0 discovery data and compare ablation results across nine models on MMAU-test-mini and MMAR. Figures 10 and 11 show accuracy under both query scopes at budgets of 2.5%, 5%, 7.5%, and 10%, alongside baseline and global random controls.

For MiMo-Audio, targeted ablation causes larger accuracy drops than random ablation on both benchmarks, under both query scopes, and at all four budgets. At the 5% budget, MMAU accuracy is 59.4% under T_Q ablation and 47.7% under T_{All} ablation, compared with 74.8% for random ablation. The corresponding MMAR accuracies are 42.0%, 30.7%, and 63.6%. These results show that a small subset of MiMo heads identified through a controlled auditory evidence task makes a substantial behavioral contribution on both downstream audio benchmarks.

For the other models, targeted ablation produces only a modest gap from random ablation in most conditions, while the few larger effects depend more strongly on query scope and budget. For example, Qwen2-Audio's T_Q curve stays close to random, whereas T_{All} ablation causes greater degradation on both benchmarks. For Step-Audio-2-mini, the effect of targeted ablation becomes markedly stronger at the 10% budget. These complete curves extend the 5% comparison in the main text, showing how each model's dependence on the selected heads varies with the scope of ablation.

E Causal Analysis of Audio–Language Integration

E.1 State Interventions

Setup. We perform state replacement in nine models using 480 pairs of 10-s P0 counterfactual evaluation examples. Each pair A, B shares the question, answer choices, and waveform outside the evidence interval, but differs in the auditory evidence and correct answer. At 11 layer positions uniformly spaced across backbone depth, we replace the auditory evidence states, matched non-evidence states, or all text states in B with the corresponding states from A, then continue computing the answer. Replacement occurs at the selected backbone layer's output; all other positions retain their states from B.

Let a and b be the correct answers for A and B, respectively. We define answer preference as

$$m(X) = z_a(X) - z_b(X),$$

where z denotes a valid answer option's logit. The replacement effect is

$$\delta = m(\widetilde{B}) - m(B),$$

where $\widetilde{B}$ is B after state replacement. Larger δ indicates a stronger shift toward A's answer. Figure 12 shows the full depth curves for all three position sets, with paired-bootstrap 95% confidence intervals.

Results. The eight non-Gemma models share a depth pattern: auditory evidence-state replacement has stronger effects early and weaker effects later, while all text-state effects increase with depth. The transition in Qwen3-Omni occurs later and spans a wider range. Matched N curves provide a reference using the same number of non-evidence positions. Related work on audio-visual LLMs identifies sink tokens that store integrated cross-modal information (Jung et al., 2026).

Speech state and access results, including those for Gemma 4, are reported in Appendix H.

E.2 Critical Evidence-Access Intervals

Prefix/Suffix localization. Using 480 P0 discovery examples, we control direct attention access from all text tokens to auditory evidence tokens. At selected layers, we set the attention logits from these queries to E to $-\infty$ before softmax, which renormalizes attention over the remaining accessible positions. This intervention applies to all query heads in each selected layer.

Let the backbone contain L layers, indexed from zero. For a boundary b, Prefix permits this access only before layer b, whereas Suffix permits it only at or after layer b. Moving b and measuring answer-margin recovery reveals how far early access must continue and how late access can begin.

Here, the margin compares the correct answer with the strongest incorrect option. Let $M(c)$ be the mean margin across examples under condition c. We define recovery as

$$R(c) = \frac{M(c) - M(\text{Block E})}{M(\text{Normal}) - M(\text{Block E})},$$

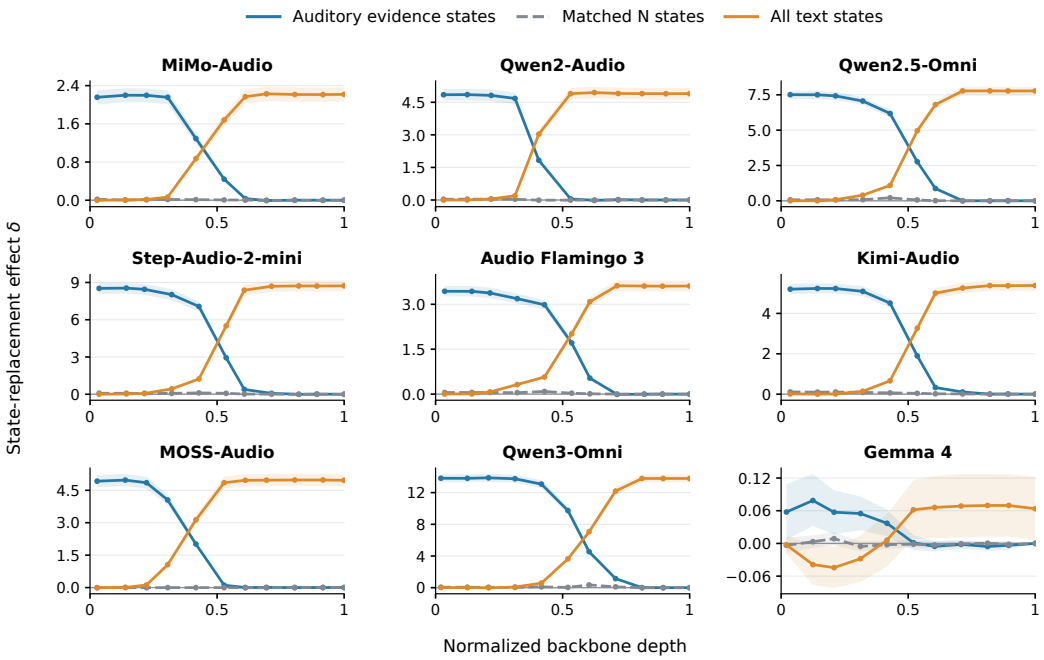


Figure 12: State-replacement effects at auditory evidence, matched non-evidence, and all text positions across nine models on 480 P0 10-s evaluation pairs per model. Bands show pointwise 95% confidence intervals. Vertical scales vary by model.

where Block E removes all text-to-E access throughout the backbone. Thus, $R = 0$ and $R = 1$ correspond to full-depth blocking and normal access, respectively. A model qualifies for interval localization when full-depth E blocking causes a positive margin reduction and a greater reduction than blocking the same number of N positions. Both paired contrasts must have a positive lower bound on their 95% confidence intervals.

For both curves, we identify adjacent measured boundary segments whose recovery ranges intersect 20–80%, and use their outermost boundaries to form the access envelope. Figure 13 shows the localization curves. Table 4 lists the access intervals used for subsequent pathway validation and joint state–pathway interventions.

Model	L	Access interval	Primary probe
MiMo-Audio	36	[13,20)	21
Qwen2-Audio	32	[10,15)	16
Qwen2.5-Omni	28	[12,18)	19
Step-Audio-2-mini	28	[12,16)	16
Audio Flamingo 3	28	[11,18)	19
Kimi-Audio	28	[12,16)	16
MOSS-Audio	36	[9,17)	18
Qwen3-Omni	48	[20,33)	33
Gemma 4	48	—	—

Table 4: Access intervals used for pathway validation and joint state–pathway interventions, with the primary text-state probes. The primary probe is the first prespecified grid layer at or beyond the interval endpoint. Layer indices are zero-based and interval endpoints are exclusive. Gemma 4 does not meet the sound-task endpoint criteria.

Evidence and stage specificity. On an independent set of 480 P0 evaluation examples, we compare four blocking conditions: E within the critical interval, matched N within the same interval,

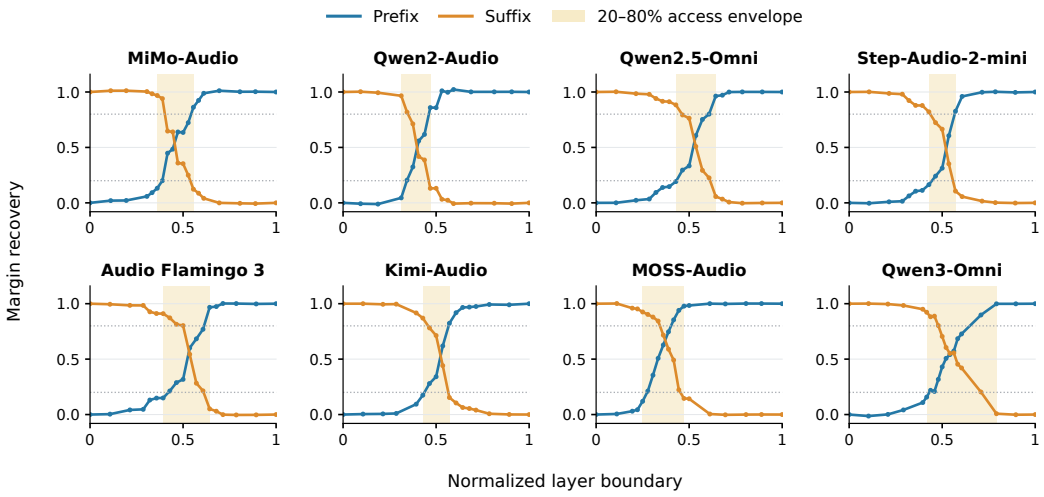


Figure 13: Prefix/Suffix recovery curves for all text-to-auditory-evidence access on 480 P0 discovery examples per model. Shading marks the 20–80% access envelopes; horizontal dotted lines mark the recovery thresholds. Lines connect measured boundaries.

and E within equally wide early and late intervals. If the critical interval has width w, the early and late intervals are $[0, w)$ and $[L - w, L)$, respectively.

In all eight models, blocking E within the critical interval causes a larger margin reduction than each of the other three conditions; all 24 prespecified paired comparisons pass Holm correction. Critical-interval E blocking reduces accuracy by 26.7–39.6 percentage points. These results establish both evidence and stage specificity for intermediate-layer text-to-auditory-evidence access. Figure 14 shows the four margin reductions and their 95% confidence intervals within each model.

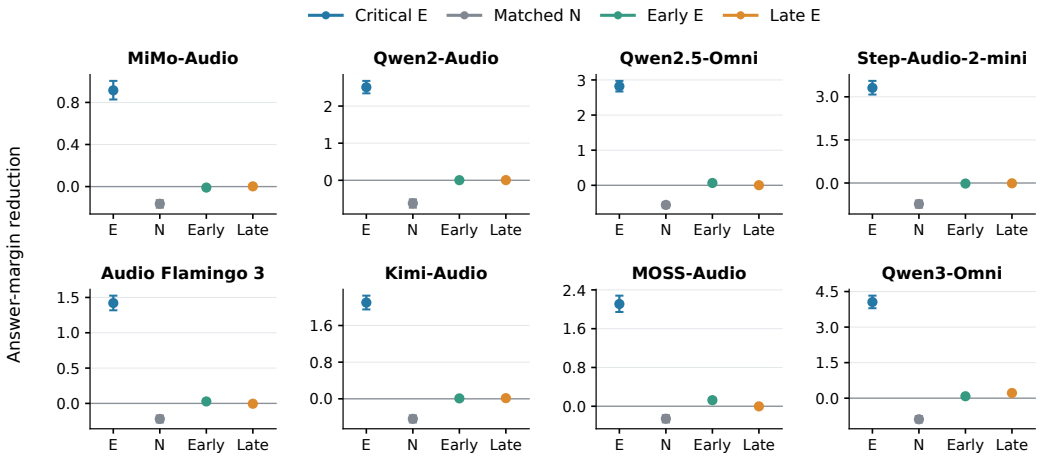


Figure 14: Answer-margin reductions from blocking auditory evidence in the critical interval, matched non-evidence in the same interval, or evidence in equally wide early and late intervals. Points and bars show means and pointwise 95% confidence intervals on 480 P0 evaluation examples per model. Vertical scales vary by model.

E.3 How Evidence Access Affects Text States

We process A under three conditions: normal access, blocking all text-to-E access within the critical interval, and blocking all text-to-N access within that interval. We then extract all text states at the same depth, replace the corresponding states in normal B, and measure δ as defined in Appendix E.1. All three conditions use the same normal B, so their differences arise from the text states formed in A.

At probe layer ℓ, we block only layers in the critical interval that strictly precede ℓ. We retain the 11-layer grid from the state experiment and prespecify the first grid layer at or beyond the interval endpoint as the primary probe. Figure 5 shows the full depth curves; Table 5 reports the primary-probe results.

Model	Normal	Block E	Block N	$N - E$ [95% CI]
MiMo-Audio	2.168	1.286	2.334	1.048 [0.956, 1.143]
Qwen2-Audio	4.896	2.571	5.310	2.740 [2.563, 2.927]
Qwen2.5-Omni	7.777	5.339	8.227	2.887 [2.719, 3.058]
Step-Audio-2-mini	8.376	5.321	8.814	3.493 [3.241, 3.751]
Audio Flamingo 3	3.619	2.197	3.879	1.682 [1.568, 1.802]
Kimi-Audio	5.004	2.974	5.340	2.365 [2.191, 2.544]
MOSS-Audio	4.851	2.714	5.168	2.454 [2.271, 2.647]
Qwen3-Omni	12.194	7.960	12.925	4.966 [4.639, 5.296]

Table 5: All text-state replacement effects at the primary probes, using 480 pairs per model. Values are mean A-versus-B margin changes relative to normal B. The final column reports the paired difference between Block N and Block E, with its 95% confidence interval. All eight comparisons pass Holm correction.

Across all eight models, text states formed under E blocking produce smaller replacement effects than those formed under normal access or N blocking. Together with the state-depth curves, these results show that evidence access within the critical interval enables auditory evidence to be integrated into text states and affect subsequent answers.

E.4 Evidence Access Through Selected Heads

Within-interval head budgets. We reuse the G_E rankings from Section 5.1 and select the highest-ranked heads within each critical access interval, separately for T_Q and T_{All}. Budgets are 5%, 10%, 20%, 40%, and 80% of the query heads within the interval, with head counts rounded to the nearest integer. Smaller sets are nested within larger sets. For each selected set, we separately block all text-to-E and all text-to-N attention, leaving other paths unchanged. The two rankings differ in how heads are selected; both interventions use all text positions as queries. Each of the eight models uses 480 P0 evaluation examples.

We report margin reductions relative to normal inputs separately for E and N blocking. Here, m_{gold} is the correct option's logit minus the strongest incorrect option's logit. To characterize concentration, we compare the mean E-blocking effect of a selected head set H with that of all heads H_I in the interval:

$$D_E(H; X) = m_{\mathrm{gold}}(X) - m_{\mathrm{gold}}(X; \mathrm{block}_E(H)), \qquad C(H) = \frac{\overline{D_E(H)}}{\overline{D_E(H_I)}} \times 100\%.$$

The overbar denotes the mean across evaluation examples. Numerator and denominator use the same examples and blocking operation. This ratio measures relative intervention effects, rather than an additive share of information. We obtain pointwise 95% confidence intervals from 5,000 paired bootstrap resamples, resampling numerator and denominator together for $C(H)$.

Concentration across models. MiMo-Audio shows pronounced concentration at small budgets. Selecting 10% of interval heads (22 of 224 heads after rounding) yields 82.4% and 81.8% of the full-interval E-blocking effect under T_Q and T_{All}, respectively. Qwen2.5-Omni reaches 75.5% and 71.1% at a 20% budget, while Step-Audio-2-mini reaches 61.1% and 58.7% at 10%. Qwen3-Omni

reaches 78.1% and 84.5% at 40%. In Qwen2-Audio, Kimi-Audio, and MOSS-Audio, the T_Q effect increases more gradually with budget. The curves also reveal query-scope dependence and non-monotonic effects in some models. Figure 15 shows the raw margin reductions across budgets, with full-interval E blocking as a reference.

Evidence and non-evidence access. Across eight models, two rankings, and five budgets, 77 of 80 conditions show greater margin damage from E blocking than from N blocking, with positive pointwise 95% confidence intervals and Holm-adjusted $p < 0.05$. The adjustment covers all 80 two-sided paired t-tests. The remaining conditions are Qwen2-Audio with T_{All} at 5%, and Audio Flamingo 3 with T_Q at 5% and 10%.

Blocking matched non-evidence access often increases the correct-answer margin, suggesting that such access can interfere with evidence-based prediction. Figure 15 reports the two effects separately, distinguishing the damage from blocking auditory evidence from the gains associated with blocking non-evidence access.

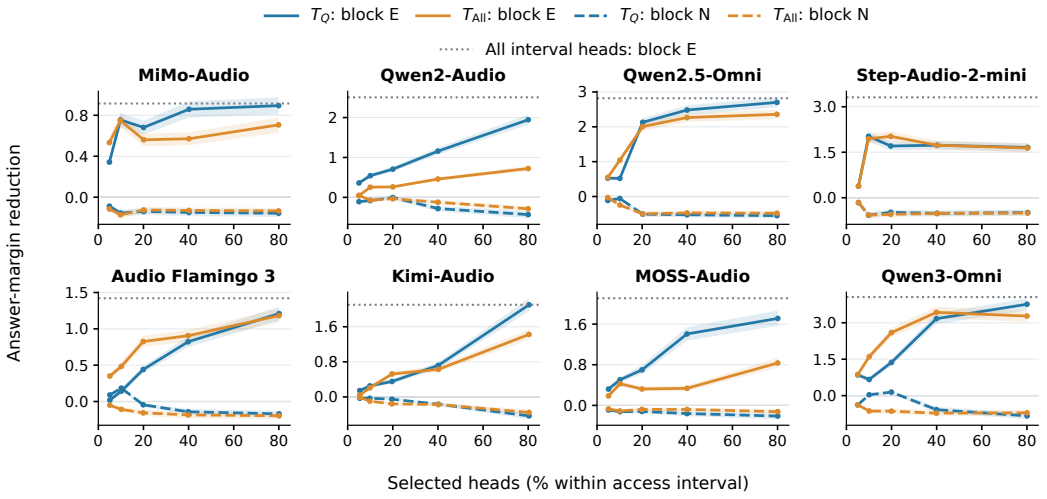


Figure 15: Answer-margin reductions from blocking all text attention to E or matched N in selected heads. The horizontal axis gives the selected fraction of heads within each access interval. Colors denote ranking query scopes; solid and dashed curves show E and N blocking. Dotted lines mark E blocking across all heads in the interval. Bands show pointwise 95% confidence intervals. Negative values indicate increased correct-answer margins. Each model uses 480 P0 examples; vertical scales vary by model.

E.5 Evidence Attention and Critical Access

Figure 16 overlays the mean signed G_E across heads at each layer with the access envelope. MiMo-Audio's high-scoring region aligns closely with its access interval, whereas strong attention in Step-Audio-2-mini, Qwen2-Audio, and MOSS-Audio also extends to later layers. Thus, strong evidence attention does not always identify the most critical access stage.

F Generalization Across Tasks and Audio Durations

F.1 Data and Experimental Setup

For cross-task evaluation, we use 240 10-s examples each from P1 event verification, P2 temporal ordering, and P3 event counting, with the 480 P0 retrieval examples as a reference. P1 uses examples in which the target event is present. For P2, evidence positions E cover both events involved in the temporal comparison; for P3, E covers all occurrences of the target event, with counts of one or two in this set. Matched non-evidence positions N come from the same audio, exclude E, and contain the same number of audio positions.

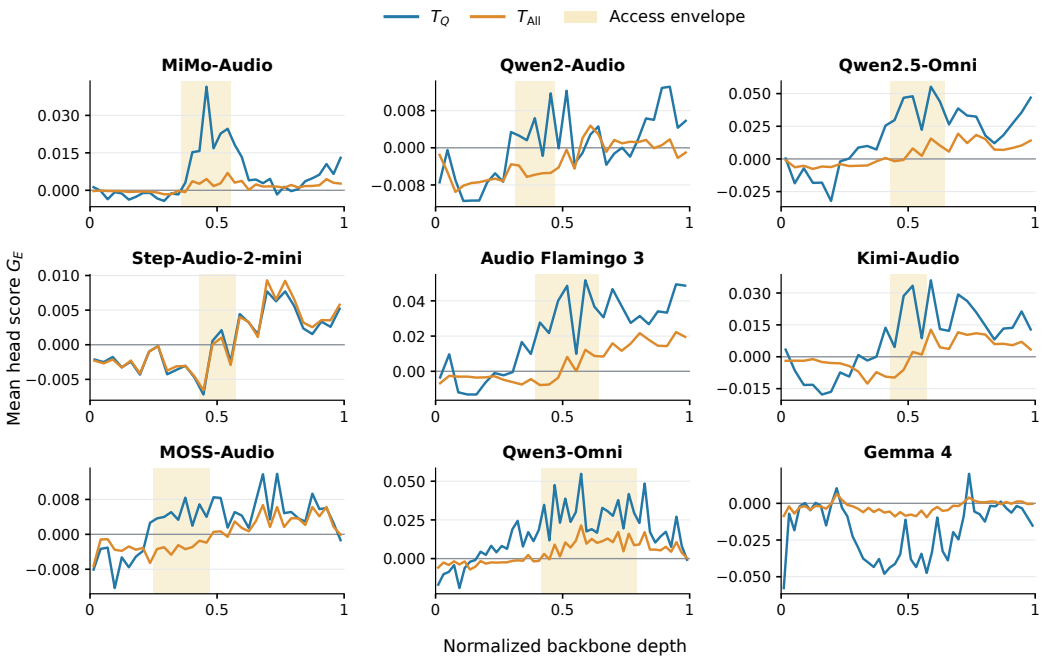


Figure 16: Layer-wise means of signed question-conditioned evidence-attention scores for both query scopes, overlaid with independently located all text access envelopes. Vertical scales vary by model. Gemma 4 has no qualifying access envelope.

For duration generalization, we use paired P0 10-s and 30-s inputs constructed from the same source examples. Each 30-s version preserves the original sound content, question, choices, and correct answer, expanding the temporal span by repositioning sound blocks and adding silent gaps. State replacement covers nine models and 480 counterfactual A/B pairs per duration. Path blocking covers the eight models with reliable P0 access intervals, using the corresponding 480 A inputs at each duration. Matched non-evidence positions N are selected within each duration and need not correspond to the same sound segment across durations.

Path experiments retain the critical access intervals identified on P0 10-s inputs and compare four conditions: blocking attention from all text tokens to E within the critical interval (Critical E), blocking attention to N within the same interval (Matched N), and blocking attention to E within equal-width early or late intervals (Early E and Late E). State experiments retain the 11 backbone layers spaced evenly across depth and replace states at auditory evidence or all text positions.

State effects are measured by the change in the A-versus-B answer margin induced by replacement. Path effects are measured by the reduction in the correct-answer margin relative to normal computation, with accuracy changes reported alongside them. Error bars and shaded bands show paired-bootstrap 95% pointwise confidence intervals. Holm correction is applied to the 72 primary paired margin tests across tasks and, separately, to the 24 tests on 30-s inputs.

F.2 Transfer to Verification, Temporal Ordering, and Counting

Figure 17 compares the effects of the fixed P0 intervals across four auditory tasks. On P1 verification and P2 temporal ordering, Critical E causes larger margin reductions than Matched N, Early E, and Late E in all eight models. All 24 primary comparisons for each task pass correction. Thus, the critical access stage identified through retrieval also supports event verification and temporal ordering.

Counting results depend more strongly on the model. All three primary comparisons support the contribution of critical-interval evidence access in MiMo-Audio, Qwen2.5-Omni, Step-Audio-2-mini, Kimi-Audio, and Qwen3-Omni. Audio Flamingo 3's comparisons with the early and late

intervals, and MOSS-Audio's comparison with the early interval, have the same direction but do not pass multiple-comparison correction. Qwen2-Audio shows the opposite effect: blocking evidence access within the critical interval increases the correct-answer margin by 0.578 and raises accuracy from 12.08% to 20.00%. Overall, 18 of the 24 P3 comparisons are positive and pass correction, showing model-dependent contributions of critical-interval evidence access to correct counting.

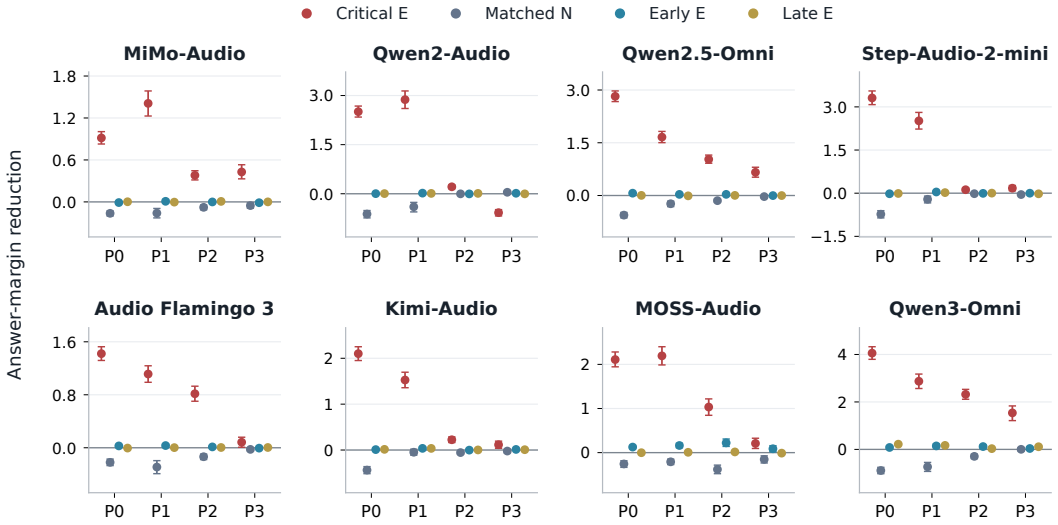


Figure 17: Answer-margin reductions across P0–P3 in eight models. Colors distinguish Critical E, Matched N, Early E, and Late E. Error bars show 95% confidence intervals; vertical scales vary by model. Negative values indicate increased margins after blocking.

F.3 State Patterns and Critical Access Across Audio Durations

Figure 18 shows state-replacement curves for paired 10-s and 30-s inputs. The eight non-Gemma models retain the same overall depth pattern: auditory evidence states have stronger effects early, which diminish as text-state effects increase with depth. Effects are generally smaller for 30-s inputs, while the transitions between the two state effects occur at similar depths. The curve intersections are close in most models, with a larger shift in MOSS-Audio. Gemma 4 shows weaker absolute effects on this P0 sound task.

Figure 19 further tests the transfer of the critical access stage across durations. With the intervals fixed from 10-s inputs, Critical E still produces larger margin reductions than all three controls on 30-s inputs in all eight models, and all 24 primary comparisons pass correction. Blocking evidence access in the critical interval reduces accuracy by 14.58–34.79 percentage points. The original critical access stage therefore remains important for prediction when the audio spans a longer duration.

F.4 Transfer of the Critical Access Stage to External Benchmarks

We further apply the P0 access intervals to MMAU-test-mini (Sakshi et al., 2025) and MMAR (Ma et al., 2025), using 1,000 examples from each benchmark. As these benchmarks lack localized evidence annotations, we block attention from all text tokens to all audio positions and compare the critical interval with equal-width early and late intervals.

On both benchmarks, blocking within the critical interval causes larger mean answer-margin reductions than early or late blocking in all eight models, with all 16 margin comparisons per benchmark passing correction. Figure 20 shows the corresponding accuracy changes: critical-interval blocking reduces accuracy by 3.6–9.6 percentage points on MMAU and 3.8–9.9 points on MMAR. All accuracy comparisons pass correction on MMAU, as do 14 of 16 on MMAR; MOSS-Audio and Qwen3-Omni show weaker differences from early blocking on the latter. These results support the transfer of the critical audio-access stage from controlled retrieval to external benchmarks.

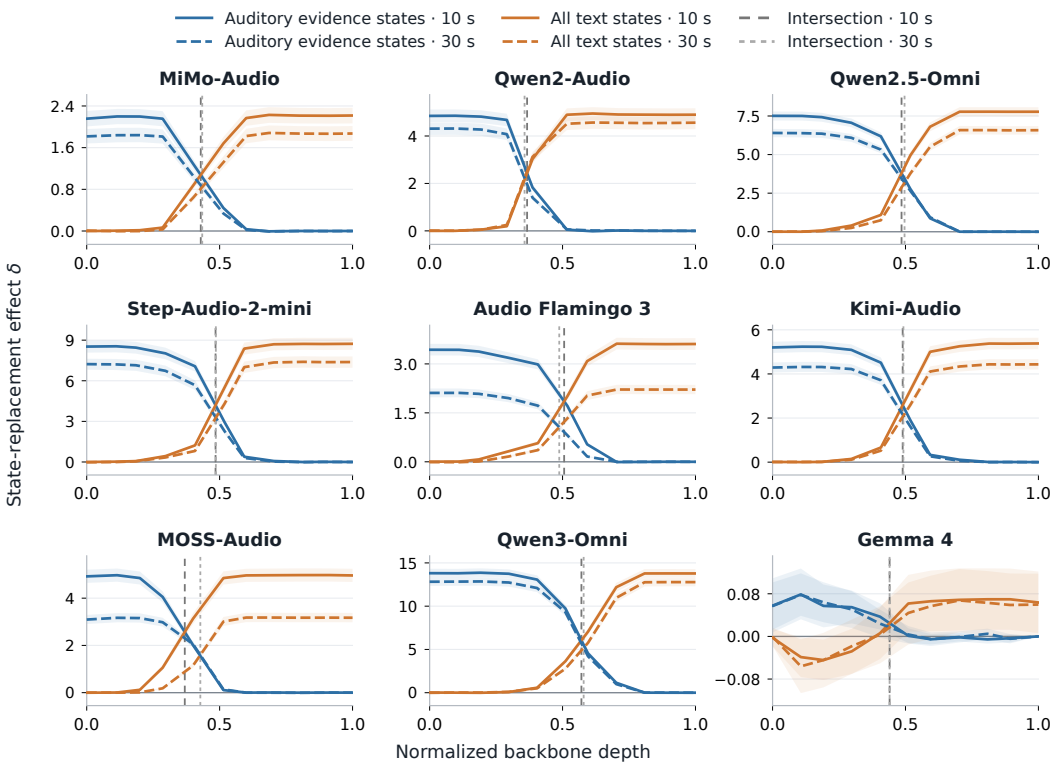


Figure 18: State-replacement curves for paired 10-s and 30-s inputs across nine models. Colors distinguish auditory evidence and all text states; solid and dashed curves denote 10-s and 30-s inputs. Gray vertical lines mark the intersections of the two mean curves at each duration, linearly interpolated between sampled layers. Shading shows 95% confidence intervals; vertical scales vary by model.

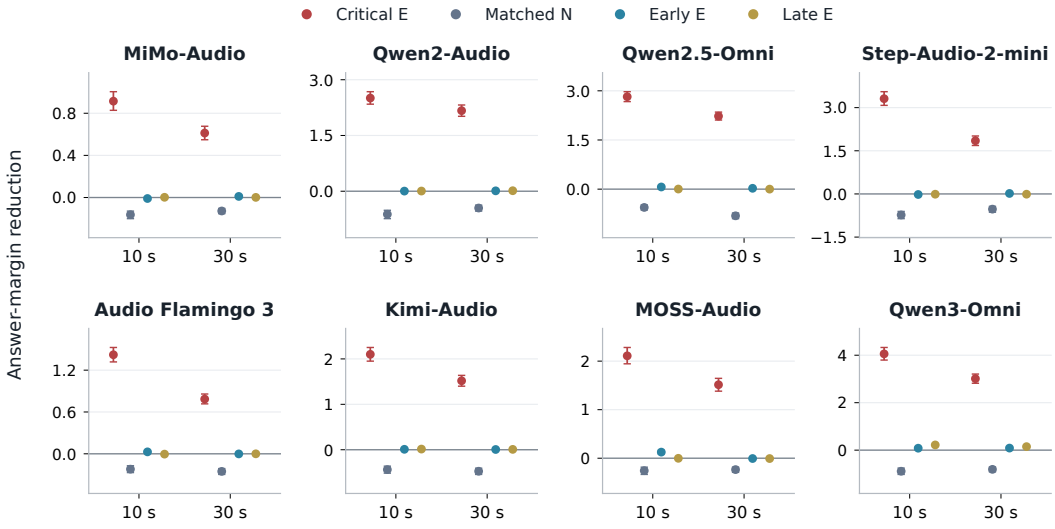


Figure 19: Answer-margin reductions on the A inputs from paired P0 10-s and 30-s examples. Matched N positions are selected separately at each duration. Colors distinguish the four blocking conditions. Error bars show 95% confidence intervals; vertical scales vary by model.

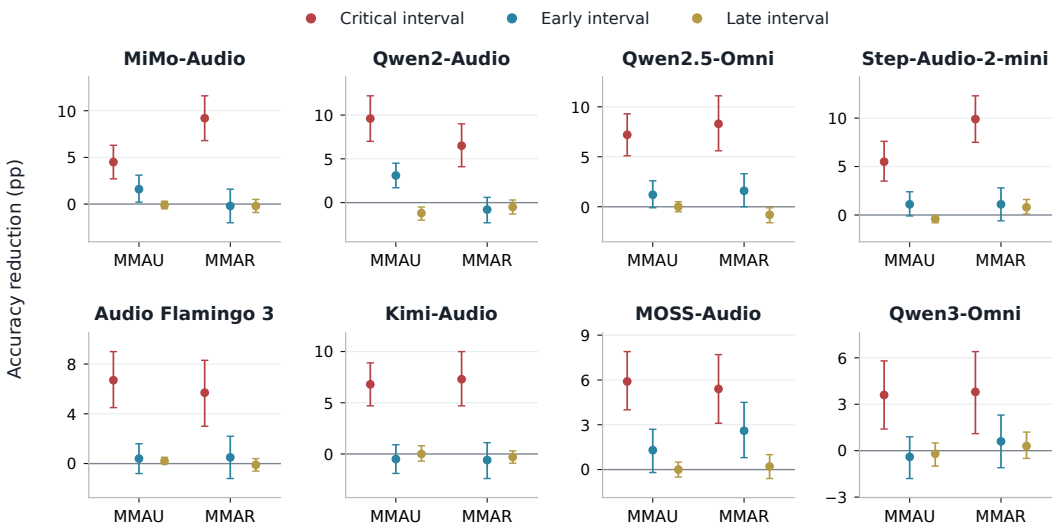


Figure 20: Accuracy reductions on MMAU-test-mini and MMAR from blocking attention from all text tokens to all audio positions in the P0-defined critical interval or equal-width early and late intervals. Error bars show 95% confidence intervals; vertical scales vary by model.

G Inference-Time Control: Supplementary Methods and Results

G.1 Intervention and Calibration

We use MiMo-Audio as the main case study and scale attention-head outputs at inference time. We select heads with the same evidence-attention scoring and global ranking procedure used in the preceding mechanistic experiments, without restricting their layers. For control, we compute the ranking on a larger AudioEvidence discovery sample constructed with the same sampling and screening procedures and drawn from the same distribution as the 480-example P0 sample used for ablation. We fix this ranking before calibration and downstream evaluation.

Let $o_{l,h,t}$ denote the output of head h in layer l at token position t, and let H be the selected head set. Before the attention output projection, we apply

$$\widetilde{o}_{l,h,t} = \begin{cases} \alpha o_{l,h,t}, & (l,h) \in H, \\ o_{l,h,t}, & (l,h) \notin H. \end{cases}$$

Scaling applies at all token positions, leaves model parameters unchanged, and requires no evidence annotations at inference time. The setting $\alpha = 1$ recovers normal computation.

We calibrate on 480 examples, with 120 from each of P0–P3, using $K \in \{20, 40, 60, 80, 100\}$ and $\alpha \in \{0.8, 0.9, 1, 1.05, 1.1, 1.2, 1.3\}$. Among amplification candidates with $\alpha \geq 1$, we select the highest mean accuracy across the four tasks, breaking ties by proximity of α to 1 and then by smaller K. The selected configuration, Top80 with $\alpha = 1.3$, increases calibration correctness from 295/480 to 313/480. We separately calibrate scaling all 224 heads in the interval $[13, 20)$, selecting $\alpha = 1.2$. Both configurations are then fixed for downstream evaluation (Figure 21).

Multiple-choice evaluations compare valid option logits at the first answer position. ASR follows the official English transcription prompt and evaluation pipeline of MiMo-Audio-Eval (Xiaomi LLM-Core Team, 2025), with greedy decoding for both global and local samplers and a 256-step generation cap. We use 5,000 paired cluster-bootstrap resamples for uncertainty: exact-audio SHA256 groups for HalluAudio and AHa, and speakers for ASR. Reported intervals are pointwise 95% confidence intervals without multiple-comparison adjustment.

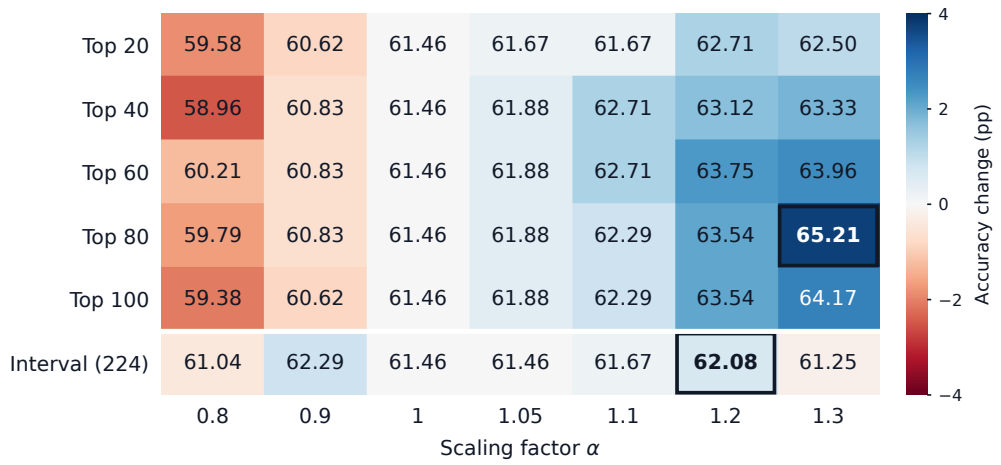


Figure 21: MiMo-Audio calibration accuracy (%) across head sets and scaling factors. Colors show changes from the 61.46% baseline in percentage points; boxes mark selected amplification settings. Values below $\alpha = 1$ are diagnostic and excluded from amplification selection.

G.2 Hallucination Benchmarks and Error Types

We evaluate the fixed configurations on 5,653 HalluAudio (Zhao et al., 2026) questions and 886 AHa-Bench (Cheng et al., 2025) questions, covering 4,336 and 369 audio clusters, respectively. Table 6 reports both Top80 amplification and full-interval scaling. Top80 improves HalluAudio accuracy. The AHa point estimate also increases, although its difference interval includes zero. Full-interval scaling shows no clear gain on either benchmark.

Benchmark	Configuration	Accuracy (%)	Change (pp) [95% CI]
HalluAudio	Baseline	67.999	—
	Top80, $\alpha = 1.3$	69.238	$+1.238$ $[+0.433, +2.096]$
	Interval, $\alpha = 1.2$	67.734	-0.265 $[-0.711, +0.178]$
AHa-Bench	Baseline	56.772	—
	Top80, $\alpha = 1.3$	58.239	$+1.467$ $[-0.227, +3.143]$
	Interval, $\alpha = 1.2$	56.659	-0.113 $[-1.395, +1.094]$

Table 6: MiMo-Audio accuracy with fixed scaling configurations. Changes are intervention minus baseline; interval scaling uses all 224 heads in $[13, 20)$.

We further measure false-positive rates (FPR) and false-negative rates (FNR) on 4,040 HalluAudio yes/no questions and all 886 AHa questions (Table 7). Both benchmarks show fewer false positives and more false negatives. Head amplification thus changes the model's tendency to judge sounds as present, with a trade-off between the two error types.

	FPR (%)		FNR (%)	
Benchmark subset	Baseline	Top80	Baseline	Top80
HalluAudio yes/no	66.54	55.68	3.94	7.75
AHa-Bench	66.52	59.66	17.38	21.90

Table 7: Error-type changes under Top80 scaling with $\alpha = 1.3$. HalluAudio's binary subset contains 4,040 questions, whereas its overall accuracy uses 5,653 questions.

G.3 Directional Effects in an Exploratory AHa Grid

A separate exploratory AHa experiment evaluates 55 configurations: $K \in \{20, 40, 60, 80, 100\}$ and $\alpha = 1.00, 1.05, \ldots, 1.50$, on the same 886 questions. It uses a historical ranking obtained with the question and answer choices omitted, and deterministic first-position full-vocabulary argmax followed by Yes/No parsing. This ranking and readout differ from the fixed-configuration evaluation above, so the grid is reported with its own baseline.

Figure 22 reveals distinct directions of behavioral change. For Top20, increasing α from 1.0 to 1.5 reduces FPR from 34.12% to 27.47% while increasing FNR from 35.95% to 40.48%: fewer absent sounds are accepted as present, but more present sounds are missed. This shifts sound-presence judgments toward a more conservative response tendency. Top40 instead lowers both error rates at $\alpha = 1.4$, to 33.69% FPR and 32.38% FNR. The direction therefore depends on the selected head set as well as scaling strength.

The highest observed grid accuracy is 67.16% at Top80/$\alpha = 1.3$, compared with the grid baseline of 65.01%; its FPR and FNR are 30.47% and 35.48%. Larger scaling factors do not consistently improve accuracy. These same-set search results characterize the response to intervention, rather than an independent test of the best observed configuration. These results demonstrate head-level control over sound-presence judgments, with different head sets enabling different changes in error rates.

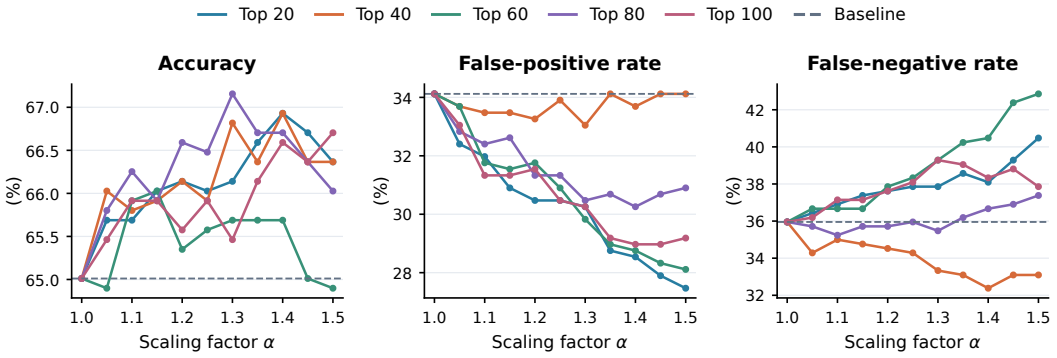


Figure 22: Exploratory AHa accuracy, false-positive rate, and false-negative rate across scaling strengths for five head budgets. Each configuration uses one deterministic run on 886 questions. Dashed lines mark the experiment's own unscaled baseline.

G.4 Speech Recognition and Sources of Improvement

We evaluate the same fixed configurations on the 2,620 utterances from 40 speakers in LibriSpeech test-clean (Panayotov et al., 2015). Our primary metric is corpus WER: total substitutions, deletions, and insertions divided by the total number of reference words. Top80 scaling reduces corpus WER from 3.441% to 2.866% (Table 8), with total word errors decreasing from 1,809 to 1,507. Error counts improve for 68 utterances, worsen for 55, and remain unchanged for 2,497.

Configuration	Corpus WER (%)	Change (pp) [95% CI]
Baseline	3.441	—
Top80, $\alpha = 1.3$	2.866	-0.574 [-1.002, -0.224]
Interval, $\alpha = 1.2$	3.454	$+0.013$ [-0.082, $+0.097$]

Table 8: LibriSpeech test-clean results for MiMo-Audio. Confidence intervals resample the 40 speaker clusters.

A post-hoc examination of outputs shows that 17 baseline utterances containing Chinese characters account for 269 of the 302 net word errors removed (89.1%; Table 9). These cases include Chinese

translations or paraphrases of English speech instead of the requested English transcription. The number of outputs containing Chinese characters decreases from 17 to 6. Among the other 2,603 utterances, WER changes from 2.7589% to 2.6957%, a difference of -0.0633 percentage points with 95% CI [-0.1483, $+0.0133$].

Post-hoc group	n	Baseline errors	Top80 errors	Net reduction
All utterances	2,620	1,809	1,507	302
Baseline contains Chinese characters	17	370	101	269
Remaining utterances	2,603	1,439	1,406	33

Table 9: Word-error changes grouped by baseline output. Groups are defined after evaluation to examine the source of the aggregate ASR gain.

Substitutions change from 1,097 to 1,091, deletions from 571 to 267, and insertions from 141 to 149. Chinese outputs aligned against English references can produce many deletion errors. The overall WER gain is driven primarily by fewer hallucination-related errors and task deviations, particularly translation in place of transcription. Figure 23 provides example outputs.

G.5 Positive Cases in Additional Models

We also evaluate full-interval and interval-restricted Top-K scaling in seven additional models. Table 10 shows the configurations whose HalluAudio accuracy improvements have pointwise 95% intervals above zero. These are positive cases selected from the extension results; other configurations have intervals including zero or show decreases.

Each model uses 120 calibration examples per task from P0–P3, then fixes its configuration for evaluation on 5,653 HalluAudio questions. Full-interval scaling acts on all heads in the interval. For Top-K scaling, heads are ranked within the interval using T_Q or T_{All}, with candidate counts corresponding to 2.5%, 5%, 7.5%, or 10% of all query heads in the model. Configurations are selected among $\alpha \geq 1$ candidates by calibration accuracy, with ties resolved by α closest to 1 and then smaller K.

Model	Selection	K	α	Accuracy (%)	Change (pp) [95% CI]
Qwen2-Audio	Full interval	160	1.3	58.323	$+1.468$ [$+0.963$, $+1.999$]
	Interval T_Q	77	1.3	57.456	$+0.601$ [$+0.072$, $+1.125$]
	Interval T_{All}	77	1.2	57.385	$+0.531$ [$+0.071$, $+0.995$]
Qwen2.5-Omni	Interval T_{All}	39	1.2	69.839	$+0.655$ [$+0.227$, $+1.106$]
Qwen3-Omni	Interval T_{All}	38	1.2	71.944	$+0.601$ [$+0.123$, $+1.098$]

Table 10: Selected positive HalluAudio cases beyond MiMo-Audio. Baseline accuracies are 56.855%, 69.185%, and 71.343% for Qwen2-Audio, Qwen2.5-Omni, and Qwen3-Omni, respectively. Intervals use 5,000 paired audio-cluster bootstrap resamples without multiple-comparison adjustment.

Head-output scaling also improves HalluAudio accuracy in Qwen2-Audio, Qwen2.5-Omni, and Qwen3-Omni under the configurations in Table 10.

G.6 ILLUSTRATIVE ASR EXAMPLES

Figure 23 shows three post-hoc examples from LibriSpeech test-clean: recovery of the requested output language, improved transcription coverage, and a word-segmentation regression. Reference and model outputs are reproduced from the saved evaluation records. Gray, amber, and green/red panels distinguish references, baseline outputs, and scaled outputs, respectively.

(a) Restoring English transcription

LibriSpeech test-clean — 2300-131720-0004

Word errors: 44 → 0 / 44 — **WER: 100.0% → 0.0%**

Reference OWING TO HIS INSISTENCE ON LOW PRESSURE DIRECT CURRENT FOR USE IN DENSELY POPULATED DISTRICTS AS THE ONLY SAFE AND TRULY UNIVERSAL PROFITABLE WAY OF DELIVERING ELECTRICAL ENERGY TO THE CONSUMERS EDISON HAS BEEN FREQUENTLY SPOKEN OF AS AN OPPONENT OF THE ALTERNATING CURRENT

Baseline 因此由于他坚持使用低压直流电作为在人口密集区使用的唯一安全且真正通用盈利的方式向消费者输送电能爱迪生经常被说成是交流电的反对者

Scaled ($K = 80$, $\alpha = 1.3$) Owing to his insistence on low pressure direct current for use in densely populated districts as the only safe and truly universal profitable way of delivering electrical energy to the consumers edison has been frequently spoken of as an opponent of the alternating current

The baseline paraphrases the speech in Chinese; scaling returns the requested English transcription.

(b) Recovering the omitted opening

LibriSpeech test-clean — 908-31957-0015

Word errors: 15 → 1 / 40 — **WER: 37.5% → 2.5%**

Reference THAT WAS THE CHRISM OF LOVE WHICH LOVE'S OWN CROWN WITH SANCTIFYING SWEETNESS DID PRECEDE THE THIRD UPON MY LIPS WAS FOLDED DOWN IN PERFECT PURPLE STATE SINCE WHEN INDEED I HAVE BEEN PROUD AND SAID MY LOVE MY OWN

Baseline The third upon my lips was folded down in perfect purple state since when indeed i have been proud and said my love my own

Scaled ($K = 80$, $\alpha = 1.3$) That was the chalice of love which love's own crown with sanctifying sweetness did precede the third upon my lips was folded down in perfect purple state since when indeed i have been proud and said my love my own

Scaling restores the 15-word opening omitted by the baseline, with one remaining substitution: CHRISM → chalice.

(c) A regression in word segmentation

LibriSpeech test-clean — 1580-141083-0009

Word errors: 0 → 2 / 27 — **WER: 0.0% → 7.4%**

Reference THE ALTERNATIVE WAS THAT **SOMEONE** PASSING HAD OBSERVED THE KEY IN THE DOOR HAD KNOWN THAT I WAS OUT AND HAD ENTERED TO LOOK AT THE PAPERS

Baseline The alternative was that **someone** passing had observed the key in the door had known that i was out and had entered to look at the papers

Scaled ($K = 80$, $\alpha = 1.3$) The alternative was that **some one** passing had observed the key in the door had known that i was out and had entered to look at the papers

Scaling splits "someone" into "some one", producing one substitution and one insertion under word-level scoring.

Figure 23: Reference, baseline, and Top80-scaled outputs for three MiMo-Audio ASR examples. Word-error counts and WER use the full utterances. The same fixed configuration ($K = 80$, $\alpha = 1.3$) is used in all examples.

H SPEECH EXTENSION AND GEMMA 4

H.1 CONTROLLED SPEECH COUNTERFACTUALS

To examine whether the depth patterns observed on sound tasks extend to speech, we construct a controlled English speech dataset with counterfactual pairs. Each recording states that an event occurred because of a particular cause, and the question asks the model to select that cause from four options. For example, for the question "Why did the library close early?", the paired recordings state "The library closed early because the electricity supply failed." and "The library closed early because a water pipe burst." The two recordings share the same sentence prefix and differ only in the cause segment. The question, choices, temporal placement, and waveform outside the replacement interval remain fixed, while the correct answer changes with the cause.

We synthesize speech with IndexTTS2 (Zhou et al., 2026) using a fixed reference voice and a calm speaking style. The shared prefix and the two cause segments are synthesized separately. The shorter cause segment is padded with trailing silence so that both recordings have the same replacement boundaries. Complete sentences are placed at early, middle, or late positions within 10-s recordings. We check audio duration, evidence boundaries, and waveform identity outside the replacement interval, and conduct spot listening checks.

The discovery and evaluation splits each contain 480 pairs from 40 scenarios, with 12 pairs per scenario and no shared scenarios between splits. Each split contains 160 pairs in each of the three temporal-position groups, with balanced answer labels within each group.

We analyze two evidence scopes: the complete target sentence and the cause segment replaced between A and B. Both use the same audio examples. Qwen3-Omni uses 120 pairs per split, selected to balance scenarios and temporal positions; the other eight models use the full data. State replacement uses the independent 11-layer grid to compare auditory evidence states with all text states. Path experiments use Prefix/Suffix interventions to control attention from all text tokens to evidence, followed by layer-by-layer scans within the initially identified access-transition range. Confidence intervals use 5,000 bootstrap resamples clustered by the 40 scenarios.

H.2 GEMMA 4: FROM SOUND TO SPEECH

Gemma 4 supports multilingual speech recognition and translation (Gemma Team et al., 2026). Our speech extension evaluates its evidence integration on controlled spoken counterfactuals. In this task, Gemma 4 reliably distinguishes the counterfactual answers. Discovery accuracy is 100%, as is evaluation accuracy on both A and B inputs. Blocking text attention to the cause segment across all backbone layers reduces accuracy to 69.17% and the correct-answer margin by 9.99 (95% CI: [9.45, 10.54]). With the complete sentence as evidence, blocking reduces accuracy to 75.21% and the margin by 8.94 (95% CI: [8.36, 9.52]).

State replacement reveals a clear depth pattern: replacing auditory evidence states at early layers shifts predictions toward A, but this effect rapidly diminishes as text-state effects strengthen at intermediate layers. Replacing all text states at layer 19 produces A's answer in 72.71% of examples, rising to 100% at layer 24 (zero-based indexing). The finer Prefix/Suffix scans locate the 20–80% recovery envelope at $[16, 26)$ for the cause segment and $[16, 25)$ for the complete sentence. Figure 24 presents these results alongside the P0 sound-state curves.

These results show that Gemma 4 also exhibits the depth-wise shift from auditory evidence state effects to text-state effects in speech. Further scaling of its encoder-free audio interface with broader audio training offers a promising direction for studying unified audio–language computation.

H.3 SPEECH STATE AND ACCESS RESULTS ACROSS NINE MODELS

Figure 25 shows speech state-replacement results across all nine models. With the cause segment as auditory evidence, every model exhibits stronger auditory evidence state effects early and increasing text-state effects later, with the transition depth varying across models. Gemma 4 thus shares this speech-state pattern with the other models.

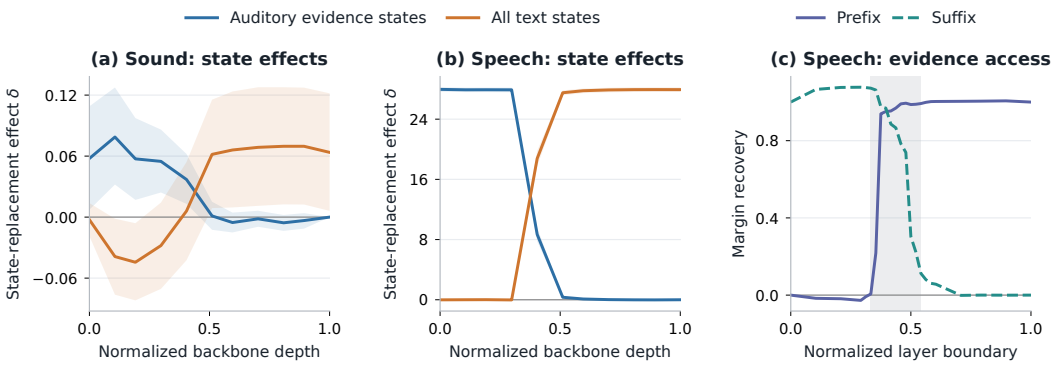


Figure 24: Gemma 4 state-replacement curves on P0 sound inputs and the speech extension, alongside Prefix/Suffix access curves for the speech cause segment. Shading in the state panels shows 95% confidence intervals; the gray region in the access panel marks the 20–80% recovery envelope. Vertical scales differ across panels.

Path experiments also detect evidence-access effects in all nine models: blocking text attention to the cause segment across all backbone layers reduces the correct-answer margin, with all nine 95% confidence intervals above zero. Finer Prefix/Suffix scans show model-dependent access stages. For example, the cause-segment access envelopes are $[10, 20)$ for MiMo-Audio, $[9, 15)$ for Qwen2-Audio, $[10, 33)$ for Qwen3-Omni, and $[16, 26)$ for Gemma 4. Access transitions in speech can begin earlier or span more layers than in sound tasks. Together, the state and path results show that this depth-dependent organization of auditory evidence use also appears in controlled speech understanding.

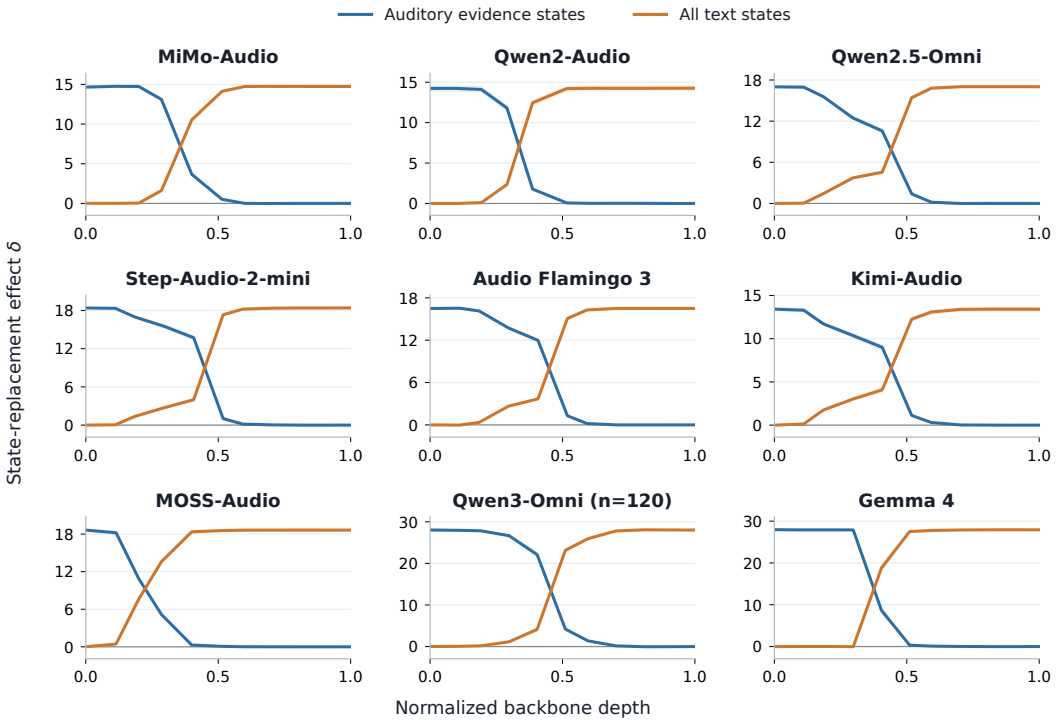


Figure 25: State-replacement curves for speech counterfactual pairs across nine models. Blue denotes auditory evidence states at the cause segment, and orange denotes all text states. Shading shows 95% confidence intervals from scenario-cluster bootstrap resampling. Qwen3-Omni uses 120 pairs; the other models each use 480. Vertical scales vary by model.

This extension uses synthetic English speech with a fixed voice to examine state effects across depth and evidence-access effects. The two evidence scopes are alternative views of the same data.