CAN WE READ THE MIND OF AN AUDIO LLM? A VERBALIZABLE, MULTILINGUAL MIDDLE-LAYER WORKSPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent interpretability work shows that language models maintain a small, privileged set of verbalizable representations in their middle layers—readable as words, reused across internal computation, and broadcast to later processing—that behaves like a global workspace, so far characterized for text inputs. We show that an audio language model forms such representations from sound. Reading an off-the-shelf Qwen3-Omni model with a logit lens at the audio-token positions, which precede and cannot see the question, we find answer-relevant concepts readable as words from the early and middle layers on, before any token is emitted. A waveform-swap control fixes the question and options and swaps only the audio (the real clip, a mismatched clip, or silence): the clip-level readout is best with the real clip, weaker with a mismatched clip, and at chance with silence, so it is driven by the sound rather than by the lens's preference for some option tokens. Traced across depth, this sound-driven signal is null at the input, becomes significant about a tenth of the way up, and persists to the output. Patching the audio-position activations into a run whose audio was removed restores the answer from the early and middle bands but not from the late layers, so the audio content is used, and committed before the late layers. The readable content is conceptual rather than a re-transcription: one concept surfaces in several languages and scripts, and a speaker's emotion is read at the audio positions although the model's own transcript and caption usually do not name it. On a clip about a president forced to resign, *Watergate* and *scandal* are followed by the answer, *Nixon*, at positions that cannot see the printed options. In this model, the verbalizable representations that text models form thus also form from sound.

1 INTRODUCTION

An audio language model has been a black box: we see what it says, never what it thinks. Chain-of-thought helps only if the model writes its reasoning out, and a model that answers a spoken question directly writes nothing down. In this work we read an audio LLM during the forward pass, as it listens and before it answers. Intermediate concepts, such as the source of a sound, a speaker's emotion or an inferred event, take shape from its early and middle layers on and become legible as words.

Prior work shows that intermediate computation is, to a useful degree, legible. By projecting intermediate states into vocabulary space, the logit lens (nostalgebraist, 2020) and its trained successors (Belrose et al., 2023) surface partial predictions well before those tokens are produced. Gurnee et al. (2026) argue that these readable states are not incidental: language models keep a small, privileged set of *verbalizable* representations in their middle layers that is used for internal reasoning, broadcast onward and selectively recruited, a functional global workspace (Baars, 1988; Dehaene & Naccache, 2001). They describe three functional bands: *sensory* early layers, a *workspace* band from about a third of the way up to shortly before the output, and *motor* late layers aligned to the output. This was shown for text inputs, including code; audio was not tested.

Whether this extends to audio matters for oversight. Speech agents consume raw audio, may use paralinguistic content that a transcript discards, and can act directly. If their decision-relevant concepts

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

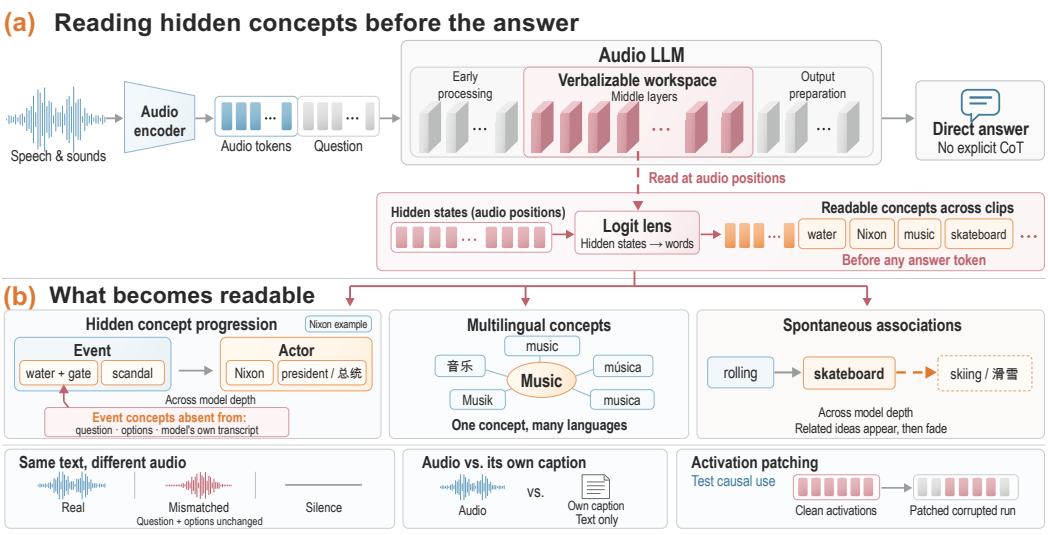


Figure 1: **Reading an audio LLM before it answers.** **(a)** The audio encoder writes features at audio-token positions, which precede the written question; we apply the model's own output head early at those positions and read the word each state would produce, before any answer token exists. **(b)** What the readout contains: event concepts in neither the question, the options, nor the model's own transcript; the same concept in several scripts; and semantic neighbours that appear after the concept that evokes them and are then dropped. Bottom: the controls—swapping only the waveform, replacing the audio with its caption, and patching clean activations into an audio-free run.

were legible before they act, they could be monitored where a chain-of-thought monitor (Baker et al., 2025; Korbak et al., 2025) sees nothing: stated reasoning need not be produced and can be unfaithful (Turpin et al., 2023; Lanham et al., 2023). We therefore ask: when an audio LLM listens to a spoken question, does it form a middle-layer, verbalizable structure from which the sound-derived answer can be read, or does what we read merely reflect which option tokens the lens favours regardless of the sound?

We study Qwen3-Omni-30B-A3B-Instruct (Qwen Team, 2025) as released, whose audio encoder injects features at audio-token positions into a 48-layer text *Thinker*. These positions precede the written question, so a readout taken there cannot see the question or the options. We apply a logit lens at those positions and, as our central control, keep the question and options byte-identical while swapping only the waveform (the real clip, a mismatched clip, or silence). Figure 1 summarizes the setup. We find that these verbalizable middle-layer representations extend to audio. Our contributions are:

- **A controlled readout for audio.** We formulate a clip-level logit-lens readout at the audio positions and a waveform-swap control, new to our knowledge for lens readouts of audio LLMs, that separates what the sound supplies from the lens's preference for some option tokens (Section 3).
- **Readable, sound-driven concepts.** Answer-relevant concepts, including intermediate ones in neither the prompt nor the model's own transcript, are readable before any output token (Section 4.2). The sound-driven part of the readout is null at the input, opens early, and persists to the output (Section 4.3).
- **Causal use.** Patching clean activations into a run whose audio was removed restores the answer from the sensory and workspace bands but not from the motor band, so the audio content is used, not merely decodable (Section 4.4).
- **Conceptual content.** The readout is multilingual, and it carries what the words lack: a speaker's emotion is read at the audio positions although the model's own transcript and caption usually do not name it (Sections 4.5 and 4.6).

2 Related Work

Reading transformer internals. The logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2023) decode intermediate states into vocabulary space; circuit- and feature-level work (Elhage et al., 2021; Templeton et al., 2026; Lindsey et al., 2025) studies how such content is computed. On text, lens readouts can expose latent multi-hop reasoning (Yang et al., 2024; Biran et al., 2024). In vision-language models, object words become readable at image-token positions and flow to the text positions (Jiang et al., 2025; Neo et al., 2025); lenses have also been applied to text-to-image encoders (Toker et al., 2024). We ask the same of audio, where the sound also carries content that no transcript holds. Gurnee et al. (2026) characterize a verbalizable middle-layer global workspace and introduce the Jacobian lens, a principled correction to the logit lens, to map it. Chain-of-thought monitors (Baker et al., 2025; Korbak et al., 2025) instead depend on stated reasoning, which can be unfaithful (Turpin et al., 2023; Lanham et al., 2023).

Latent language and audio lenses. Text models represent meaning in a shared "semantic hub" across languages, often closest to English (Wu et al., 2025; Wendler et al., 2024). Two lines of work apply a vocabulary projection to audio inputs: the Semantic Hub Hypothesis (Wu et al., 2025) shows audio content is lens-readable in the model's dominant language, and AudioLens (Yang et al., 2025) tracks how surface auditory attributes (*e.g.* gender, emotion) evolve across layers. Probing studies show which acoustic and phonetic information speech encoders hold layer by layer (Pasad et al., 2021). We ask not whether *any* audio content is lens-readable, but whether the workspace organization of Gurnee et al. (2026), a verbalizable middle band carrying task-relevant, *inferred* content, appears for audio. We also add a waveform-swap control, without which a lens readout cannot be attributed to the sound rather than to the lens's preference for some option tokens. To our knowledge no prior lens study of audio LLMs uses such a control.

Audio LLMs. Audio LLMs couple an audio or speech encoder to a language-model backbone (Chu et al., 2024; Tang et al., 2024; Gong et al., 2024); Qwen3-Omni injects encoder features at dedicated audio-placeholder positions, a natural target for position-resolved lensing. We read such a model as it answers directly, with no reasoning written out, and ask what is already legible inside it.

3 Method

3.1 Problem Formulation

We study an audio LLM that answers a multiple-choice question about an audio clip. An input consists of a waveform a and a text prompt $x = (q, o_1, \ldots, o_m)$, a question and m answer options (m=4 in every readout). An audio encoder maps a to a sequence of features, which are injected into a decoder-only language model, the *Thinker*, at a contiguous block of audio-placeholder positions $\mathcal{P}_{\mathrm{aud}}$; the prompt occupies the text positions $\mathcal{P}_{\mathrm{txt}}$ that follow. We write $h_p^{(\ell)}(a, x) \in \mathbb{R}^d$ for the residual-stream state at position p after layer ℓ, where $\ell = 0$ is the embedding output and $\ell = 1, \ldots, L$ are the Thinker layers, and $\delta(\ell) = \ell/L$ for its relative depth. Because attention is causal and the audio positions precede the prompt, a state at an audio position cannot depend on the question or the options:

$$h_p^{(\ell)}(a, x) = h_p^{(\ell)}(a, x') \qquad \text{for all } p \in \mathcal{P}_{\mathrm{aud}},\ \ell \in \{0, \ldots, L\},\ \text{and prompts } x, x'. \tag{1}$$

A readout taken at $\mathcal{P}_{\mathrm{aud}}$ is therefore a function of the sound (and a fixed template prefix), not of the printed text.

The global-workspace account of Gurnee et al. (2026) holds that a language model keeps a privileged set of verbalizable representations in a middle band of layers, between *sensory* layers that encode the input and *motor* layers that prepare the output. For an audio LLM this account makes three testable predictions, which organize our method: (i) concepts on which the answer depends are readable, as words, at the audio positions before any output token is produced; (ii) the part of this readout that the sound supplies, as opposed to the option-token prior defined below, is present in the middle band; and (iii) this content is used by the computation that produces the answer, rather than being merely decodable. Sections 3.2 and 3.3 define the readout used for (i), Section 3.4 tests (ii), and Section 3.5

tests (iii). Section 3.6 then asks what the readable content adds beyond the words of the clip. We test verbalizability, depth and use, not broadcast or selective recruitment.

3.2 Logit-Lens Readout at the Audio Positions

The logit lens (nostalgebraist, 2020) reads an intermediate state by applying the model's own final normalization and unembedding $W_U \in \mathbb{R}^{|V|\times d}$ before the last layer. At layer ℓ and position p this gives vocabulary logits and, for any token $t \in V$, a rank:

$$z_p^{(\ell)} = W_U \,\mathrm{norm}\big(h_p^{(\ell)}\big) \in \mathbb{R}^{|V|}, \qquad \mathrm{rank}_p^{(\ell)}(t) = 1 + \big|\{v \in V :\; z_{p,v}^{(\ell)} > z_{p,t}^{(\ell)}\}\big|. \tag{2}$$

Rank 1 means that t is the top readout. We work with ranks rather than probabilities because the scale of the logits drifts with depth, whereas ranks are comparable across layers. We call a pair (ℓ, p) a *cell*, and the tokens of rank at most k at a cell its top-k readout. No lens parameter is trained, so every readout needs only a forward pass of the model.

A concept rarely corresponds to a single token. We identify a concept c with a set of surface tokens $T(c)$—its word pieces and, where stated, its forms in other languages and scripts—and define its rank at a cell, and its best rank over a set of positions $\mathcal{P}$, as

$$\mathrm{rank}_p^{(\ell)}(c) = \min_{t \in T(c)} \mathrm{rank}_p^{(\ell)}(t), \qquad r_c^{(\ell)}(\mathcal{P}) = \min_{p \in \mathcal{P}} \mathrm{rank}_p^{(\ell)}(c). \tag{3}$$

The curve $\ell \mapsto r_c^{(\ell)}(\mathcal{P}_{\mathrm{aud}})$ traces where in depth a concept becomes readable from the sound. We take the minimum over positions rather than the mean because an audio clip spans many positions of which only a minority carry a given concept (Section 4.2); an average would dilute a signal that is present. To measure how widely a concept is represented, we also count the cells of a band at which it is in the top-k readout, and compare this count against that of unrelated placebo words. The logit lens is the identity-Jacobian special case of the Jacobian lens that Gurnee et al. (2026) use to map the workspace, so early-layer readouts are noisy by construction (Appendix A.1).

3.3 Functional Bands and the Clip-Level Readout

We use the three functional bands of Gurnee et al. (2026) and partition the $L + 1$ readout depths with boundaries fixed *a priori* and not tuned on our data: a sensory band $\mathcal{B}_{\mathrm{sen}} = \{\ell : \delta(\ell) < 0.35\}$, a workspace band $\mathcal{B}_{\mathrm{ws}} = \{\ell : 0.35 \le \delta(\ell) \le 0.80\}$, and a motor band $\mathcal{B}_{\mathrm{mot}} = \{\ell : \delta(\ell) > 0.80\}$. Their workspace likewise begins about a third of the way up but ends later, shortly before the output.

To turn the readout into a prediction we score each answer option by its first token. Let $t_1, \ldots, t_4$ be the first tokens of the four options. The readout of a band $\mathcal{B}$ scores option i by its best rank over the band and the audio positions, and predicts the best-ranked option:

$$r_i(\mathcal{B}) = \min_{\ell \in \mathcal{B}} \min_{p \in \mathcal{P}_{\mathrm{aud}}} \mathrm{rank}_p^{(\ell)}(t_i), \qquad \hat{\imath}(\mathcal{B}) = \operatorname*{arg\,min}_{i \in \{1,\ldots,4\}} r_i(\mathcal{B}). \tag{4}$$

The procedure is given as Algorithm 1 in Appendix B. It makes one prediction per clip, so no audio position is treated as an independent trial. By Eq. (1) the prediction cannot read the printed options, but it can inherit an *option-token prior*: some option tokens are readable at the audio positions whatever the audio contains. Answer positions are also imbalanced, so that always choosing the most frequent position scores about 46% on our data rather than 25%. We therefore score the readout by balanced accuracy over the four answer positions,

$$\mathrm{BA}(\mathcal{B}) = \frac{1}{4}\sum_{j=1}^{4} \frac{1}{|\mathcal{D}_j|} \sum_{n \in \mathcal{D}_j} \mathbf{1}\big[\hat{\imath}_n(\mathcal{B}) = j\big], \tag{5}$$

where $\mathcal{D}_j$ is the set of clips whose correct option is in position j. Any constant guess scores exactly 25% under Eq. (5). As a threshold-free complement we also report the median, over clips, of the best rank $r_{i^\star}(\mathcal{B}_{\mathrm{ws}})$ of the correct option $i^\star$ in the full vocabulary.

3.4 Separating Sound from the Option-Token Prior: the Waveform-Swap Control

To measure how much of the readout the sound itself supplies, we run every clip under three audio conditions while the prompt x is held byte-identical: the real clip a, a mismatched clip $\tilde{a}$ drawn from

the same category, and silence $a_\varnothing$ (Appendix C.2). Writing $A_\kappa(\mathcal{B})$ for the accuracy of the prediction in Eq. (4) under condition κ, measured by Eq. (5) or, where stated, as raw accuracy, we define two audio-driven gaps for a band $\mathcal{B}$:

$$\Delta_{\mathrm{sil}}(\mathcal{B}) = A_{\mathrm{real}}(\mathcal{B}) - A_{\mathrm{sil}}(\mathcal{B}), \qquad \Delta_{\mathrm{mis}}(\mathcal{B}) = A_{\mathrm{real}}(\mathcal{B}) - A_{\mathrm{mis}}(\mathcal{B}). \tag{6}$$

Under silence the audio positions carry nothing of the clip, so A_{sil} is what the option-token prior achieves on its own, and Δ_{sil} is the part of the readout that this prior cannot supply. Silence, however, yields fewer audio positions than a typical real clip, and Eq. (4) takes a minimum over positions, so a longer audio span offers more chances to rank an option well. A mismatched clip has nearly as many audio positions as a typical real clip, so Δ_{mis} largely removes this confound and asks whether the *right* sound beats a *wrong* one.

We evaluate Eq. (6) for each band and, setting $\mathcal{B} = \{\ell\}$, for each of the $L + 1$ depths separately, which traces where in depth the sound-driven signal appears. If the sound's content is verbalized in the workspace, Δ_{sil} should be small in the sensory band, where acoustic features are present but not yet verbalized, and clearly positive in the workspace band. All comparisons are paired at the clip level. We test them with the exact McNemar test on paired correct/incorrect outcomes: two-sided for the condition and band contrasts, and one-sided (real $>$ silence, the direction the account predicts) for the per-depth sweep, with Benjamini–Hochberg control across the $L + 1$ depths.

3.5 From Decodable to Used: Activation Patching

A readable concept need not be used. We test use on the clips that the model answers correctly with the real audio and incorrectly once the audio is removed. For such a clip we form a clean run on (a, x) and a corrupt run on (a_0, x), in which the audio mel-features are zeroed while the token sequence is kept identical, so that the two runs align position by position. A patched run repeats the corrupt pass but overwrites the residual stream at the audio positions of one band with the clean states,

$$h_p^{(\ell)} \leftarrow h_p^{(\ell)}(a, x) \qquad \text{for all } \ell \in \mathcal{B},\ p \in \mathcal{P}_{\mathrm{aud}}, \tag{7}$$

and lets the forward pass continue. The patch *restores* the answer if the patched run generates the correct option. As a graded measure we report the fraction of the clean–corrupt gap in the correct answer's output logit g that the patch recovers,

$$R(\mathcal{B}) = \frac{g_{\mathrm{patch}}(\mathcal{B}) - g_{\mathrm{corr}}}{g_{\mathrm{clean}} - g_{\mathrm{corr}}}. \tag{8}$$

A patch at one layer propagates to every later layer, so Eq. (7) localizes coarsely: restoration from band $\mathcal{B}$ shows that the audio content is used, and committed, no later than $\mathcal{B}$, not that $\mathcal{B}$ alone is responsible. As a control we apply Eq. (7) to the workspace band at the text positions instead of the audio positions.

3.6 What the Sound Adds: the Two-Mind Comparison

The swap control asks whether the readout depends on the sound; the two-mind comparison asks what the sound contributes beyond the words it contains. For each clip we take the model's own emotion-free caption $\gamma(a)$, a written description of the clip that is meant to omit paralinguistic detail such as the speaker's emotion. We then run the model twice with the identical prompt x: an *audio mind* that receives the waveform, and a *caption mind* that receives $\gamma(a)$ as text and no waveform. For a concept c we compare its best rank across depth in the two runs, writing the input of each run after the semicolon:

$$r_{c,\mathrm{aud}}^{(\ell)} = r_c^{(\ell)}\big(\mathcal{P}_{\mathrm{aud}};\, a, x\big), \qquad r_{c,\mathrm{cap}}^{(\ell)} = r_c^{(\ell)}\big(\mathcal{P}_{\mathrm{cap}};\, \gamma(a), x\big), \tag{9}$$

where $\mathcal{P}_{\mathrm{cap}}$ are the positions of the caption tokens. Neither readout includes the positions of the printed options, so an option's text cannot be read back as the concept. A concept that forms in the audio mind but not in the caption mind is content that the sound carries and a transcript-like description of it does not. We also compare the two minds behaviourally, by the accuracy of the answer that each run produces. On emotion clips we add two reference runs with the same prompt: one given the model's own verbatim transcript of the clip as text, and one given silence. If the sound carries affect that words do not, accuracy should drop from audio to transcript (words only), drop again for the caption (meant to omit emotion), and reach chance under silence.

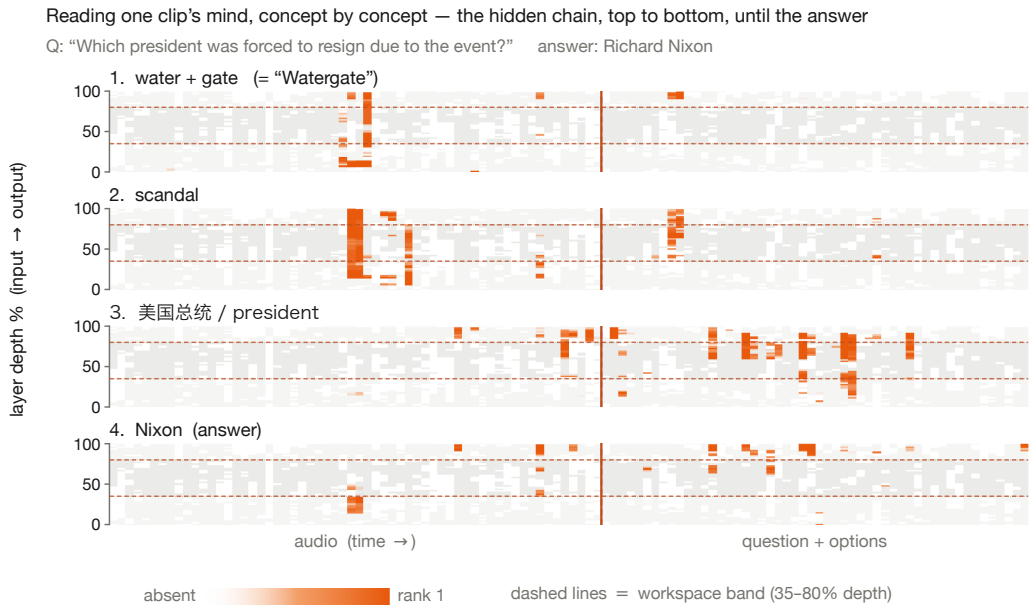


Figure 2: **Reading one clip concept by concept** (Nixon clip). Each row is one concept's logit-lens rank at every position (x: audio positions, then question+options) and depth (y); darker = better rank, and the dashed lines bound the workspace band. The event concepts (water/gate, scandal) reach rank 1 at the audio positions in the early layers; Nixon first reaches rank 1 there at 27–38% depth, and the role Měiguó zǒngtǒng ("US president") only from about 62% depth.

4 Experimental Results

4.1 Experimental Setup

Model. We read Qwen3-Omni-30B-A3B-Instruct (Qwen Team, 2025) without fine-tuning. Its Thinker is a mixture-of-experts decoder with L=48 layers, width d=2048, a vocabulary of 152,064 tokens and an untied unembedding, so the bands of Section 3.3 are layers 0–16 (sensory), 17–38 (workspace) and 39–48 (motor). Each readout is one bfloat16 forward pass on two H200 GPUs (Appendix C).

Data. Clips come from the MMAU benchmark (Sakshi et al., 2025), whose multiple-choice questions cover sound, speech and music; one boundary check uses a spoken version of TriviaQA (Joshi et al., 2017) (Appendix D.8). Appendix C.3 lists the items each experiment uses, and Appendix C.7 gives the statistical tests.

4.2 Answer-Relevant Concepts Are Readable Before the Answer

An event-to-actor chain. Our clearest case asks which president was forced to resign (answer: Nixon). The model's own transcription is empty garbling ("The ote — [End of Audio]") and its caption recovers only a generic political scene ("…president of…politician…"), never *Watergate*, *scandal* or *Nixon*. Yet the readout at the audio positions assembles the chain (Figure 2). In the early layers water and gate reach rank 1 at adjacent audio positions, reconstructing *Watergate*, which is not a single token, and scandal reaches rank 1 as well. The actor comes later: Nixon reaches rank 1 across the transition into the workspace band, and the role Měiguó zǒngtǒng ("US president") from the upper workspace band on, so the chain runs from event to actor. *Nixon* is a printed option, but by Eq. (1) its readout at the audio positions can come neither from the printed options nor from the question: the chain is what the model draws from the sound. (The question refers to an event "mentioned by the speaker", so the event words may be spoken; Appendix A.1.)



Figure 3: **One audio position, read down the layers** (skateboard clip). Each row lists the top readout tokens at one layer (pure lens noise filtered, noise-only layers omitted). Blue marks the answer category, orange the neighbour *skiing* (huáxuě) and downhill.

The pattern recurs, specifically and sparsely. A treaty clip surfaces negotiation before *Egypt and Israel*. An archaeology clip surfaces a dig scene, although *Howard Carter* itself is read only at the text positions, and sound clips name their source, often in another language (Appendix D.5). These cases are illustrative; the population-level test is the swap control (Section 4.3). Ten unrelated placebo words (*banana*, . . .) are essentially absent from the same cells, except *trumpet* on a whip crack (Appendix D.3). Only a minority of audio positions carry a concept, which is why Eq. (4) takes a minimum over positions.

One audio position, read down the layers. Figure 3 follows one audio position of a skateboard clip. The rolling sound is verbalized first (gǔndòng, “rolling”); the category then forms as skate and skateboard and leads the readout. Beneath it sits a semantic neighbour, huáxuě (“skiing”), spoken nowhere in the clip and in neither the question nor the options; it stays in the top readout for several layers without reaching rank 1 and then drops out. In every association family we could check, the neighbour appears after the concept that evokes it, and the most distant associate, lǎohǔ (“tiger”) on the lion clip, appears latest, close to the output. The richest cases are whole neighbourhoods: the archaeology clip reconstructs a dig scene (burial, excavation, treasures, guān “coffin”, yízhǐ “ruins”; Appendix D.6).

4.3 The Readout Is Driven by the Sound from Early Layers On

The sound drives the readout. With the question and options held byte-identical, the real audio lifts the workspace readout clearly above chance; a mismatched clip lowers it, and silence drops it to chance (Table 3 and Figure 9). Both gaps of Eq. (6) are significant, and Δ_{mis} largely controls for the length confound of silence: the right sound beats a wrong sound of the same category and nearly the same length. The threshold-free median rank of the correct option follows the same order. The effect is carried mainly by sound-source clips (Appendix A.1).

The signal opens early and stays. Depth by depth (Figure 8), real audio and silence are indistinguishable through the first several layers, although the encoder has already written acoustic features into the stream. The gap opens about a tenth of the way up and stays open to the output, significant at most workspace depths and every motor depth. Band by band (Table 4), silence reads out as well as real audio in the sensory band but falls to chance in the workspace band, where the option-token prior alone fails; the advantage over a mismatched clip is largest in the motor band.

4.4 The Audio Content Is Causally Used Before the Late Layers

On the clips whose answer depends on the audio, patching clean states at the audio positions of the sensory band restores the correct answer on every clip, and patching the workspace band restores it on all but one, a counting question (Table 2); both recover most of the logit gap of Eq. (8). Patching the motor band alone restores no clip. The audio content is therefore used, and it is committed before the late layers. Because an early patch propagates downstream, this places its use before the motor band rather than in one band alone; a workspace-band patch at the text positions, which have

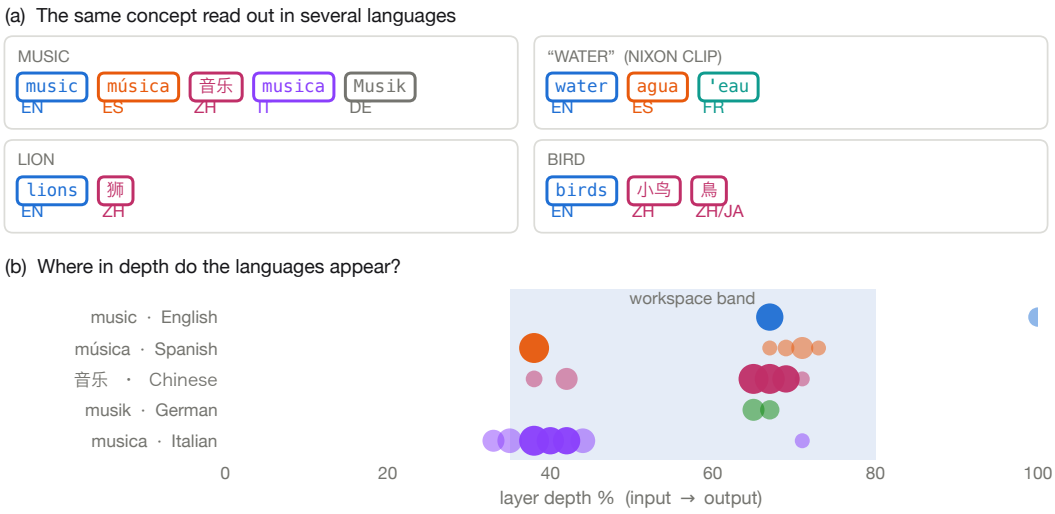


Figure 4: **The workspace is multilingual. (a)** Tokens read at the audio positions for four concepts, each coloured by its language. **(b)** On one music clip, every depth at which each form of *music* is read; dot size and opacity track its rank, and the shaded span is the workspace band. All five forms surface inside that band.

already attended to the audio, restores the answer as well (Appendix A.1). Single-layer deletion is in Appendix C.6.

4.5 The Readout Is Multilingual

The clips and prompts are in English, yet one concept is often read in several languages (Figure 4). On a music clip, *music* appears in five forms inside the workspace band (English music, Chinese yīnyuè, Spanish música, German Musik, Italian musica); on the Nixon clip *water* is read in English, Spanish and French, a lion clip reads lions and Chinese shī ("lion"), and a bird clip birds and Chinese xiǎoniǎo ("little bird"). Across the corpus the top workspace readout is mostly English, with Chinese a strong second (Table 5). Since the logit lens may favour frequent CJK tokens, we check that the Chinese forms of *music* and *bird* are the top readout far more often on the clips whose content they name than elsewhere, which a globally frequent token would not show (Appendix D.4). The readout thus reflects a concept, not the English word for it.

4.6 The Audio Carries What Its Caption Lacks

Emotion is read from the voice. Emotion is the cleanest test of whether the readout holds more than the words: an emotion-flip question (did a speaker's emotion change during the dialogue?) and its options name no emotion, and the model's own caption is meant to leave emotion out. Yet on the clips of Figure 5 the emotion word becomes the top readout at the audio positions, which precede the question, and apart from dips keeps that rank from the workspace band, or earlier, into the last layers. Across the emotion clips, the emotion word the audio mind reads rarely appears in the model's own verbatim transcript and is in its caption only on a minority of clips; most strongly read emotions are absent from the transcript (Table 1). On a sarcastic clip, for instance, the readout reaches frustration, a word absent from the question, the transcript and the caption. On a separate emotion-recognition slice, behaviour agrees: the model answers best from the audio, less well from its transcript, worse from its caption and at chance from silence, though at this size only the gap to silence is significant (Table 1). Unlike attribute tracking (Yang et al., 2025), this tests the emotion read from the voice against the model's own words for the clip.

Two minds on the same clip. The same asymmetry holds beyond emotion. A roar is read as lion across the workspace band in the audio mind, while in the caption mind lion appears almost only at

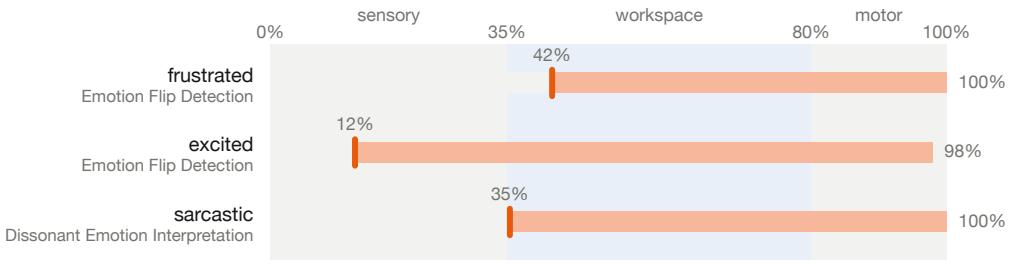


Figure 5: **Affect is held in the workspace.** For three emotion clips, the span of depths over which the emotion word is the top readout at the audio positions, from its first to its last rank-1 layer (workspace band shaded); within a span the word can briefly fall back, most visibly on the sarcastic clip. The two emotion-flip questions name no emotion; the third asks why a statement is sarcastic, but the audio positions precede the question and cannot attend to it.



Figure 6: **Two minds, one clip.** The roaring clip read two ways, with *lion* pinned (darker = better rank). **Top** (audio): lion lights up across the workspace over the audio span; the model answers *Lion*. **Bottom** (the model's own emotion-free caption): lion appears almost only at the printed option, and the model answers *Wolf*.

the printed option and the model answers *Wolf* (Figure 6): its caption describes growls and a howl but names no lion. On a speaker-role clip the audio mind reads priest where the caption has only “father”. The audio mind also forms the correct speaker role more often, and answering from the audio beats the caption on every MMAU-mini track (Table 1). The disagreement can also flag an error (Appendix D.7), and where the sound adds nothing, listening does not help (Appendix D.8).

5 Conclusion

We asked whether the verbalizable middle-layer representations found in text language models also form in an audio LLM. Reading Qwen3-Omni with a logit lens at the audio positions, we find that they do: answer-relevant concepts become readable before any token is emitted, the sound drives this readout from the early layers to the output, and the audio content is causally used before the late layers. The content is conceptual rather than a re-transcription: it is multilingual, carries a speaker's emotion that the model's own transcript and caption usually do not name, and reaches semantic neighbours of what is heard. This is exactly what a chain-of-thought monitor cannot see. Appendix A.1 sets out the limits of this evidence: a proxy lens, one model, a sound-leaning sample, and coarse causal localization.

AI USE STATEMENT

The object of study is a publicly available model, Qwen3-Omni-30B-A3B-Instruct (Qwen Team, 2025), read without any fine-tuning (Appendix C). Beyond the model under study, generative AI tools were used only to polish the writing of the manuscript (grammar and phrasing). They did not contribute to the research questions, the methodology, the experiments, or the findings; every readout shown is the model's actual activation read through Eq. (2), and every number comes from the authors' own experiment outputs. All AI-polished text was reviewed by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics. This work analyzes a publicly available model on public benchmarks, MMAU (Sakshi et al., 2025) and a spoken version of TriviaQA (Joshi et al., 2017); it involves no human subjects and releases no new datasets or models.

REPRODUCIBILITY STATEMENT

The lens in Eq. (2) has no trained parameters, so every readout needs only a forward pass of a public model. The method is specified in Section 3, and Algorithm 1 (Appendix B) gives the clip-level readout. Experimental details are in Appendix C: model and lens (Appendix C.1), audio conditioning and the swap conditions (Appendix C.2), item selection (Appendix C.3), the two-mind comparison (Appendix C.4), activation patching (Appendix C.5), layer deletion (Appendix C.6), statistics (Appendix C.7), and compute (Appendix C.8). Full control results, per-band statistics, the placebo control, and per-clip case details are in Appendix D.

REFERENCES

Bernard J Baars. A cognitive theory of consciousness. 1988.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *CoRR*, abs/2503.11926, 2025. doi: 10.48550/ARXIV.2503.11926. URL https://doi.org/10.48550/arXiv.2503.11926.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112, 2023. doi: 10.48550/ARXIV.2303.08112. URL https://doi.org/10.48550/arXiv.2303.08112.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 14113–14130. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.781. URL https://doi.org/10.18653/v1/2024.emnlp-main.781.

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. *CoRR*, abs/2407.10759, 2024. doi: 10.48550/ARXIV.2407.10759. URL https://doi.org/10.48550/arXiv.2407.10759.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and

Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=nBZBPXdJlC.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul C. Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *CoRR*, abs/2607.15495, 2026. doi: 10.48550/ARXIV.2607.15495. URL https://doi.org/10.48550/arXiv.2607.15495.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=94kQgWXojH.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca D. Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Madry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vladimir Mikulik. Chain of thought monitorability: A new and fragile opportunity for AI safety. *CoRR*, abs/2507.11473, 2025. doi: 10.48550/ARXIV.2507.11473. URL https://doi.org/10.48550/arXiv.2507.11473.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023. doi: 10.48550/ARXIV.2307.13702. URL https://doi.org/10.48550/arXiv.2307.13702.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Blog post.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pp. 914–921. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688093. URL https://doi.org/10.1109/ASRU51503.2021.9688093.

Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025. doi: 10.48550/ARXIV.2509.17765. URL https://doi.org/10.48550/arXiv.2509.17765.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam S. Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *CoRR*, abs/2605.29358, 2026. doi: 10.48550/ARXIV.2605.29358. URL https://doi.org/10.48550/arXiv.2605.29358.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 9713–9728. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.524. URL https://doi.org/10.18653/v1/2024.acl-long.524.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15366–15394. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.820. URL https://doi.org/10.18653/v1/2024.acl-long.820.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FrFQpAgnGE.

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 2025, Honolulu, HI, USA, December 6-10, 2025*, pp. 1–8. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434723. URL https://doi.org/10.1109/ASRU65441.2025.11434723.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 10210–10229. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.550. URL https://doi.org/10.18653/v1/2024.acl-long.550.

A DISCUSSION

The main text reads an audio LLM at the positions where the sound enters and finds what the text-model account predicts for verbalizability, depth and use: answer-relevant concepts that are verbalizable from the early and middle layers on, driven by the waveform rather than by the printed options, and causally used before the output. This section states how far that evidence reaches (Appendix A.1).

A.1 LIMITATIONS

Proxy lens. We read with the logit lens (nostalgebraist, 2020), which corresponds to an identity Jacobian in the formulation of Gurnee et al. (2026), not with their Jacobian lens, which averages the linearized layer-to-output effect over many contexts. Early layers are therefore noisy by construction and the band boundaries are approximate. A Jacobian lens would sharpen these measurements.

One model. We read a single model, Qwen3-Omni. Whether audio LLMs with other encoders and other ways of injecting audio show the same organization is untested.

Sample size and composition. The clip-level core has $n{=}140$ items, because the readout needs single-token, non-counting MMAU items, which are rare for speech-content questions; the set therefore leans toward sound-source identification. The position-matched real-versus-mismatch effect is carried mainly by the sound-source clips, while on speech-content clips a mismatched clip reads out about as well (Appendix D.1), so we scope the "read from the sound" claim to sound-source perception. Option positions are imbalanced (A/B/C/D $= 64/36/15/25$), so the naïve 25% chance line is misleading: a majority-class guess already matches raw real-audio accuracy (Table 3). We therefore rest the evidence on balanced accuracy and paired McNemar tests, which both cancel the positional prior. Even then the effect is well below ceiling: among correctly read clips the answer rises clearly *within* the workspace band only in a minority and resolves closer to the motor layers for the rest, and the signal is carried by a few audio positions rather than spread across the clip.

Band specificity. The sound-driven signal is not confined to the middle band. In the per-depth sweep the real$-$silence gap turns significant already within the sensory band and stays open to the output (Figure 8), although at band level the sensory gap is null; the position-matched real$-$mismatch gap grows with depth and is largest in the motor band (Table 4). What is specific to the middle band is that the silence readout falls to chance there. Band-level and per-depth results can also differ: the band readout takes the best rank over all of a band's layers, so a few layers where the option-token prior is strong can lift the silence readout for the whole band, while the per-depth test scores each layer on its own. The bands are fixed *a priori* and differ in size, and the best-rank readout favours larger bands; a size-controlled analysis and a Jacobian lens would test whether the middle band is distinct in this model.

Coarse causal localization. Activation patching shows that the audio-derived content is causally used and committed before the motor band, but it localizes the effect only coarsely: an early-band patch propagates downstream, and patching workspace activations at the text-option positions, which have already attended to the audio, also restores the answer. It also runs only on the clips where audio is decisive, and all ten of those are sound-track clips; the multi-hop clips we checked, such as the Nixon clip, are answered correctly from the text prior even without audio, so they cannot enter this test. The causal claim is therefore about audio content being used in general, not about the multi-hop chains specifically.

Inference versus transcription. The controls show the readout is audio-driven and absent from the prompt, but they do not by themselves separate a concept the model infers from one it hears spoken. This is why the Nixon clip carries the multi-hop claim: both of the model's own audio-to-text routes fail to name the chain—its verbatim transcription is empty garbling and its free-form caption recovers only a generic political scene—yet the readout assembles it. That rules out the lens merely echoing the model's own transcription, but it does not prove the words were never spoken: degraded audio could defeat both routes while still permitting partial, sub-transcription recognition (the adjacent-position water/gate readout is consistent with either). For the association

families we inspected the available transcripts and dropped any neighbour that is literally voiced; that inspection is not paraphrase-proof, so association is stated as a qualitative reading. A forced-alignment transcript would close this gap.

Scope of the language. We use "workspace", "access", and "mind" in the functional sense of Baars (1988) and Dehaene & Naccache (2001)—information made available for report and downstream use. We make no claim about consciousness or subjective experience.

B READOUT ALGORITHM

Algorithm 1 Clip-level four-way workspace readout

Require: clip audio a; text prompt x (question and options $o_{1:4}$ with first tokens $t_{1:4}$); workspace band $\mathcal{B} = \{17, \ldots, 38\}$; audio-pad positions $\mathcal{P}_{\text{aud}}$

Run one forward pass of the Thinker on (a, x); cache residual states $h_p^{(\ell)}$ for $\ell \in \mathcal{B}$, $p \in \mathcal{P}_{\text{aud}}$
for $\ell \in \mathcal{B}$, $p \in \mathcal{P}_{\text{aud}}$ **do**
    $\text{logits} \leftarrow W_U \,\text{norm}(h_p^{(\ell)})$ ▷ logit lens, Eq. (2)
    for $i \in \{1, 2, 3, 4\}$ **do**
        $\text{rank}(t_i \mid \ell, p) \leftarrow 1 + |\{v : \text{logits}_v > \text{logits}_{t_i}\}|$
    end for
end for
for $i \in \{1, 2, 3, 4\}$ **do**
    $r_i \leftarrow \min_{\ell \in \mathcal{B}} \min_{p \in \mathcal{P}_{\text{aud}}} \text{rank}(t_i \mid \ell, p)$ ▷ best rank over the grid
end for
return $\hat{\imath} \leftarrow \arg\min_i r_i$

The minimum over positions (rather than the mean) is used because the signal is sparse across the audio span (Section 4.2). One prediction is made per clip, so no audio position is treated as an independent trial. The per-band analyses restrict $\mathcal{B}$ to the layers of one band.

C EXPERIMENTAL DETAILS

C.1 MODEL AND LENS

We use Qwen3-Omni-30B-A3B-Instruct (Qwen Team, 2025) as released, without fine-tuning, and read its Thinker text stack (L=48 layers, d=2048, $|V|$=152,064, untied language-modeling head). The logit lens applies the model's own final RMSNorm and unembedding to the hidden state at each layer (Eq. (2)); the rank of a target token is one plus the number of vocabulary entries with strictly greater lens logit. No lens is trained.

C.2 AUDIO CONDITIONING AND THE SWAP CONDITIONS

Audio is encoded and injected at <|audio_pad|> positions between <|audio_start|> and <|audio_end|>. The audio encoder emits float32 features while the Thinker runs in bfloat16, so injected features are cast to the model dtype before the forward pass. For the mismatch condition we substitute a different clip of the same MMAU category; for silence we pass a fixed 2 s zero waveform through the same encoder path, which yields fewer audio positions than a typical real clip. The audio span is therefore not identical across conditions; what is identical is the written question and every option token. Because the best-rank readout takes a minimum over positions, a longer span offers more draws, and across clips more audio positions do correlate with better best-ranks. We therefore do not rest the claim on real-versus-silence alone: the decisive contrast is real-versus-mismatch, which compares the right clip with a wrong clip of the same category and similar length and still favors real audio on raw as well as balanced accuracy (Table 3).

C.3 ITEM SELECTION AND DATA

The readout controls use the MMAU (Sakshi et al., 2025) sound and speech items whose correct answer is a single vocabulary token that no other option shares as its first token; counting and numeric questions are left out, since their answer probability can be driven by output formatting rather than the audio. Activation patching uses the sound-track clips answered correctly only with the audio (Appendix C.5). Corpus-level statistics use larger MMAU pools: a corpus of 948 clips whose top-ranked tokens are stored at every cell, the 1000-clip MMAU-mini, 55 emotion clips, a 44-clip emotion slice and the 27 speaker-role clips on which the two minds disagree, and the fact-recall boundary uses a 500-question spoken TriviaQA (Joshi et al., 2017).

C.4 TWO-MIND COMPARISON

Each clip is run twice with the identical question and options: once with the real waveform, and once with the model's own emotion-free caption of that clip supplied as text, with no waveform. The caption is meant to discard paralinguistic detail, although it still names the emotion on a minority of emotion clips (Table 1). Readouts in the caption run are taken over the caption tokens; speaker-role statistics restrict the readout to the input positions so that option text cannot leak in. On the speaker-role clips where the two minds disagree, the audio mind forms the correct role on more clips than the caption mind (Table 1). On most emotion clips the audio mind carries a strong affect readout, and the emotion word it reads rarely appears in the model's own verbatim transcript and appears in its emotion-free caption only on a minority of clips (Figure 5 and Table 1). Behaviourally the audio route leads on the emotion slice and on every MMAU-mini track (Table 1); on the emotion slice the ordering runs audio, then the model's verbatim transcript, then its emotion-free caption, then silence, which is the order the two-mind reading predicts, although at this size only the audio–silence difference is significant.

Table 1: **Two-mind behavioural comparison.** Accuracy answering from the real waveform against answering from the model's own caption of the same clip (meant to omit emotion), with the question and options identical in both runs. On the 44-clip emotion slice the audio–transcript and audio–caption differences are not significant at this size (paired McNemar p=0.125 and p=0.065; audio–silence p=0.0009), so the ordering is descriptive. Of the 55 emotion clips, the emotion word read at the audio positions appears in the verbatim transcript on 6 and in the caption on 13; 39 clips carry a strong affect readout, and on 33 of them the affect is absent from the transcript. On the 27 speaker-role clips where the two minds disagree, the input-position readout forms the correct role on 88.9% of clips with the audio and on 70.4% with the caption.

	n	Audio	Caption	Other
Emotion slice	44	59.1%	43.2%	transcript 47.7%, silence 22.7%
MMAU-mini, sound	333	80.8%	69.7%	
MMAU-mini, speech	333	74.5%	68.5%	
MMAU-mini, music	334	75.4%	69.8%	

C.5 ACTIVATION PATCHING

For each clip we form a clean run (real audio) and a corrupt run in which the audio mel-features are zeroed while the token sequence is held byte-identical, so the two runs align position-for-position. We re-run the corrupt pass while overwriting, at one band of layers and at the audio positions, the residual stream with the clean activations, and read the model's actual next-token answer. We study the clips where audio is necessary—correct with audio, wrong once it is zeroed. There are 10 such clips, all from the sound track: eight sound-source identifications, one sound-duration and one counting question. The single workspace-band miss is the counting clip. The motor band is the top 20% of layers. A control patches workspace-band activations at the text positions instead of the audio positions.

Table 2: **Activation patching.** Clips whose answer flips back to correct when the clean run's residual stream is copied into the corrupted run at one band, and the fraction of the clean−corrupt logit gap recovered. The single workspace-band miss is the counting clip. The last row is the text-position control.

Patched band	Positions	Clips restored	Logit gap recovered
Sensory	audio	$10/10$	95.7%
Workspace	audio	$9/10$	89.1%
Motor	audio	$0/10$	4.8%
Workspace	text	$9/10$	93.1%

C.6 LAYER DELETION

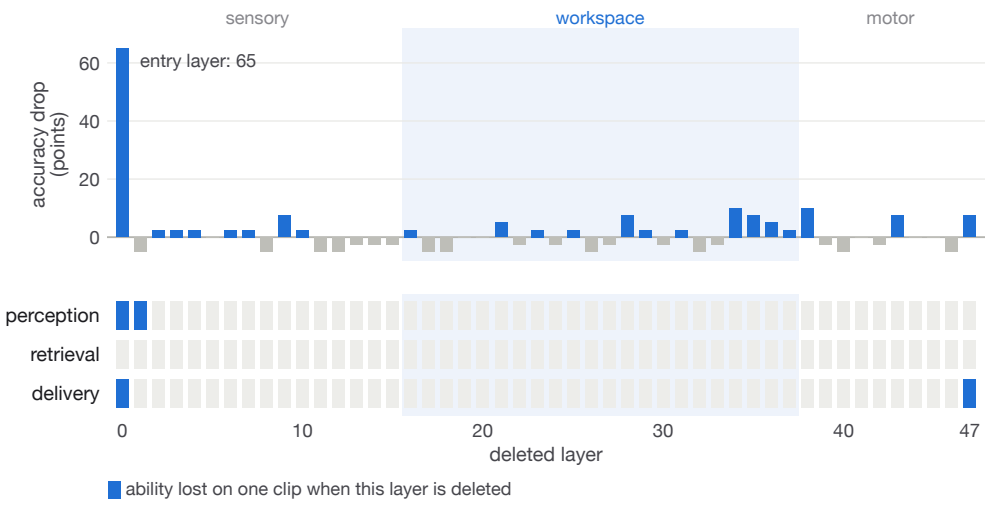


Figure 7: **Deleting one layer at a time. Top:** accuracy drop from bypassing each layer, over 40 clips; only deleting the entry layer causes a large drop (65 points), and every interior layer costs at most ten (grey bars are small gains). **Bottom:** which ability is lost on one clip (blue cells), on the same layer axis. Perception fails only at the entry layers and retrieval at no single layer; delivery is flagged at the entry, where the output collapses, and at the last layer, where the correct option is still emitted but in a changed format.

Reading is correlational, so we also delete layers. We bypass one Thinker layer at a time with an identity map (deletion indices count layers 0–47; lens depths count residual states 0–48, with 0 the embedding output) and measure behavioral accuracy over 40 MMAU clips. Only the entry layer, where the encoder features enter the text stack, is necessary: deleting it collapses the output, while every interior layer costs far less (Figure 7, top). No single interior layer is necessary for the answer, yet patching a whole band restores it (Section 4.4), so the interior carries the decision jointly rather than through one critical layer. On one clip we also track which ability a deletion breaks (Figure 7, bottom): perception (does the event still form at the audio positions?), retrieval (does the recalled answer concept still surface?) and delivery (is the answer still emitted in the expected format?). Only deleting the entry layers stops the event from forming, no single layer stops the answer concept from surfacing, and delivery changes only at the two ends: at the entry the output collapses, while deleting the last layer leaves the correct option letter in place and changes only its format.

Expert routing in this bf16 mixture-of-experts makes the generated answer mildly nondeterministic even without intervention: on one clip, repeating the un-ablated forward pass three times already gives more than one answer, and repeating a full 48-layer graded-rank sweep three times gives a different "most-damaging" layer each time, with only the entry layer recurring. The answer-flip version is stable: across three repeats only the entry layer and the final layer change the output in every repeat, and the final layer's change is a formatting shift (the correct option letter is still emitted, only prefixed). We therefore report only this reproducible conclusion and no per-layer critical layers.

C.7 STATISTICS

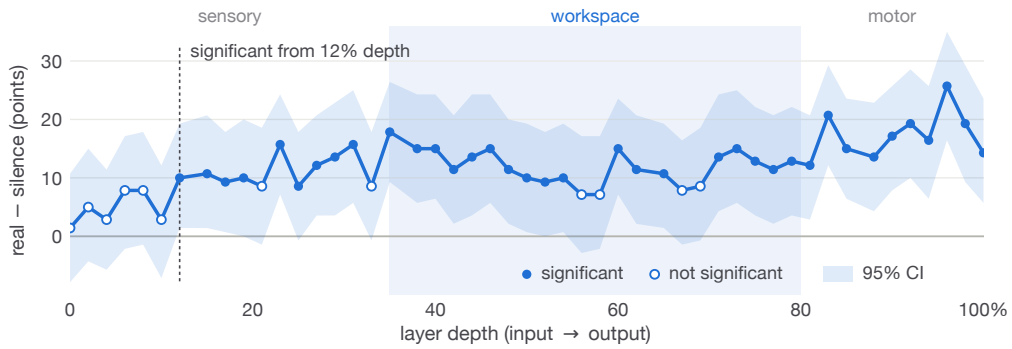


Figure 8: **The audio-driven gap over the option-token prior, layer by layer.** Raw four-way readout accuracy (clip-level) with real audio minus the same quantity with silence, with a 95% bootstrap band; filled markers are depths where the paired test is significant. The gap is null at the input, becomes significant at $\sim$12% depth, and stays open at 18 of the 22 workspace layers and all 10 motor layers, against 9 of the 17 sensory layers; the gap does not fade toward the output. One-sided: 37 of 49 depths reach $p<0.05$, 35 after Benjamini–Hochberg control. Two-sided: 32 of 49 (28 after correction; 16 of 22 workspace and 6 of 17 sensory).

All accuracies are clip-level. Condition contrasts use the exact McNemar test on the paired correct/incorrect outcomes. The aggregate real-versus-silence and real-versus-mismatch contrasts and the per-band tables use the two-sided test. The per-layer depth analysis (Figure 8) tests the directional alternative real$>$silence with the one-sided test, with Benjamini–Hochberg control across the 49 depths. The counts of significant depths under the one-sided and the two-sided test are given in the caption of Figure 8; the qualitative conclusion holds under either convention. The confidence band in Figure 8 is a 95% bootstrap interval.

C.8 COMPUTE

Each condition is a single forward pass of the 30B model on two H200 GPUs. No training or backward pass is involved in any readout.

D ADDITIONAL RESULTS

D.1 AUDIO-SWAP CONTROL: FULL READOUT ACCURACY

Table 3 and Figure 9 give the full control behind Section 4.3. The single-token restriction only gives the lens a clean target; scoring the first token of each option instead lets the same control run over every four-option MMAU-mini clip, where the readout is again ordered real $>$ mismatch $>$ silence in balanced accuracy and silence sits at the chance line (we report only this ordering for that set). Two bounds apply: silence is also a shorter, blank input, so the cleaner length-matched contrast is real-versus-mismatch; and that length-matched effect is carried by the sound-source clips, while on speech-content clips a mismatched clip reads out about as well. The silence condition is also, in effect, a text-only ablation—full question and options, a blank (zero) waveform—and it sits at chance.

Table 3: **Workspace readout accuracy under the audio-swap control** (MMAU clips, one prediction per clip). The question and options are byte-identical across conditions; only the waveform changes. Because the correct option's position is imbalanced, we report balanced accuracy (majority guess $= 25\%$); raw accuracy does not exceed the majority baseline and is shown only for completeness. Paired tests are exact two-sided McNemar.

Condition	Raw acc.	Balanced acc.	Paired test vs. real
Real audio	45.0%	**40.0%**	—
Mismatched audio	32.1%	32.2%	p=0.015
Silence	22.9%	21.8%	p=4.7 × 10^{-5}
Position-only baselines: label-permutation null 29%; majority class 45.7%.			

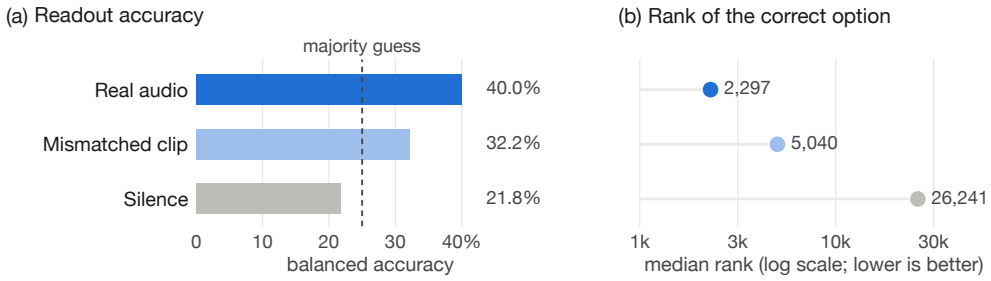


Figure 9: **The audio-swap control. (a)** Balanced readout accuracy by condition; a majority guess scores the dashed line. **(b)** The same ordering without a threshold: median rank of the correct answer among the 152,064 vocabulary tokens, over the workspace band and the audio span. Only the waveform changes between conditions.

D.2 PER-BAND RAW ACCURACY

Table 4 is the raw-accuracy companion to the balanced four-way accuracy of Table 3. Real-audio accuracy rises with depth while the silence baseline dips to chance only in the workspace band, which is why the workspace band is where the sound-driven signal separates most cleanly from the option-token prior.

Table 4: **Raw four-way readout accuracy by functional band** (clip-level; the majority baseline is $\sim$46%). Each band's prediction uses the best rank within that band's layers.

Band (depth)	Real	Mismatch	Silence	Real$-$Silence	McNemar p
Sensory (0–35%)	37.9%	30.7%	35.7%	+2.1	0.78
Workspace (35–80%)	45.0%	32.1%	22.9%	+22.1	4.7×10^{-5}
Motor (80–100%)	51.4%	30.7%	34.3%	+17.1	1.5×10^{-3}

D.3 SPECIFICITY: PLACEBO CONTROL

We compare the clip's topical concept against ten unrelated placebo words using exact top-k membership over the same audio positions and workspace layers (Figure 10). Placebo words are essentially absent, while the concept fills a clear share of cells on every clip except the whip clip, where both are rare; the one visible placebo hit there, *trumpet*, plausibly reflects the resemblance of a whip crack to a brass transient. This shows the surfaced content is clip-specific rather than generic word frequency. It does not by itself show specificity to the correct answer over the clip's topic, which the question and all options share; against that stricter baseline the correct option's literal token does not reliably beat the distractors'. Answer-specificity rests instead on the audio-swap control and on the multi-hop concepts that are absent from the prompt.

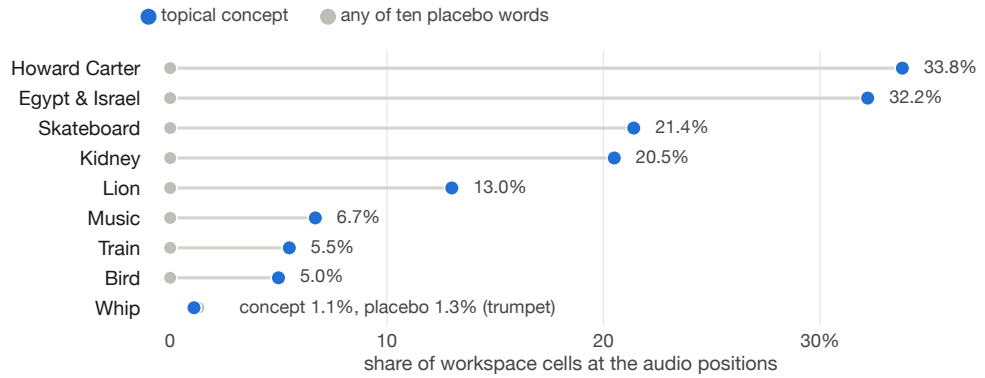


Figure 10: **Placebo control.** For each of nine case clips, the fraction of workspace-band cells over the audio positions that contain the clip's topical concept (blue), against the fraction that contain any of ten unrelated placebo words (grey). The placebo share rounds to zero on every clip except the whip clip, where *trumpet* reaches 1.3%, slightly above the concept.

D.4 THE MULTILINGUAL READOUT: CORPUS TALLY AND FREQUENCY CONTROL

Tallying the language of the top-1 readout over all audio-region workspace cells in the 948-clip readout corpus (restricted to real-word readouts), the inner vocabulary is English-dominant with Chinese a clear second and every other language far behind (Table 5). Because the logit lens may favour frequent CJK tokens, we check that the Chinese forms of *music* and *bird* are clip-specific rather than globally frequent: each form is read far more often on the clips whose content it names than on the rest of the corpus. A globally frequent token would show no such lift.

Table 5: **The multilingual readout. Left:** language of the top-1 readout over $863{,}770$ workspace cells on 948 clips. **Right:** frequency control—how much more often a Chinese form is the top readout on the clips whose content it names than on the rest of the corpus (music-track clips for yīnyuè, birdsong clips for the bird forms).

Top-1 readout language	
English	52.6%
Chinese	38.5%
any other	$\leq 4.3\%$

Clip-specific lift	
yīnyuè (music)	$5.0\times$
niǎolèi (birds)	$9.1\times$
niǎo (bird)	$20.4\times$

D.5 ADDITIONAL CASE STUDIES

Table 6: **Five example cases.** For each we report the strongest (lowest-rank) workspace-band readout of a task-relevant concept over the audio span, the layer (of 48) where it is most stably rank 1, its depth, its rank among the 152,064-token vocabulary, and the number of workspace cells with the concept in the top-8. All five are sound-source identifications.

Answer	Top readout	Layer	Depth	Rank	# ws. hits
Lion	lions	L19	40%	1	37
Bird	niǎolèi (birds)	L33	69%	1	12
Music	música	L18	38%	1	11
Train	rail	L19	40%	1	27
Whip	biān (whip)	L32	67%	1	7

Three regimes. The Nixon case is a single clip, but the readout structure is not unique to it: Table 6 collects five clips that illustrate the range, for which a task-relevant concept is read at rank 1 in the workspace band over the audio positions before any output token. (The population-level claim is the waveform-swap and band statistics of Section 4.3, not these examples.) Together with the clips below, the cases span three regimes. In *multi-hop inference* an intermediate concept absent from the prompt is surfaced over the audio and only then resolved: a treaty clip surfaces negotiation before *Egypt and Israel*, and an archaeology clip surfaces a dig scene, although *Howard Carter* itself is read only at the text positions. In *speech-content recovery* the readout recovers a spoken detail: on a Kidney clip the specific organ, named only in the audio, is read out but reaches rank 1 only in the late layers, so it is not in Table 6, and the question-echoed transplant does not count. In *sound-source identification* (Lion, Bird, Music, Train, Whip) the readout names the source directly from the acoustics. The recurrence of non-English forms at rank 1—niǎolèi (birds), *música* (music)—is the multilingual signature of Section 4.5.

D.6 FREE-ASSOCIATION CONTROLS

A neighbour counts only if it is absent from the clip's transcript on inspection and corpus-rare ($\leq$8 of the 948 clips); literally spoken words are excluded, for example a *crisis* field on a clip that says the word. Because the lens projects onto W_U, a strengthening concept would also lift its embedding neighbours, so we check each associate's cosine similarity to the concept that evokes it: all but one fall outside that concept's top-100 cosine neighbours, and the one near pair (*skiing*, a close neighbour of *skate*) still appears strictly after *skate*, stays below rank 1, and is then dropped—a schedule static similarity does not encode. Two associates in the timetable are not rare words (*subway* and especially *wolf*); for those the control is within-clip specificity: *wolf* fires only on the late howl frames of the lion clip and never enters the answer slot. On the skateboard clip the raw sound is read as gǔndòng ('rolling') before skate forms, and huáxuě follows skate without ever reaching rank 1 (Figures 3 and 11).

(a) Neighbour words read at the audio positions

clip	answer	absent from the clip's transcript, in no question or option; corpus-rare
Archaeology	Howard Carter	burial 棺 coffin 遗址 ruins excavation treasures Egypt
Watergate	Nixon	scandal -gate controversy
Telephone patent	1876	telecommunications wired
Skateboard	Skateboard	滑雪 skiing

(b) When each word is first read

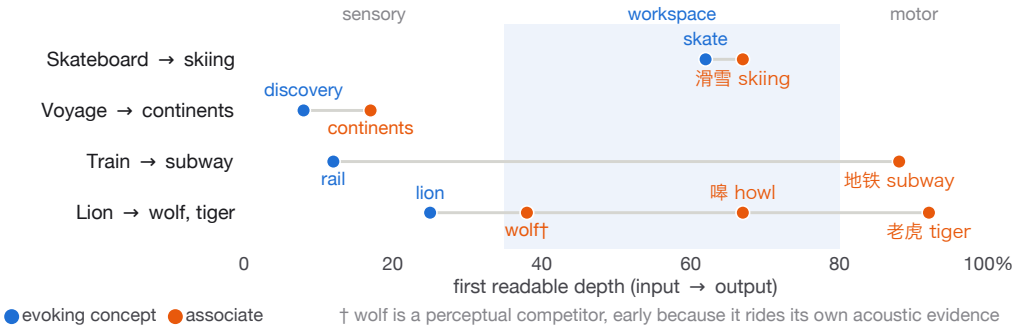


Figure 11: **Semantic neighbours at the audio positions.** **(a)** Spontaneous neighbour words that light up at the audio positions—absent from the clip's transcript on inspection, in no question or option; corpus-rare. **(b)** Timetable: for each family, the first readable layer of the evoking concept (blue) and of each associate (orange); the associate always follows the concept that evokes it.

D.7 HALLUCINATION DETECTION BEYOND ONE CLIP

Catching a hallucination before it is spoken. Reading both minds, their disagreement flags an error before either run speaks. On a clip asking what major event is captured (answer: *an explosion*), the model's own caption describes two people talking and never mentions a blast. Given the audio, explosion becomes readable at the audio positions and reaches rank 1 inside the workspace band; the model answers correctly. Given only the caption, explosion never enters the top-6 readout at a caption position at any depth we read, and the model answers *a conversation*—the wrong option (Figure 12). We rest the claim on this asymmetry and make no aggregate claim about where in depth a wrong answer first appears, since early signal in such traces is largely the wrong option echoed from the prompt, as the clips below show.

The asymmetry in Figure 12 recurs on other clips the model answers correctly from audio and wrongly from its caption: the caption drops the decisive acoustic detail, and the caption mind falls back on what the caption does carry—one clip is answered *traffic noise* instead of the correct *loud braying*, another *music* for a ding—while the audio mind holds the correct concept at the audio positions. A naive trace of the first depth at which the wrong word tops the readout can suggest an early onset, but much of that early signal is the wrong option echoed from the printed choices and from the caption itself; once those input-echo positions are excluded, the wrong concept does not reliably surface early across the set. We therefore report the detection asymmetry and make no aggregate claim about where in depth a fabricated answer originates.

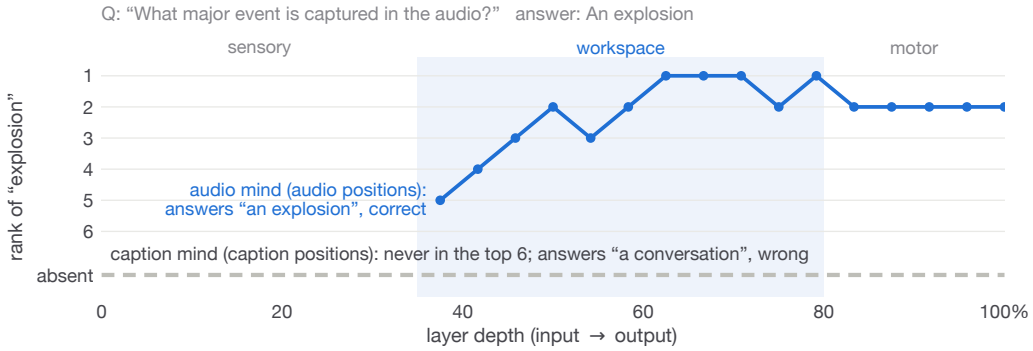


Figure 12: **Catching a hallucination.** One clip ("what major event is captured in the audio?", answer *an explosion*) read two ways: the best rank of explosion in the top-6 readout at every second layer (workspace band shaded). In the audio mind (blue, audio positions) explosion enters in the early workspace, reaches rank 1 inside the band and stays in the top two to the output; the model answers correctly. In the caption mind (grey, positions of the model's own caption) explosion never enters the readout, and the model answers *a conversation*.

D.8 SPOKEN FACT RECALL: WHERE HEARING DOES NOT HELP

We synthesize spoken TriviaQA from the written questions with Qwen3's text-to-speech model. Each question thus exists as a spoken clip and as the identical written question, and the audio carries no extra information. The model answers correctly more often from text than from speech, the expected direction for a lossier route to the same words (Table 7). On the items answered correctly from text and wrongly from speech, the answer concept is present in the speech mind's workspace on most of them and in the text mind's on more still, so "listening never retrieves the answer" is false, but no single mechanism emerges.

Table 7: **Spoken fact recall.** 500 spoken TriviaQA questions, each also supplied as the identical written question. The right column is over the 87 items answered correctly from text and wrongly from speech. At the generation position the final-layer top token matches the generated first token exactly on 31% of runs, is a sub-word prefix of it on 50%, and disagrees on 19%.

	Accuracy	Answer concept in the workspace
From text	65.2%	$72/87$
From speech	56.4%	$54/87$

At the generation position, the final-layer top token, which is the model's own next-token prediction, matches the generated first token exactly on only a minority of runs (Table 7). We did not resolve this mismatch between the readout position and generation, so we make no mechanistic claim from this task and use it only to mark the boundary of the effect: where the sound adds nothing, listening does not help.