LISTEN-TO-REASON: LISTEN WITH EXPERTS, RETRIEVE OVER A GRAPH, REASON WITH LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large audio-language models (LALMs) fuse an audio encoder into a large language model (LLM) through multi-stage training. This coupling means that a new domain or a stronger LLM requires retraining, and their answers cannot be traced to what the model heard: a chain-of-thought is a post-hoc account. We propose Listen-to-Reason (L2R), an interpretable-by-design pipeline that passes audio to the LLM through an explicit, human-readable tree: small heads on frozen expert encoders map each chunk of a clip to semantically meaningful nodes on the tree (for speech, music and environmental sound), and a frozen text-only LLM answers from these nodes and an ASR transcript without hearing the clip. Every answer can therefore be traced to the nodes and transcript it read, and the nodes are causal: replacing the deciding node with a distractor overturns 78% of correct answers on SAKURA. With a 7B reader, L2R outperforms all LALMs we compare against on SAKURA and trails them by 6–12 points on MMAU and MMAR, despite training about 1,400$\times$ fewer parameters on orders of magnitude less audio data. However, because any LLM can serve as the reader, we show that a stronger reader narrows this gap without retraining any audio component. A new domain is added with one small head: with five labelled clips per species, it outperforms QLoRA fine-tuning of an LALM on the same clips by 13–26 points. Examples and an interactive view of the tree are available on our project page.

1 INTRODUCTION

Large audio-language models (LALMs) extend large language models (LLMs) to audio (Cui et al., 2025; Arora et al., 2025). A typical LALM encodes the audio with a pretrained encoder, maps the encoder output into the input space of the LLM through an adaptor, and trains the combined model on audio-text data (Deshmukh et al., 2023; Gong et al., 2024; Tang et al., 2024; Chu et al., 2023; 2024; Ghosh et al., 2024; Kong et al., 2024). They range from models that answer in text (Gong et al., 2023; Hu et al., 2024; Ghosh et al., 2025b;a) to models that also speak (Zhang et al., 2023; Défossez et al., 2024; Zeng et al., 2024; Xie & Wu, 2024; Fang et al., 2025; Wang et al., 2025a; Yu et al., 2025) and omni models that add vision (Fu et al., 2024; Abouelenin et al., 2025; KimiTeam et al., 2025; Huang et al., 2025; Xu et al., 2025a;b).

These abilities come from long, multi-stage training: the audio encoder is first aligned with the LLM on perception data such as automatic speech recognition (ASR) and audio captioning, then tuned on instruction and question-answer pairs, and finally taught to reason with chain-of-thought (CoT) data or reinforcement learning (RL) (Tang et al., 2024; Gong et al., 2024; Ghosh et al., 2025a; Xie et al., 2025; Xu et al., 2025b). Each stage updates the adaptor, and often the encoder and the LLM as well, with large data and compute budgets (Gong et al., 2024; Ghosh et al., 2025a; Xu et al., 2025a;b). The encoder and the LLM are therefore coupled: replacing either one, with a stronger LLM or a better encoder, means repeating the training. Adapting the model to a new domain likewise requires another fine-tuning run, which can degrade what it already knows (Luo et al., 2025).

Recent benchmarks test audio reasoning directly: MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2026), MMAR (Ma et al., 2025b) and SAKURA (Yang et al., 2025b) ask questions over speech, sound and music that need expert knowledge or multi-step inference. To do well on them, LALMs are trained to write a CoT (Wei et al., 2022) before answering, by supervised fine-tuning on reasoning traces or by RL (Xie et al., 2025; Kong et al., 2025; Li et al., 2025; Wen et al., 2025;

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

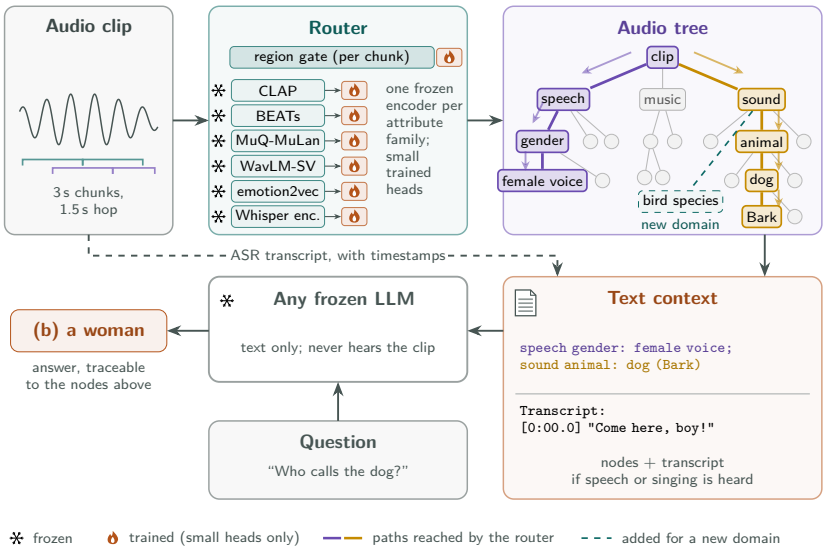

Figure 1: **Overview of LISTEN-TO-REASON. Listen:** a four-level tree describes speech, music and sound. **Retrieve:** router heads on frozen encoders map each chunk to tree nodes, which join an ASR transcript in a readable context. **Reason:** a frozen text-only LLM answers from this context. **Adapt:** a new domain is a new encoder–head branch (dashed).

Rouditchenko et al., 2025; Wu et al., 2026; Wijngaard et al., 2025b; Tian et al., 2025; Xiong et al., 2025), and models with such thinking modes, including Audio Flamingo 3 and Qwen3-Omni, are among the strongest open models (Ghosh et al., 2025a; Xu et al., 2025b). CoT does not always help, however: it can lower accuracy on hard items and shift attention away from the audio (Ma et al., 2025a; Bonzi et al., 2026).

Whether CoT helps or not, it does not explain the answer. In text LLMs, CoT explanations can be unfaithful to the decision they accompany (Turpin et al., 2023; Lanham et al., 2023). In LALMs, reasoning chains often agree with the prediction but are weakly grounded in the audio, and they change under hallucination-inducing and adversarial inputs (Mousavi et al., 2026a). The trace is generated by the same network that makes the decision, after the audio has been fused into its hidden states, so it is a post-hoc account that can change with the prompt or the sampling. Learning what an LALM actually used requires separate analyses, such as alignment scores, layer-wise probes, cue-reliance tests and hallucination benchmarks (Mousavi et al., 2026b; Xiang et al., 2025; Yang et al., 2025a; Chen et al., 2026a; Foo et al., 2026; Kuan et al., 2024; Kuan & Lee, 2025); these sit outside the model, whose own output does not state what it heard.

A different line of work puts the evidence in the model's input. Tool-using LLMs call other models, programs or APIs and read their outputs (Schick et al., 2023; Yao et al., 2023; Shen et al., 2023; Lu et al., 2023; Patil et al., 2024; Qin et al., 2024; Wu et al., 2023; Yang et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023), and audio agents apply this idea to audio understanding (Huang et al., 2024; Wijngaard et al., 2025a; Wang et al., 2026a). Symbolic cascades describe the audio in text and let a text LLM reason over the description (Zeng et al., 2023); closest to our work, SAR-LM converts audio into human-readable features across speech, sound and music and uses them to trace errors, but it concatenates the separate label spaces of off-the-shelf extractors rather than one vocabulary (Taheri et al., 2025). Retrieval-augmented generation (RAG) conditions the answer on retrieved documents (Lewis et al., 2020; Gao et al., 2023), and graph-based RAG retrieves and traverses a graph built over the knowledge source (Edge et al., 2024; Gutiérrez et al., 2024; He et al., 2024; Peng et al., 2025). Retrieval improves accuracy on questions that need outside knowledge, and it improves interpretability, because the evidence the model conditions on is part of its input and can be read.

Knowledge graphs (KGs) give this evidence an explicit structure. They store facts as typed nodes and edges, and combining them with LLMs grounds generation in these facts and exposes the reasoning path (Pan et al., 2024; Jiang et al., 2023; Baek et al., 2023; Sun et al., 2024; Luo et al., 2024). In vision, multimodal KGs, scene graphs and retrieved knowledge support visual question answering

and multimodal reasoning (Liu et al., 2019; Wang et al., 2020; Alberts et al., 2021; Zhu et al., 2024; Krishna et al., 2017; Marino et al., 2019; 2021; Gui et al., 2022; Chen et al., 2022; Hu et al., 2023; Lee et al., 2024; Abootorabi et al., 2025). Audio has few counterparts. iKnow-audio refines the zero-shot predictions of CLAP, a contrastive audio–text model rather than an LLM, with a knowledge graph of relations between sound events (Olvera et al., 2025), and Audiopedia links the transcript of a clip to Wikidata entities and adds their facts to the prompt of an LALM (Penamakuri et al., 2025). Graph retrieval has used audio only as one track of video (Park et al., 2026), and audio RAG retrieves clips by embedding similarity (Chen et al., 2025). To our knowledge, no multimodal KG covers speech, music and environmental sound with nodes that a clip can be routed to directly from its waveform.

Such a graph would make audio understanding interpretable by design. A model is interpretable when a person can see why it made a decision (Doshi-Velez & Kim, 2017; Lipton, 2018). Post-hoc methods explain a trained black box, such as an audio classifier (Paissan et al., 2024b;a), after the fact, and their explanations can misrepresent it: saliency maps can be insensitive to the model's weights, and attention weights are not reliable explanations (Adebayo et al., 2018; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Rudin, 2019). A model that is interpretable by design decides from an intermediate representation a person can read, so the explanation is the computation itself (Rudin, 2019; Koh et al., 2020; Chen et al., 2019; Oikarinen et al., 2023). This matters where errors are costly (Topol, 2019), as in healthcare: clinicians need to check the evidence behind a prediction (Tonekaboni et al., 2019; Amann et al., 2020; Ghassemi et al., 2021), speech is increasingly used to assess depression, Parkinson's and Alzheimer's disease (Cummins et al., 2015; Tsanas et al., 2012; Luz et al., 2020; Low et al., 2020; Fagherazzi et al., 2021; Botelho et al., 2024), and audio models used there should state the evidence behind each decision (Akman & Schuller, 2024; Chen & Hirschberg, 2026).

We present LISTEN-TO-REASON (L2R), which bottlenecks the audio representation through a multimodal knowledge-tree. This tree connects audio to an LLM and makes the evidence behind each answer explicit (Figure 1). The tree describes a clip in named nodes that are grounded in audio, and a frozen text-only LLM answers from these nodes, together with an ASR transcript of any speech or singing, without hearing the clip. The encoders and the LLM communicate only through text, which decouples perception from reasoning: each can be replaced or extended without retraining the other. Our contributions are as follows:

- **Introducing LISTEN-TO-REASON.** We propose L2R, an audio question-answering model in which the audio reaches the LLM only as text: named tree nodes (§2.1) and an ASR transcript. A router of small heads on frozen expert encoders traverses the tree for each chunk of a clip and retrieves the nodes it hears as a readable context (§2.2), and a frozen text-only LLM answers from these nodes and the transcript without hearing the clip (§2.3). Any LLM can serve as the reader: across 26 readers, a stronger one narrows the gap to LALMs with no change to the audio side (§4.2).
- **Interpretability by design.** Every answer is conditioned only on the nodes and transcript printed in the reader's prompt, and each prediction links to a traceable retrieval path through the tree (§2.3). A wrong answer can therefore be located in a node and corrected without touching the reader. The nodes are causal: replacing the deciding node with a distractor overturns 78% of correct answers, and injecting the gold node repairs 91% of wrong ones (§4.4).
- **Few-shot adaptation at low training cost.** A new domain is added as one attribute of the tree, with a frozen expert encoder and a small head trained on a few labelled clips; the reader and the other heads are not updated, so there is no catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017): with both new domains added, accuracy on SAKURA, MMAU and MMAR drops by at most 0.6 points (§2.4, Appendix G). With five clips per species, this outperforms QLoRA fine-tuning of an LALM on the same clips while training 120–240$\times$ fewer parameters.

We evaluate on MMAU, MMAR, SAKURA and MMAU-Pro, and on two new bioacoustic domains (§3, §4). L2R is a transparent alternative to end-to-end LALMs that is cheap to train: it trains 7.7M parameters, about 1,400$\times$ fewer than Qwen2.5-Omni-7B, on orders of magnitude less audio data (§4.5). Demos, code and checkpoints, worked examples and an interactive viewer of the tree are available on our project page.

1.1 Related work

Audio LLMs and audio agents. LALMs range from early task-as-generation models (Deshmukh et al., 2023; Gong et al., 2024) to multi-encoder and mixture-of-experts designs (Tang et al., 2024; Shan et al., 2025; Wang et al., 2025b); their encoder output enters the LLM as continuous embeddings, so adding a domain or changing the LLM requires further fine-tuning, typically with LoRA or QLoRA (Hu et al., 2022; Dettmers et al., 2023). Audio agents keep an LALM in the loop as reasoner or tool, and several train it to plan tool calls (Wijngaard et al., 2025a; Lee et al., 2026; Wang et al., 2026a; Chen et al., 2026b; Wang et al., 2026b). Listen-to-Reason also combines several frozen encoders, but their outputs reach the LLM only as named nodes in text; no LALM runs at inference, the same heads run on every clip without planning, and a new domain is a new head on a frozen encoder rather than a fine-tune, as in few-shot audio classification from prototypes or pretrained embeddings (Snell et al., 2017; Kahl et al., 2021; Ghani et al., 2023; Rauch et al., 2025; Hagiwara et al., 2023); we compare the two routes directly on the same clips.

Text interfaces, concept bottlenecks and retrieval. Cascades describe audio in text for a text LLM, through zero-shot model composition (Zeng et al., 2023), captions (Sakshi et al., 2025; Lu et al., 2026) or agent-refined descriptions (Rong et al., 2025). SAR-LM (Taheri et al., 2025) writes the outputs of off-the-shelf extractors for speech, sound events and music as text, lets a frozen LLM reason over them, and traces errors to features. Listen-to-Reason differs from it in four ways. First, it describes every clip with one hierarchical vocabulary rather than the separate label spaces of several extractors. Second, it predicts the nodes with small heads trained for this vocabulary, which back off from leaf to class when the leaf is uncertain, whereas SAR-LM uses its extractors as they are. Third, it adds a new domain as one attribute with a head trained on a few labelled clips. Fourth, we test its interpretability quantitatively: an audio LLM judge scores how related the selected nodes are to the clip, and interventions on the nodes test whether the answer depends on them.

Concept bottleneck models predict human-defined concepts before the label, so a person can inspect and correct them (Koh et al., 2020; Oikarinen et al., 2023), including for voice-based health assessment (Chen & Hirschberg, 2026); in audio, post-hoc methods instead explain classifiers with listenable components or maps (Parekh et al., 2022; Paissan et al., 2024b;a). Graph-based retrieval reaches knowledge from a text query (Edge et al., 2024; Gutiérrez et al., 2024) or from audio and multimodal keys (Chen et al., 2025; Park et al., 2026; Mousavi et al., 2025). Knowledge graphs have also been paired with audio models: iKnow-audio re-ranks the zero-shot predictions of CLAP with a graph of relations between sound events (Olvera et al., 2025), and Audiopedia adds Wikidata facts about entities named in a clip's transcript to the prompt of an LALM (Penamakuri et al., 2025). Such graphs store facts about sounds or entities, whereas our tree describes what is heard in a clip across speech, music and sound, and the reader supplies the world knowledge. Listen-to-Reason is a concept bottleneck for open-ended audio question answering. The waveform, not a text query, is the entry point; some heads are distilled from teachers that run only at training time (Hinton et al., 2015); and the predictor above the bottleneck is any frozen LLM, which we vary across 26 readers.

2 Method

Listen-to-Reason (Figure 1) achieves interpretable, modular reasoning by constructing a hierarchical multimodal tree (*Listen*), mapping input clips to tree nodes to generate human-readable context (*Retrieve*), using a frozen LLM to answer queries based on this context (*Reason*), and efficiently expanding to new domains by training lightweight projection heads (*Adapt*).

2.1 Listen: Building the Multimodal Tree

We organize audio concepts into a four-level hierarchical tree: *regions* (speech, music, environmental sound), *attributes* characterizing each region (e.g., speaker gender, tempo, sound source), closed *classes* under each attribute (e.g., `angry`, `jazz`, `engine`), and open-vocabulary *leaves* that provide granular refinements (e.g., `engine knocking`). Domain knowledge informs the attributes, while environmental-sound identities align with the AudioSet ontology (Gemmeke et al., 2017).

To populate the nodes, we segment training clips into short chunks and annotate them using a hybrid approach. Each chunk is evaluated by single-task expert classifiers (e.g., speaker verification, emotion recognition) and a Large Audio Language Model. Rather than answering free-form queries, the LALM completes a strict template to fill predefined attributes: it identifies active regions, selects

an existing class (or *unsure*) for each attribute, and optionally appends a brief 1–3 word leaf (e.g., {`is: happy, detail: sarcastic`}).

We merge the expert and LALM predictions depending on whether an attribute allows multiple concurrent choices (e.g., overlapping instruments or sound events) or only a single choice (e.g., gender or tempo). For attributes allowing multiple choices, we simply take the union of all predicted classes. For a single-choice attribute, we use the expert's label when its confidence exceeds a per-attribute threshold, and the LALM's label otherwise. Finally, to maintain a clean taxonomy, free-text leaves are instantiated as nodes only if they recur across multiple clips. An external text LLM merges synonyms based on lexical and acoustic checks. The annotation prompts, the full list of attributes and the leaf rules are given in Appendix A.

2.2 Retrieve: Routing Audio into the Tree

The router maps each short chunk x of a clip to nodes of the tree. Each attribute is read by a frozen pretrained encoder suited to it (e.g., CLAP for sounds and music, emotion2vec for emotion), and attributes that share an encoder form one *router*: a small trainable body ϕ on the frozen embedding $h = E(x)$, with one head g_a per attribute a. A head has one output per node of its attribute and is trained to score highest the node that matches the chunk. For example, the emotion head has outputs for `happy`, `happy (sarcastic)`, `sad`, and so on; for a chunk spoken happily but sarcastically it should score `happy (sarcastic)` highest and `sad` low. The score of a class v combines its own node and its leaves $\mathcal{L}_{a,v}$, so `happy` is scored from both `happy` and `happy (sarcastic)`:

$$s_a(v \mid x) = \log \sum_{n \in \{v\} \cup \mathcal{L}_{a,v}} \exp g_a\big(\phi(h)\big)_n, \tag{1}$$

so a confident leaf also supports its class. Each router is trained on all of its attributes at once. Since a chunk is labelled for only some attributes (e.g., gender but not language), the loss is masked,

$$\mathcal{L}(x) = \sum_a m_a(x) \Big[\ell_a^{\text{class}}(x) + \ell_a^{\text{leaf}}(x) \Big], \tag{2}$$

where $m_a(x) = 1$ only if attribute a is labelled on x, and the two terms are cross-entropies over the classes and over the leaves of the labelled class. At inference, a region gate decides which regions each chunk contains, each open attribute serves its most likely class, and a leaf is added only when it is confident; otherwise the router backs off to the class (e.g., `engine` instead of `engine knocking`). Details are in Appendix B.

2.3 Reason: Answering from the Tree

The router serves nodes for each chunk (§2.2). We collect them over the whole clip into one text context, listing each node once and grouping the nodes of each attribute together. For example, if a woman speaks in some chunks and a dog barks in others, the context reads `speech gender: female voice; sound animal: dog (Bark)`. When speech or singing is heard, we add the ASR transcript, and for questions about the order, duration or number of events, we add the list of detected events with their times. A frozen text LLM then receives this context with the question and answers it. Because the reader only sees text, any LLM can be used without training, and each answer can be traced to, and changed by editing, the lines of its context.

2.4 Adapt: Adding a Domain

A new domain, such as bird species, becomes a new attribute under a region. It needs a frozen encoder suited to the domain, either one of the existing encoders or a new one, and a head over that encoder trained on k labelled clips per class. We also train the head on clips without birds, so it learns to give them a low bird probability and the router does not enter the new domain for them.

3 Experimental Details

Data. The tree is generated from 4,054 training clips from 19 sources, labelled by expert models and an audio LLM annotator (Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)); the router heads also use human labels from AudioSet, FSD50K, ESC-50, MusicCaps, IRMAS and FLEURS (Appendix C). The tree has 39 attributes (speech 6, music 18, sound 15), 363 classes and 668 leaves (Appendix A).

Table 1: **Accuracy (%) on four audio reasoning benchmarks**, full sets. Our readers never hear the clip; they read the same cached tree and transcript. B = parameters in billions (for audio LLMs, total size from the released weights). On MMAU-Pro the options are shuffled per item for all systems (Appendix D).

		B	SAKURA↑	MMAU↑	MMAR↑	MMAU-Pro↑
Input ablation (Qwen2.5-7B)	Blind (question and options only)	7.6	32.6	50.2	37.8	38.9
	Transcript only	7.6	51.8	60.1	47.8	44.6
	Tree only	7.6	74.7	57.5	40.3	39.3
L2R (ours) (tree + transcript) by reader	Qwen2.5-7B	7.6	80.1	68.3	51.6	45.7
	Qwen2.5-14B	14.7	83.3	69.4	55.6	48.5
	Qwen2.5-32B	32.8	85.1	71.4	55.7	49.8
	Qwen2.5-72B	72.7	85.3	70.5	56.6	50.1
	Phi-4	14.7	83.5	71.7	54.9	46.1
Audio LLMs (hear the clip)	Qwen2.5-Omni-7B	10.7	73.3	75.1	63.6	57.4
	Audio Flamingo 3	8.3	68.2	75.5	57.8	49.1
	Qwen3-Omni-30B-A3B	35.3	73.0	74.0	62.0	58.2

Benchmarks. We evaluate on four multiple-choice audio reasoning benchmarks, scored by accuracy: MMAU test-mini (Sakshi et al., 2025) (1,000 items over sound, music and speech), MMAR (Ma et al., 2025b) (1,000 items over speech, sound, music and their mixtures), SAKURA (Yang et al., 2025b) (4,000 single- and multi-hop items on gender, language, emotion and animal), and the single-audio multiple-choice subset of MMAU-Pro (Kumar et al., 2026) (4,133 items). For domain adaptation we build two species-identification tasks on which current LALMs are near chance: birds (48 species, 938 test items, from BirdSet (Rauch et al., 2025)) and marine mammals (31 species, 678 test items, from BEANS (Hagiwara et al., 2023)). Each item is a multiple-choice question with four options, the correct species and three taxonomically close ones, so chance is 25% (details in Appendix D).

Readers and baselines. Because the router's output is cached as text, the same contexts are evaluated by 26 frozen text LLMs from 0.5B to 72B parameters, from the Qwen2.5, Qwen3, Llama, Vicuna, Mistral, Phi, OLMo, Falcon, SmolLM and gpt-oss families (Appendix F). Transcripts come from Whisper-large-v3 (Radford et al., 2023). We compare against three LALMs that hear the clip: Qwen2.5-Omni-7B (Omni-7B) (Xu et al., 2025a), Audio Flamingo 3 (AF3) (Ghosh et al., 2025a) and Qwen3-Omni-30B-A3B (Xu et al., 2025b). Input ablations give a reader the question only (*blind*), the transcript only, the tree only, or both. For new domains, we compare our head with three uses of the same labelled clips by Omni-7B: zero-shot, clips given in context, and QLoRA fine-tuning (Dettmers et al., 2023). Training and inference details are in Appendix B.

4 Results

4.1 Main evaluation on reasoning benchmarks

Table 1 compares Listen-to-Reason, read by five frozen text LLMs, with three audio LLMs that hear the clip. With the Qwen2.5-7B reader, Listen-to-Reason outperforms every audio LLM on SAKURA (80.1 against 73.3 for the best, Omni-7B) and trails them on the other three benchmarks, by 6.8 points to Omni-7B on MMAU and by 12.0 on MMAR. Note that these gaps come with a model whose every answer can be traced to the nodes it read, and which trains about 1,400× fewer parameters than Omni-7B on a training set orders of magnitude smaller (§4.4, §4.5). Measured from a blind reader that sees only the question and options (50.2 on MMAU, 37.8 on MMAR), Listen-to-Reason recovers nearly three quarters of Omni-7B's gain on MMAU, about half on MMAR and about a third on MMAU-Pro, and more than all of it on SAKURA.

The input ablation shows what each part of the context contributes. The tree carries the identity of what is heard: on SAKURA, whose questions ask for an animal, emotion, gender or language, the tree alone reaches 74.7 and adds 28.3 points over the transcript. The transcript carries spoken content, which matters on MMAU and MMAR, where questions mix speech with sound and music. The two channels are complementary: together they outperform either one alone on every benchmark.

Because the audio side is fixed, a stronger reader narrows these gaps without any retraining. Reading the same cached contexts, Phi-4 closes half of the MMAU gap to Omni-7B (71.7), Qwen2.5-72B

closes about 40% of the MMAR gap (56.6), and the 32B and 72B readers pass AF3 on MMAU-Pro (49.8 and 50.1 against 49.1). We study the reader in §4.2 and the question types behind the remaining gap in Appendix H.

4.2 SCALING THE READER

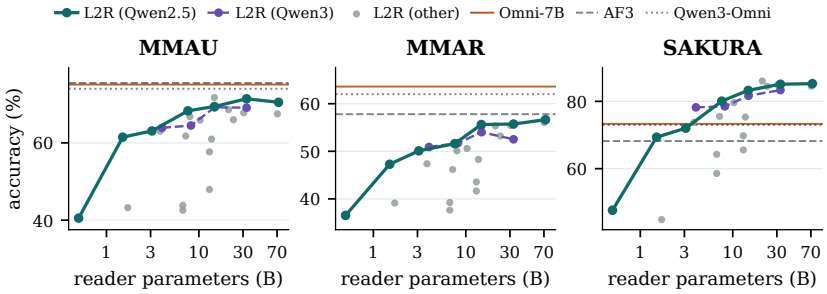

Figure 2: **Accuracy of LISTEN-TO-REASON as a function of the frozen reader.** Every reader reads the same cached tree and transcript contexts. Lines connect the Qwen2.5 (solid) and dense Qwen3 (dashed) families; grey points are the other 15 readers (Llama, Vicuna, Mistral, Phi, OLMo, Falcon, SmolLM, gpt-oss and the mixture-of-experts Qwen3-30B-A3B). Horizontal lines are audio LLMs that hear the clip. Full table in Appendix F.

The router and the ASR run once per clip and their output is cached, so any text LLM can read the same context, and differences between the 26 readers are differences in reasoning over a fixed description of the audio (Figure 2, Appendix F). Accuracy rises steeply with size up to 7B and more slowly beyond: within the Qwen2.5 family, the 7B reader already reaches 68.3 on MMAU and 80.1 on SAKURA, MMAR gains the most from 7B to 14B (4.0 points, to 55.6) because more of its questions require multi-step inference over the context, mostly the transcript, and above 14B all three benchmarks change by at most two points.

The generation of the reader matters as much as its size: Phi-4 (14.7B) is the best reader on MMAU (71.7), above Qwen2.5-72B (70.5), while Llama-2 and Vicuna of similar size stay close to their blind scores on MMAU and MMAR, barely using the identical context. SAKURA shows how much the context alone carries: every reader released since 2024 with at least 7B parameters exceeds Omni-7B, which hears the clip. For single-hop questions the answer is usually the name of a node; for multi-hop questions the node name is enough because the reader's own knowledge supplies the rest: once the node identifies the animal as a dog, a follow-up question about dogs needs no further evidence from the audio. Stronger readers can therefore be adopted as they appear, and each one reads the same cached description.

4.3 FEW-SHOT ADAPTATION TO NEW DOMAINS

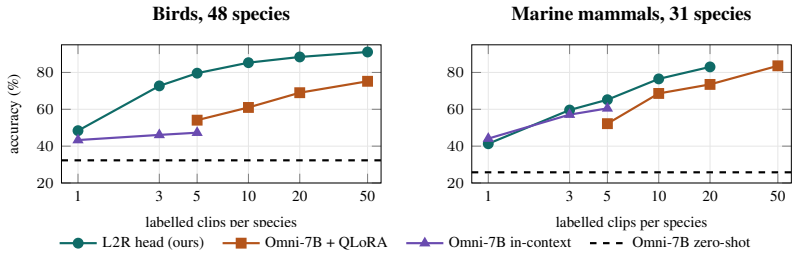

Figure 3: **Data efficiency on new domains.** One small head (0.34M parameters for birds, 0.17M for marine mammals) on a frozen encoder (BirdNET for birds, CLAP for marine mammals), read by a frozen Qwen2.5-7B, against Omni-7B fine-tuned with QLoRA on the same clips and Omni-7B given the clips in context.

Next, we look at how L2R adapts to a new domain from a few labelled clips, a setting in which an LALM would normally have to be fine-tuned. We use the two species-identification tasks of §3, birds (48 species) and marine mammals (31 species), in which each question offers four candidate species. Both are out of distribution for current audio LLMs: Omni-7B scores 32.3 and 25.8, barely above

chance (25). LISTEN-TO-REASON adds each domain by training one head on a frozen encoder suited to it: BirdNET (Kahl et al., 2021), trained on bird vocalisations, for birds, and CLAP, already used by the router, for marine mammals. The heads have 0.34M and 0.17M parameters and train in under a minute on one GPU; the tree, the other heads and the reader stay unchanged. We compare this head with two ways of giving the same clips to Omni-7B: fine-tuning it with QLoRA, and providing a few example clips with their labels in its prompt, known as in-context learning, which the LALM's context length limits to a few clips (at most five per candidate species here).

This head makes far better use of a few labelled clips than fine-tuning the LALM (Figure 3). With five clips per species, it reaches 79.6 on birds and 65.2 on marine mammals, against 54.1 and 52.2 for a QLoRA fine-tune of Omni-7B on the same clips, which trains 120–240× more parameters (40.4M) and takes 7–29 GPU-minutes per point. The advantage persists as data grows: on birds, QLoRA with 50 clips per species (75.2) stays below our head with five, and on marine mammals it needs 50 clips to match what our head reaches with 20 (83.6 against 83.0). Giving Omni-7B the clips in context helps less than QLoRA on birds and falls between the two methods on marine mammals, whose calls are short enough for the prompt to hold whole calls.

Because the domain node carries the answer, the reader barely matters: with the five-clip head, 24 readers from 1.5B to 72B all score 78–82 on birds and 64–68 on marine mammals. Adding a domain also leaves the original tasks intact: with both heads added, the Qwen2.5-7B reader scores 79.6 on SAKURA (80.1 without), 67.7 on MMAU (68.3) and 51.8 on MMAR (51.6); exact numbers are in Appendix G.

4.4 INTERPRETABILITY

We now turn to interpretability. Every answer of LISTEN-TO-REASON is conditioned only on text that a person can read, the served nodes and the transcript, so the explanation of an answer is its own input. For this explanation to be useful, the nodes must describe the audio correctly, and the answer must depend on them. We test the first with audio LLMs as judges and the second with interventions on the nodes, and illustrate both with examples. An interactive page plays example clips and shows the router's path through the tree, chunk by chunk, with the probability of every node and the text the reader receives.[1]

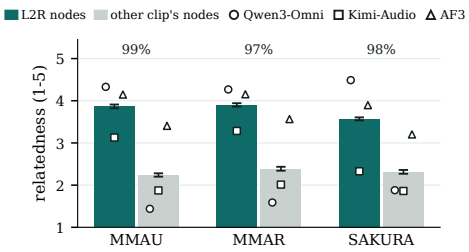

Figure 4: **Relatedness of the served nodes to the audio** (1–5). Bars: mean of three judges with 95% bootstrap intervals; markers: each judge; numbers: share of items where the L2R nodes score higher.

Audio LLMs as judges. To check that the served nodes describe the audio, three audio LLMs (Qwen3-Omni-30B-A3B, Kimi-Audio-7B and Audio Flamingo 3) hear each clip and rate, on a 1–5 scale, how related the served nodes are to what is audible, without seeing the question. They rate 1,200 held-out items (400 each from MMAU, MMAR and SAKURA, 100 per SAKURA track), and, as a control, the nodes of a clip from a different category (e.g. music nodes for a speech clip). All three judges rate the served nodes well above the swapped ones on every benchmark, and higher on 97–99% of items (Figure 4). They differ in scale, one rating even the swapped nodes above the midpoint, but they agree in direction; the prompt and protocol are in Appendix I.

Case study. Figure 5 in Appendix J shows what the reader actually receives for three MMAU items, and our project page shows many more similar examples. In the two correct items, the node that decides the answer is present, and the reader picks the option that names it. In the wrong item, the router reached the `water` class but served `Rain` instead of `Waterfall`, and the reader followed it, so the error can be traced to a single node, which is where a fix belongs.

The nodes are causal for the answer. The judges show that the nodes fit the audio; interventions show that the answer depends on them. When we serve the tree of another clip together with the item's own transcript, Qwen2.5-7B drops from 68.3 to 55.0 on MMAU and from 51.6 to 46.4 on MMAR, below its transcript-only scores (60.1 and 47.8): the reader uses the nodes it is given, and a wrong tree is worse than none. On SAKURA, with the tree of a clip from a different class, accuracy

[1] `https://anonymous.4open.science/w/listen-to-reason/index.html#examples`

falls from 80.1 to 40.0; gender falls to 11.7%, far below chance, because the reader follows the wrong gender node. To isolate single nodes, we edit the context of SAKURA single-hop items, whose question asks for exactly one attribute, and change only that attribute's clause. Replacing the routed node with the node of a distractor option breaks 78.3% of the correct answers of Qwen2.5-7B, and replacing it with the gold node repairs 90.9% of the wrong ones. For animal, emotion and gender, the swap breaks 93.8%, 96.9% and 99.8% of correct answers, and the inject repairs 87.5%, 94.8% and 97.8% of wrong ones. Language is the exception: the swap breaks only 8.0%, because the transcript also shows which language is spoken, and the reader follows the words over a contradicting node.

4.5 Efficiency and resource use

Beyond accuracy, we compare what Listen-to-Reason and the audio LLMs cost (Table 13 in Appendix K): training cost decides how easily a system can be rebuilt, extended or moved to a new LLM, and inference cost decides what it takes to deploy it. Parameters are summed from the released weights, without the pretraining data of the components each system starts from; MACs (PyTorch FLOP counter) and seconds are averaged over 20 MMAU items on one A100 GPU.

Listen-to-Reason is far cheaper to train, in both parameters and data. It trains 7.7M parameters, about 1,400$\times$ fewer than Omni-7B, and it learns from a training set orders of magnitude smaller than those of the audio LLMs: about 0.2M labelled clips in total. The audio LLMs, in contrast, train every parameter on vast audio corpora: Omni-7B uses 300 billion tokens of audio-related data in its second pretraining stage alone (Xu et al., 2025a), Audio Flamingo 3 uses 9.6M question–answer pairs and 13.2M audio–caption pairs (Ghosh et al., 2025a), and Qwen3-Omni trains its audio encoder on 20 million hours of audio and its first pretraining stage on 0.77 trillion audio tokens (Xu et al., 2025b). The saving carries over to extending the system: adding a domain updates 120–240$\times$ fewer parameters than QLoRA, and because the router's output is cached, reading the same audio with another LLM costs only that LLM's forward pass.

Inference, in contrast, is more expensive than for a dense audio LLM of similar size: per MMAU item, Listen-to-Reason costs 5.3 TMACs and 0.57 s, against 2.7 TMACs and 0.12 s for Omni-7B. Most of this cost lies outside the tree: Whisper-large-v3 accounts for half of it (2.6T) and the reader's two passes, with and without the context, for 2.1T, while all six router encoders together take 0.6T; the encoders also run one after another, which adds to the time. These figures compare a pipeline assembled from off-the-shelf components, run as they are, with audio LLMs built and tuned for inference as single networks. The largest share of our cost, the ASR model, could be replaced by a smaller or distilled one, and the encoders could run in parallel.

5 Conclusion

Listen-to-Reason replaces the fused audio representation of an audio language model with an explicit audio tree and a frozen text reader: small heads on frozen encoders map each chunk of a clip to named nodes, and a text LLM that never hears the clip answers from these nodes and a transcript. Every answer can therefore be traced to the nodes it received, and audio LLMs judge those nodes as related to what is audible. With a small reader, Listen-to-Reason outperforms audio LLMs on questions about the identity of a sound or speaker, and trails them with a relatively small gap on open reasoning over sound and music, despite being trained a small dataset, and having fewer trainable parameters. The tree and the transcript are complementary, a stronger reader narrows the gap without changing any audio component, and a new domain is added with one small head and a few labelled clips, more effectively than by fine-tuning an audio LLM.

Limitations and future work. The tree has a closed vocabulary and describes short chunks rather than individual events, so it cannot yet express counts, the order of events, musical structure or the overall setting of a scene; these questions fall back on the transcript and the reader's prior, and they account for most of the remaining gap. Event-level nodes and signal-level experts, such as music transcription, beat tracking and speaker diarization, can be added to the modular tree as new attributes. The tree also inherits the errors of the annotators and teacher classifiers used to build it, and the reader follows a wrong node rather than correcting it; better heads and a reader that weighs node confidence are natural next steps. Finally, inference runs several encoders and an ASR model before the reader, and domain addition has been tested only on identification tasks; lighter front ends and domains that require reasoning beyond identity remain to be explored.

ETHICS STATEMENT

This work uses only publicly available datasets and models, under their licenses, and collects no new data from human participants; the two bioacoustic evaluation sets are drawn from BirdSet and the Watkins collection. Some attributes of the tree describe speakers, such as gender, age group and emotion; such attributes can be misused to profile people, can be wrong for individuals, and follow the categories of their training data, which do not represent every speaker. Because the tree was built partly from audio LLM annotations, it can also inherit their biases, although every answer is conditioned on readable nodes, so such errors can be inspected and corrected. The healthcare applications that motivate this work were not evaluated, and L2R is not validated for clinical use.

REPRODUCIBILITY STATEMENT

The tree construction and annotation are described in §2.1 and Appendix A, the router, the heads and all hyperparameters in §2.2 and Appendix B, the datasets and splits in Appendix C, and the benchmarks, scoring and option shuffling in Appendix D. Results for every reader and every number of labelled clips are in Appendices F and G, and the judge prompt and protocol in Appendix I. All encoders, readers and baselines are publicly released models, no benchmark audio is used for training, and random seeds are fixed for data splits, head training and template selection. The code repository and all checkpoints needed to reproduce our results, including the tree and the trained router heads, are released on our project page.

AI USE STATEMENT

Two components of the method use generative models, and both are described and evaluated in the paper: an audio LLM (Qwen3-Omni-30B-A3B-Instruct) annotates the training chunks from which the tree is built, and a text LLM (Llama-3.3-70B-Instruct) proposes leaf groups, which are accepted only after lexical, acoustic and identity checks (§2.1, Appendix A); three audio LLMs also serve as judges of node relatedness (§4.4, Appendix I). In preparing the paper, we used an AI assistant to edit and draft passages of the text, suggest section structure, check descriptions of related work against the cited papers, and format bibliography entries. We also used it for generic tasks, such as generating figures from our results and preparing reports of the experiments; we double-checked all such output and confirmed its validity. The authors reviewed all AI-assisted text, checked every number against the experimental results and every citation against its source, and take full responsibility for the content.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, et al. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, et al. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Alican Akman and Björn W. Schuller. Audio explainable artificial intelligence: A review. *Intelligent Computing*, 3:0074, 2024.

Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, et al. VisualSem: A high-quality knowledge graph for vision and language. In *Proceedings of the Workshop on Multilingual Representation Learning (MRL)*, pp. 138–152, 2021.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, 2020.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, et al. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*, 2025.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pp. 78–106, 2023.

Francesco Bonzi, Pooneh Mousavi, Cem Subakan, and Mirco Ravanelli. Enhancing audio reasoning via semantic summary prediction. *arXiv preprint arXiv:2609.20849*, 2026.

Catarina Botelho, Alberto Abad, Tanja Schultz, and Isabel Trancoso. Speech as a biomarker for disease detection. *IEEE Access*, 12:184487–184508, 2024.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, et al. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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, 2026a.

Liyang Chen, Hongkai Chen, Yujun Cai, Sifan Li, Qingwen Ye, and Yiwei Wang. AudioRouter: Data efficient audio understanding via RL based dual reasoning. *arXiv preprint arXiv:2602.10439*, 2026b.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5558–5570, 2022.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. WavRAG: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12505–12523, 2025.

Yu-Wen Chen and Julia Hirschberg. An audio language model-based voice concept bottleneck framework for interpretable health assessment. *arXiv preprint arXiv:2607.16967*, 2026.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, et al. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, et al. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, et al. Recent advances in speech language models: A survey. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13943–13970, 2025.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, et al. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, et al. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 18090–18108, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 10088–10115, 2023.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1):78–88, 2021.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, et al. LLaMA-Omni: Seamless speech interaction with large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Leonardo Haw-Yang Foo, Chih-Kai Yang, Chen-An Li, Ke-Han Lu, and Hung-yi Lee. All that glitters is not audio: Rethinking text priors and audio reliance in audio-language evaluation. *arXiv preprint arXiv:2604.24401*, 2026.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, et al. VITA: Towards open-source interactive omni multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, et al. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13:22876, 2023.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11): e745–e750, 2021.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, et al. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6288–6313, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, et al. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 41819–41886, 2025a.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, et al. Audio Flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pp. 19358–19405, 2025b.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, Think, and Understand. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G. Hauptmann, Yonatan Bisk, et al. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 956–968, 2022.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 59532–59569, 2024.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. BEANS: The benchmark of animal sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, et al. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, et al. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572, 2024.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, et al. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23369–23379, 2023.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, et al. Step-Audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, et al. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 23802–23804, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3543–3556, 2019.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, et al. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9237–9251, 2023.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, et al. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pp. 5338–5348, 2020.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, et al. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 25125–25148, 2024.

Zhifeng Kong, Arushi Goel, João Felipe Santos, Sreyan Ghosh, Rafael Valle, et al. Audio Flamingo Sound-CoT technical report: Improving chain-of-thought reasoning in sound understanding. *arXiv preprint arXiv:2508.11818*, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. In *Proceedings of Interspeech*, pp. 4144–4148, 2024.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, et al. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 22688–22697, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 10767–10782, 2024.

Kuan-Yi Lee, Tsung-En Lin, and Hung-Yi Lee. Audio-Maestro: Enhancing large audio-language models with tool-augmented reasoning. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, et al. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43, 2018.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, et al. MMKG: Multimodal knowledge graphs. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 459–474, 2019.

Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1): 96–116, 2020.

Ke-Han Lu, Szu-Wei Fu, Chao-Han Huck Yang, Zhehuai Chen, Sung-Feng Huang, et al. How auditory knowledge in LLM backbones shapes audio language models: A holistic evaluation. *arXiv preprint arXiv:2603.19195*, 2026.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, et al. Chameleon: Plug-and-play compositional reasoning with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 43447–43478, 2023.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In *Proceedings of Interspeech*, pp. 2172–2176, 2020.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–6, 2025a.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14111–14121, 2021.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Pooneh Mousavi, Shubham Gupta, Cem Subakan, and Mirco Ravanelli. LiSTEN: Learning soft token embeddings for neural audio LLMs. In *Proceedings of Interspeech*, pp. 3244–3248, 2025.

Pooneh Mousavi, Lovenya Jain, Mirco Ravanelli, and Cem Subakan. Investigating faithfulness in large audio language models. In *Proceedings of Interspeech*, 2026a.

Pooneh Mousavi, Yingzhi Wang, Mirco Ravanelli, and Cem Subakan. ALAS: An automatic latent alignment score for audio language models. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2026b.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Michel Olvera, Changhong Wang, Paraskevas Stamatiadis, Gaël Richard, and Slim Essid. iKnow-audio: Integrating knowledge graphs with audio-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 34683–34700, 2025.

Francesco Paissan, Luca Della Libera, Mirco Ravanelli, and Cem Subakan. Listenable maps for zero-shot audio classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024a.

Francesco Paissan, Mirco Ravanelli, and Cem Subakan. Listenable maps for audio classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 39009–39021, 2024b.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36 (7):3580–3599, 2024.

Jayneel Parekh, Sanjeel Parekh, Pavlo Mozharovskyi, Florence d'Alché Buc, and Gaël Richard. Listen to Interpret: Post-hoc interpretability for audio networks with NMF. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 35270–35283, 2022.

Hyeongcheol Park, Jiyoung Seo, Jaewon Mun, Hogun Park, Wonmin Byeon, Sung June Kim, Hyeonsoo Im, JeungSub Lee, and Sangpil Kim. M^3KG-RAG: Multi-hop multimodal knowledge graph-enhanced retrieval-augmented generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14968–14978, 2026.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 126544–126565, 2024.

Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. Audiopedia: Audio QA with knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, et al. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 28492–28518, 2023.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, et al. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Yan Rong, Chenxing Li, Dong Yu, and Li Liu. AudioGenie-Reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning. *arXiv preprint arXiv:2509.16971*, 2025.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, et al. Omni-R1: Do you really need audio to fine-tune your audio LLM? In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 68539–68551, 2023.

Weiqiao Shan, Yuang Li, Yuhao Zhang, Yingfeng Luo, Chen Xu, et al. Enhancing speech large language models with prompt-aware mixture of audio encoders. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 19305–19320, 2025.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 38154–38180, 2023.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4077–4087, 2017.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, et al. Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

Termeh Taheri, Yinghao Ma, and Emmanouil Benetos. SAR-LM: Symbolic audio reasoning with large language models. *arXiv preprint arXiv:2511.06483*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, et al. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, et al. Step-Audio-R1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 106, pp. 359–380, 2019.

Eric J. Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Jennifer Spielman, and Lorraine O. Ramig. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 74952–74965, 2023.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159, 2020.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, et al. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pp. 63345–63354, 2025a.

Yucheng Wang, Jing Peng, Hanqi Li, Chenghao Wang, Wenming Tu, et al. Audio-Mind: An auditable agentic framework for audio understanding. *arXiv preprint arXiv:2605.28480*, 2026a.

Yuwen Wang, Tian-Hao Zhang, Minghao Cai, Yilin Ren, Ziyang Jiang, et al. Hear, invoke, and understand: A skill-calling multimodal agent for large audio language models. *arXiv preprint arXiv:2608.01881*, 2026b.

Ziqian Wang, Xianjun Xia, Xinfa Zhu, and Lei Xie. U-SAM: An audio language model for unified speech, audio, and music understanding. In *Proceedings of Interspeech*, pp. 2720–2724, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. SARI: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. AudioToolAgent: An agentic framework for audio-language models. *arXiv preprint arXiv:2510.02995*, 2025a.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. AudSemThinker: Enhancing audio-language models through reasoning over semantics of sound. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, et al. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, et al. Audio-Thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 33962–33970, 2026.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5187–5202, 2025.

Zhifei Xie and Changqiao Wu. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, et al. Audio-Reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 23829–23851, 2025.

Zhen Xiong, Yujun Cai, Zhecheng Li, Junsong Yuan, and Yiwei Wang. Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models. *arXiv preprint arXiv:2509.21749*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

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 *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025a.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. SAKURA: On the multi-hop reasoning of large audio-language models based on speech and audio information. In *Proceedings of Interspeech*, pp. 1788–1792, 2025b.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, et al. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, et al. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, et al. SALMONN-omni: A standalone speech LLM without codec injection for full-duplex conversation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 24617–24643, 2025.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, et al. Socratic Models: Composing zero-shot multimodal reasoning with language. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, et al. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, et al. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735, 2024.

APPENDIX

A TREE CONSTRUCTION DETAILS

This appendix follows the steps of §2.1: the attributes of the tree (A.1), the annotation of training chunks (A.2) and the processing of free-text leaves (A.3).

A.1 ATTRIBUTES

Table 2 lists the attributes of each region, the number of classes and leaves, whether each attribute is decided once per clip or per chunk, and which expert classifier labels it. Table 3 lists the identity attributes of environmental sound, whose classes follow the AudioSet ontology; their heads are trained on human labels from AudioSet, FSD50K and ESC-50, and the activity attribute also on the predictions of a VGGSound classifier.

Table 2: **Attributes of the tree.** * = multiple-choice attribute. Router encoder = frozen encoder whose embedding the attribute's head reads at training and inference. Expert labeller = expert classifier that labels the attribute when the tree is built, with its confidence threshold; it does not run at inference. Human labels = datasets whose labels are added to the training targets. †When a transcript is available, language is predicted from its text. New-domain attributes are added after the tree is built, each trained on k labelled clips per species plus clips without that domain.

Region	Attribute	Classes	Leaves	Decided per	Example classes	Router encoder	Expert labeller	Human labels
all	regions present*	3	–	chunk	speech, music, environmental sound	CLAP	PANNs (0.30)	–
speech	gender	2	–	chunk	male voice, female voice	WavLM-SV	WavLM-SV (0.05)	–
	age group	4	–	chunk	child, young adult, elderly adult	Whisper	–	–
	language	7	50	clip	western european (`french`), south asian	Whisper†	Whisper-LID (0.50)	–
	number of speakers	3	–	clip	one, two, three or more	Whisper	–	–
	emotion	7	50	chunk	angry, happy, sad, neutral	emotion2vec	emotion2vec (0.60)	–
	speaking style	8	67	chunk	conversational, public address, shouting	Whisper	–	–
music	instrument*	22	71	chunk	acoustic guitar, cello, drum kit	CLAP	–	IRMAS, MusicCaps
	vocals	6	27	chunk	female vocals, choir, no vocals	CLAP	–	MusicCaps
	vocal style*	16	–	chunk	pop or rock singing, breathy, harmonized	CLAP	–	MusicCaps
	ensemble size	5	–	clip	solo instrument, full band, orchestra	CLAP	–	–
	genre	19	31	clip	blues, classical, electronic	CLAP	–	–
	mood	10	42	clip	calm, dark, energetic	CLAP	–	MusicCaps
	tempo	4	–	clip	slow, moderate, fast	MuQ-MuLan	librosa (0.30)	–
	meter	4	–	clip	duple, triple, irregular	MuQ-MuLan	–	–
	mode	2	–	clip	major key, minor key	CLAP	librosa (0.80)	–
	texture	6	14	clip	single melodic line, dense layered	MuQ-MuLan	–	–
	melody	7	33	chunk	ascending line, repetitive motif	MuQ-MuLan	–	–
	rhythm feel	6	–	clip	syncopated, swung, march-like	MuQ-MuLan	–	–
	dynamics	6	–	clip	steady, builds up, fades out	MuQ-MuLan	–	–
	form	7	–	clip	verse and chorus, loop-based	MuQ-MuLan	–	–
	fidelity	5	–	clip	clean, noisy, muffled recording	CLAP	–	MusicCaps
	production	5	24	clip	studio, live performance, lo-fi	CLAP	–	MusicCaps
	era / culture	10	6	clip	baroque, celtic, east asian	MuQ-MuLan	–	–
	period	5	–	clip	contemporary popular, romantic era	MuQ-MuLan	–	–
sound	scene	22	77	clip	airport, farm, forest	CLAP	–	–
	action / material	10	104	chunk	footsteps, breaking, air flow	CLAP	–	–
	proximity	4	–	chunk	close, at a distance, approaching	CLAP	–	–
new domain (§4.3)	bird species	48	–	clip	northern cardinal, American crow, hermit thrush	BirdNET	–	BirdSet (k per species)
	marine-mammal species	31	–	clip	beluga, killer whale, bearded seal	CLAP	–	Watkins (k per species)

A.2 ANNOTATION

Each training chunk is labelled by the expert classifiers of Table 2 and by the LALM (Qwen3-Omni-30B-A3B-Instruct). The LALM hears the first 30 s of a clip and fills the template in two steps: it first says which regions it hears, then answers one template per heard region, choosing a listed class, *unsure* or *none* for each attribute and optionally a short leaf. Its system message is *"You are a careful audio annotator. You describe only what can be heard, and you say unsure when you cannot tell."* The box shows the first step and the speech template, abbreviated to four attributes.

Table 3: **Identity attributes of environmental sound** (all multiple-choice), following the AudioSet ontology. Each attribute is read by CLAP or BEATs, whichever scores higher on held-out AudioSet and FSD50K clips. The new-domain heads of §4.3 read BirdNET (birds) and CLAP (marine mammals).

Attribute	Classes	Leaves	Router encoder	Example classes
animal	26	96	BEATs	bird, cat, cattle, dog, insect
vehicle	12	40	CLAP	aircraft, boat, car, emergency vehicle, train
machine	2	37	CLAP	engine, mechanisms
tool	7	12	CLAP	hammer, jackhammer, power tool, sanding
household	23	38	CLAP	blender, dishes, door, tap or sink, typing
human non-speech	24	58	BEATs	applause, breathing, cough, laughter
nature	5	25	CLAP	fire, thunderstorm, water, wind
alarm and signal	13	34	CLAP	alarm clock, beep, bell, siren, telephone
impact	7	51	CLAP	explosion or gunshot, glass, liquid, surface contact
environment	8	10	BEATs	echo, inside a small room, noisy background
sound type	18	23	CLAP	clang, clicking, creak, chanting
activity (VGGSound)	10	34	BEATs	cooking or eating, construction or repair, office work

Step 1 (regions): Listen to the audio. Which of these are clearly audible? Answer with a JSON object only, e.g. {`"speech": true, "music": false, "sound": true`}. "sound" means environmental or object sounds other than speech and music. SINGING is part of music, not speech: set "speech" to true only when someone is TALKING.

Step 2 (form, one per present region): Listen to the audio and fill in this form about the speech in it. For each field choose exactly one of the listed values (copy it verbatim), or "unsure" if you cannot tell from the audio, or "none" if the field does not apply. Do not guess. Where a field is an object, "is" is the listed value and "detail" is the SINGLE most specific term (1–3 plain words of YOUR OWN, no "and") for what you hear under that value. No numbers in a detail. An empty detail `""` is the RIGHT answer whenever you can add nothing specific.
– `"gender"`: male voice | female voice
– `"emotion"`: {`"is":` one of angry | happy | sad | neutral | fearful | disgusted | surprised, `"detail": "..."`}
– `"speaking_style"`: {`"is":` one of conversational speech | read speech | public address | shouting | ..., `"detail": "..."`}
– `"language"`: {`"is":` one of western european | south asian | ..., `"detail": "..."`} – detail = the name of the language, one word
– `"notable"`: one short sentence about anything important in the audio that this form does not cover, or `""` if nothing.
Answer with one JSON object only, no other text.

For a single-choice attribute, the expert's label is used when its confidence is above its threshold and the LALM's label otherwise; for a multiple-choice attribute, the labels of both are combined. Answers outside the listed classes, *unsure* and *none* give no label, and the `detail` answers are the candidate leaves.

A.3 LEAF POST-PROCESSING

Candidate leaves become nodes in four steps (Table 4). The merging step uses an external text LLM (Llama-3.3-70B-Instruct), which proposes groups of names but never hears audio; its acoustic check compares the chunks of each proposed group in the embedding space of an encoder suited to the attribute.

Table 4: **Leaf post-processing.**

Step	Rule
Normalize	Remove filler words (`a`, `some`, `sound`) and merge names that are nearly identical and have the same number of words.
Admit	Keep a leaf only if it appears in at least 2 clips from at least 2 sources, adds detail beyond its class, and contains no number.
Merge	The text LLM proposes groups of synonyms. A group is kept only if its names do not conflict in number, negation or polarity (`high-pitched` vs `low-pitched`), its chunks sound closer to each other than to the other leaves of the class, and it does not join two different sources (`goat bleating` vs `sheep bleating`).
Prune	Fold a leaf that only adds a filler word into its sibling (`alarm sound` → `alarm`). Remove attributes whose content lies in the spoken words (topic, intent, scenario, speaker role) or that carried no usable information (recording channel, voice quality, accent, weather, number of sources).

B IMPLEMENTATION DETAILS

This section follows the pipeline: router training, router inference, the reader, compute, and finally the QLoRA baseline.

B.1 ROUTER TRAINING

All encoders are frozen, and their chunk embeddings are computed once and cached. Let y be the labelled class of attribute a on chunk x. For a single-choice attribute, class probabilities are a softmax over the class scores of Eq. 1, $p_a(v \mid x) = \exp s_a(v \mid x) / \sum_{v'} \exp s_a(v' \mid x)$, and the class term of Eq. 2 is $\ell_a^{\text{class}} = -\log p_a(y \mid x)$; for a multiple-choice attribute, $p_a(v \mid x) = \sigma\big(s_a(v \mid x)\big)$ and the class term is the binary cross-entropy against the set of labelled classes. When the label also names a leaf ℓ^* under y, the leaf term compares only the nodes inside y,

$$\ell_a^{\text{leaf}} = -\log \frac{\exp g_a\big(\phi(h)\big)_{\ell^*}}{\sum_{n \in \{y\} \cup \mathcal{L}_{a,y}} \exp g_a\big(\phi(h)\big)_n}, \tag{3}$$

so learning a leaf never moves probability between classes. A leaf becomes a node only with at least 5 training clips (3 for emotion and gender); a rarer leaf trains toward its class node, so the head learns when not to be specific. Teacher labels (e.g. activity from a VGGSound classifier) are summed onto the tree nodes they map to and used as soft targets. The environmental-sound identity heads predict three levels, the identity attribute a (e.g. `animal`), the class v (`dog`) and the leaf ℓ (`Bark`), each trained only where its parent is labelled, and serve $p(\ell \mid x) = p(a \mid x)\, p(v \mid a, x)\, p(\ell \mid v, x)$. Table 5 gives the settings of every head.

Table 5: **Head settings.** All heads use AdamW; checkpoints are selected on held-out training data.

Heads	Architecture	Loss	Learning rate	Batch	Training
speech, music, scene	1–3-layer MLP body (width 256), one head per attribute	masked CE / BCE (Eq. 2)	3×10^{-4}–10^{-3}, one-cycle	64	60–80 epochs
environmental-sound identity	LayerNorm + 1,024-unit layer, sigmoid per node	masked BCE, positive weights	10^{-3}	512	25 epochs
music properties	2-layer MLP (width 256)	class-weighted CE	10^{-3}	64	60 epochs
event timeline	linear layer on PANNs frames	BCE on strong labels	10^{-4}	256	6 epochs
new domain	1-layer MLP body, classes + `none`	CE	10^{-3}	64	40 epochs

Attributes are sampled in proportion to the square root of their number of labels, and part of each batch is class-balanced. Leaf thresholds are fitted for 0.75 precision, and the thresholds of the sound identity heads for F1, on held-out training data.

B.2 ROUTER INFERENCE

- Clips are cut into 3 s chunks with a 1.5 s hop; a chunk reads only the regions its gate opens (probability ≥ 0.5).
- Clip-level attributes average their log-probabilities over the chunks; chunk-level attributes (Table 2) decide per chunk.

- A single-choice attribute serves its best class (emotion: up to three classes with probability ≥ 0.2); multiple-choice music attributes serve every class above 0.3, and each sound identity attribute its top two classes and leaves.
- Attributes with low held-out skill (balanced accuracy < 0.3 above chance) are served only at confidence ≥ 0.5.
- If no chunk hears environmental sound, a question classifier (logistic regression on MiniLM embeddings) decides whether the answer names a sound; if so, the sound identity heads read the whole clip.
- The transcript (Whisper-large-v3, first 180 s) is added when speech or singing is heard, and language is then predicted from its text; the event timeline is added only for questions about the order, duration or number of events.
- In the context, nodes appear in the order they were first reached; time spans and probabilities stay in the traversal trace.
- A new-domain head serves a class only when $p(\texttt{none}) < 0.995$; its `none` class is trained on 2,000 chunks of the training pool.

B.3 Reader

The reader does not generate free text for the answer. With context c, question and options q and option letters $\mathcal{O}$, the prediction is

$$\hat{o} = \arg\max_{o \in \mathcal{O}} \Big[\log p(o \mid c, q) + \beta \log p(o \mid q) \Big], \tag{4}$$

where the second term is the same reader without the context and $\beta = 0.1$; this removes answer-parsing errors and adds a weak prior from the reader's own knowledge. The 26 readers are Qwen2.5 (0.5B–72B), Qwen3 (4B–32B, and 30B-A3B), Llama-2-chat (7B, 13B), Llama-3.1-8B, Llama-3.3-70B, Vicuna v1.5 (7B, 13B), Mistral-7B v0.3, Mistral-Small-24B, Phi-3.5-mini, Phi-4, OLMo-2-13B, Falcon3-10B, SmolLM2-1.7B and gpt-oss-20b. For each reader, one of four prompt templates is selected on 100 development items per benchmark and then locked; gpt-oss, which reasons before answering, generates up to 512 tokens and its answer is parsed.

B.4 Compute

All experiments run on single NVIDIA A100 (40 or 80 GB) or L40S GPUs; the tree annotation and the LLM cleaning use up to four A100-80GB GPUs. A router head trains in a few minutes and a new-domain head in under one minute; a 7B reader reads all three benchmarks in about 5 minutes. The 70B and 72B readers run in 4-bit NF4 quantization with bfloat16 computation, all others in bfloat16. We use PyTorch 2.12, Transformers 4.57, PEFT 0.19, bitsandbytes 0.50, scikit-learn 1.7 and sentence-transformers 5.6, with fixed random seeds.

B.5 QLoRA baseline

The thinker of Omni-7B is loaded in 4-bit NF4 with LoRA adapters of rank 16 ($\alpha = 32$, dropout 0.05) on the attention and MLP projections of the language model (40.4M trainable parameters); the audio encoder is frozen. Training uses AdamW with learning rate 2×10^{-4}, batch size 4 with 4 gradient-accumulation steps, a cosine schedule with 5% warm-up, and 8, 6, 4 and 2 epochs for $k = 5, 10, 20, 50$ (3 epochs at $k = 50$ on marine mammals), on items in the same format as the evaluation items.

C Datasets

Table 6 lists every dataset used to build the tree and train the router, with its role and size.

Table 6: **Datasets** used to generate the tree and to train the router. Tree pool = clips in the tree-generation pool (4,054 in total, labelled by the annotators). Router = additional labelled data used to train heads; "–" means the dataset reaches the router only through the pool's labels (the speech heads, including emotion, are trained on the pool alone). AudioSet test (9,042 clips) and FSD50K evaluation (10,161 clips) are held out. No benchmark audio is used.

Dataset	Domain	Tree pool	Router training (labels, size)
WavCaps AudioSet-SL	sound	805	sound hierarchy: AudioSet labels, 81,325 clips
Clotho	sound	645	–
FSD50K	sound	413	sound hierarchy (CLAP heads): leaf labels, 24,238 dev clips
ESC-50	sound	280	sound hierarchy (CLAP heads): class labels, 1,793 clips
WavCaps SoundBible	sound	127	–
AudioCaps	sound	107	–
Freesound	sound	82	–
YouTube (AF-Think clips)	mixed	82	–
AF-Think question–answer records	text	–	question-path classifier: 83,249 records
AudioSet strong labels	sound	–	event timeline: frame labels, 81,325 clips (the same split)
VGGSound (ECAPA classifier)	sound	–	activity: soft targets (teacher, training only)
MusicCaps	music	618	music heads: descriptors
IRMAS	music	80	music heads: instrument labels
FMA	music	82	–
GTZAN	music	80	–
Instrument samples	music	22	–
MusicCaps + IRMAS for the music heads: 8,474 chunks			
FLEURS	speech	204	language classifier: transcripts, 49 languages
RAVDESS	speech	165	–
IEMOCAP	speech	152	–
LibriSpeech	speech	30	–
TAU Urban Acoustic Scenes	scenes	80	–
BirdSet POW	birds	–	domain head: k clips per species
Watkins (BEANS)	marine	–	domain head: k clips per species

D BENCHMARKS AND EVALUATION PROTOCOL

This section details the benchmarks summarised in §3, how items are scored, and the baseline settings.

MMAU. Test-mini split, 1,000 items, about one third per domain, 27 skills that require expert knowledge.

MMAR. 1,000 items in four reasoning layers (signal, perception, semantic, cultural) and 16 subcategories. The answer is given as option text, which we match to the option letter (Appendix D); three items whose answer names two options are removed.

SAKURA. Four attributes (gender, language, emotion, animal), each with 500 single-hop and 500 multi-hop items. Single-hop items ask for the attribute; multi-hop items require the attribute and a fact about it.

MMAU-Pro. Extends MMAU with long recordings, multi-audio questions, spatial audio, open-ended answers and 49 skills. We keep the single-audio multiple-choice items (4,133, with 2 to 11 options); multi-audio, open-ended and instruction-following items are not supported by our pipeline. For recordings longer than 3 minutes, the router covers the first 60 s and the transcript the first 180 s.

The gold letters are strongly unbalanced (97% of the spatial-audio answers are option (a)), so we shuffle the options of every item with a fixed per-item permutation, the same for all systems.

Scored items. Three items per benchmark serve as in-context examples when a reader's prompt template is selected and are excluded from scoring, which leaves 997 MMAU, 994 MMAR and 3,997 SAKURA items.

New domains. Birds: 938 items over 48 species from the Powdermill (POW) subset of BirdSet, with recordists disjoint from the training clips. Marine mammals: 678 items over 31 species from the Watkins collection, through BEANS. Each item has four options, and distractors are taxonomically close species (same genus first), so the option list alone does not reveal the answer. Training clips come from the same collections, excluding the evaluation recordings and any clip above 0.95 CLAP similarity to an evaluation clip.

Baselines. Omni-7B is run with the same questions, options and scoring as our system, and hears the first 180 s of each MMAU-Pro recording. AF3 is trained with the AF-Think chain-of-thought data. In the in-context baseline, $S \in \{1, 3, 5\}$ labelled clips of each of the item's four option species are given in the prompt; QLoRA uses the same k clips per species with the same containment filter.

Scoring. MMAR's `answer` field is the option text. We map it to an option letter by matching the text to the options (995 items exact, 2 by unambiguous word overlap). Three items whose answer names two options are removed. All systems, including Omni-7B, are scored with the same mapping. On MMAU the gold letter agrees with the option text on all 1,000 items, but 16 items list the correct answer text under two letters and score only one of them; counting either letter changes Omni-7B by $+0.6$ and ours by $+0.3$ points. We report the official scoring.

E WHICH ATTRIBUTES THE READER USES

To decide which attributes to serve, we used the annotator's own labels on the 300 development clips per benchmark as a perfect router and removed one attribute group at a time from tree + transcript (Table 7, Qwen2.5-7B, before the pruning of Appendix A). Identity attributes (gender, sound source, emotion and the leaves) carry SAKURA, and music attributes carry MMAU; no group moves MMAR by more than 1.3 points (4 items). Removing the speech content attributes (topic, scenario, speaker role, intent) does not lower SAKURA or MMAU and costs 1.3 points on MMAR; we prune them because their content is already in the transcript.

Table 7: **Attribute groups.** Change in accuracy (points) when one attribute group is removed from a perfect-router tree, on the development clips (SAKURA 600 items, MMAU 300, MMAR 298; one item is 0.2, 0.3 and 0.3 points). Negative values mean the group helps.

removed group	SAKURA	MMAU	MMAR
every leaf (classes kept)	-8.0	$+0.3$	-1.3
speech gender	-7.7	-1.0	$+0.3$
sound source (event, action and material)	-7.2	-1.3	-0.3
speech emotion	-7.0	$+1.3$	-0.7
speech language	-0.8	$+1.0$	$+0.7$
music identity (genre, instrument, mood, era, vocals)	0.0	-2.3	$+0.3$
music structure (tempo, meter, mode, texture, melody, production, ensemble)	0.0	-1.3	0.0
speech voice (age, style, channel, voice quality, accent, number of speakers)	-0.2	0.0	0.0
sound context (scene, activity, weather)	-0.2	$+1.7$	-0.7
sound proximity and number of sources	0.0	$+0.3$	0.0
speech content (topic, scenario, speaker role, intent)	$+0.5$	$+0.7$	-1.3

F CHOICE OF READER

Table 8 gives the accuracy of all 26 frozen readers on the same cached contexts; Figure 2 plots them against reader size.

Table 8: **Choice of reader.** Accuracy (%) of every frozen reader on the same cached tree + transcript contexts, full sets. B = parameters in billions (Qwen3-30B-A3B has 3B active). Prompt template locked per reader on 100 development items per benchmark. Llama-2 and Vicuna are scored with an answer prefix; gpt-oss by generation. Audio LLMs: B is total size from the released weights.

Reader	B	MMAU↑	MMAR↑	SAKURA↑
Qwen2.5-0.5B	0.5	40.5	36.5	47.6
Qwen2.5-1.5B	1.5	61.5	47.3	69.3
SmolLM2-1.7B	1.7	43.2	39.1	44.9
Qwen2.5-3B	3	63.1	50.1	72.0
Phi-3.5-mini	3.8	63.0	47.4	73.8
Qwen3-4B	4	63.9	50.9	78.2
Llama-2-7B-chat	6.7	43.8	39.2	58.6
Vicuna-7B v1.5	6.7	42.5	37.6	64.3
Mistral-7B v0.3	7.2	61.8	46.2	75.5
Qwen2.5-7B	7.6	68.3	51.6	80.1
Llama-3.1-8B	8	66.8	50.1	79.5
Qwen3-8B	8.2	64.5	51.7	78.5
Falcon3-10B	10.3	65.9	50.6	79.6
Llama-2-13B-chat	13	47.9	43.6	65.5
Vicuna-13B v1.5	13	57.7	41.6	69.8
OLMo-2-13B	13.7	61.0	48.3	75.3
Qwen2.5-14B	14.7	69.4	55.6	83.3
Phi-4	14.7	**71.7**	54.9	83.5
Qwen3-14B	14.8	69.2	54.0	81.7
gpt-oss-20b	21	68.6	55.3	**86.0**
Mistral-Small-24B	24	66.0	53.2	84.5
Qwen3-30B-A3B	30.5	67.8	55.4	84.9
Qwen2.5-32B	32.8	71.4	55.7	85.1
Qwen3-32B	32.8	69.1	52.5	83.3
Llama-3.3-70B (4-bit)	70.6	67.5	56.0	84.5
Qwen2.5-72B (4-bit)	72.7	70.5	**56.6**	85.3
Qwen2.5-Omni-7B (hears)	10.7	75.1	63.6	73.3
Audio Flamingo 3 (hears)	8.3	75.5	57.8	68.2
Qwen3-Omni-30B-A3B (hears)	35.3	74.0	62.0	73.0

G NEW DOMAINS: EXACT NUMBERS

Table 9 gives the accuracy behind Figure 3 for every number of labelled clips per species.

Table 9: **Few-shot domain adaptation**, 4-way species identification, accuracy (%) by labelled clips per species (k). Ours = one new head on a frozen encoder, frozen Qwen2.5-7B reader, tree only. Omni-7B zero-shot: birds 32.3, marine 25.8. Marine has a median of 33 labelled clips per species, so k=50 is not reachable for our head.

	method (accuracy↑)	k=1	3	5	10	20	50
Birds (48 sp.)	**L2R (ours, 0.34M trainable)**	**48.4**	**72.7**	**79.6**	**85.3**	**88.4**	**91.1**
	Omni-7B + QLoRA (40.4M trainable)	–	–	54.1	61.0	69.0	75.2
	Omni-7B, clips in context	43.3	46.1	47.3	–	–	–
Marine (31 sp.)	**L2R (ours, 0.17M trainable)**	41.3	**59.6**	**65.2**	**76.5**	**83.0**	–
	Omni-7B + QLoRA (40.4M trainable)	–	–	52.2	68.6	73.5	83.6
	Omni-7B, clips in context	**44.1**	57.1	60.5	–	–	–

The encoder matters, the reader does not. With the k=5 head, 24 readers from 1.5B to 72B all score 78–82 on birds and 64–68 on marine mammals. The domain node carries the answer, so the reader only needs to copy it. The choice of frozen encoder, in contrast, decides the result. We use BirdNET for birds because it is trained on bird vocalisations, while CLAP is trained on general

audio; for marine mammals we use CLAP, which the router already runs, so this domain adds no new encoder. A modular design lets each domain use the encoder that hears it best.

Forgetting. Adding both domain heads leaves the other heads unchanged, but the new nodes can be served on clips of the original benchmarks. With both heads added, the Qwen2.5-7B reader scores 79.6 on SAKURA (80.1 without), 67.7 on MMAU (68.3) and 51.8 on MMAR (51.6). A species node is served on 17–21% of the items. All 19 SAKURA items it breaks are animal clips that now also carry a bird species (a cat served as *black-and-white warbler*). A tighter abstention rule for the domain heads is the fix.

H WHERE LISTEN-TO-REASON LAGS BEHIND AUDIO LLMS

H.1 BY QUESTION TYPE

Table 10 breaks the gap to Omni-7B down by question type, assigned from the question wording.

Table 10: **Where LISTEN-TO-REASON lacks**, by question type (assigned from the question wording), against Omni-7B and the blind reader, scored against gold. Net = items we get right minus items Omni gets right. Qwen2.5-7B reader.

	MMAU					MMAR				
question type	n	ours↑	Omni↑	blind↑	net↑	n	ours↑	Omni↑	blind↑	net↑
counting	100	47.0	53.0	43.0	-6	122	33.6	45.1	27.0	-14
order / timing	113	60.2	62.8	39.8	-3	106	47.2	57.5	37.7	-11
scene / place / activity	115	77.4	93.0	46.1	-18	111	58.6	71.2	27.0	-14
identify source / instrument / genre	148	71.6	83.8	60.1	-18	30	43.3	46.7	40.0	-1
music theory	102	57.8	61.8	52.9	-4	22	50.0	31.8	40.9	+4
speaker / voice	202	74.8	72.3	48.0	+5	74	56.8	77.0	31.1	-15
speech content	37	89.2	100.0	43.2	-4	82	58.5	65.9	46.3	-6
other	180	71.1	81.1	52.2	-18	447	54.4	68.0	40.0	-61

The losses fall into three groups.

No node carries the answer: counting. On counting questions LISTEN-TO-REASON stays close to the blind reader (MMAU 47.0 vs 43.0; MMAR 33.6 vs 27.0), a smaller gain over blind than on any other type. The router describes 3 s chunks and does not represent individual event occurrences, so no node states a count. Omni-7B is also weak here (53.0 and 45.1), which suggests that counting is hard for current audio encoders in general. Order and timing questions are closer: LISTEN-TO-REASON is within three points of Omni-7B on MMAU (60.2 vs 62.8), though 11 items behind on MMAR.

A node exists but is too coarse: identity and scene. The largest MMAU losses are in identifying a source, instrument or genre (-18 items) and in describing a scene or activity (-18). The answer to these questions is a specific name (*waterfall*, not *water*) or a description of the whole clip. In the errors we inspected, the router often reaches the right class but not the right leaf, or serves several candidate leaves. Figure 5 shows one such case.

The answer is in the transcript but needs inference: MMAR. On MMAR the largest net loss is in questions outside the listed types (-61 of 447 items). By sub-category, content analysis alone accounts for -16 items. Speaker and voice questions show the same pattern: ahead of Omni-7B on MMAU ($+5$) and 15 items behind on MMAR. These are the questions where a larger reader helps most (§4.2). LISTEN-TO-REASON is also ahead of Omni-7B on MMAR music-theory questions ($+4$ items). Because every loss is visible in the context, each group points to a specific repair: event-level nodes for counting, finer leaves and scene-level attributes for identity and scenes, and a stronger reader for inference over speech.

H.2 BY BENCHMARK SUB-CATEGORY

Tables 11 and 12 compare LISTEN-TO-REASON (Qwen2.5-7B reader) with Omni-7B on every benchmark sub-category. The MMAU losses are in acoustic-scene reasoning (-12 items), musical texture (-8), event-based sound reasoning (-7), acoustic source inference, temporal event reasoning and melodic structure (-6 each). The MMAR losses are in environmental perception (-23), content analysis (-16), counting and statistics (-15), correlation analysis (-12) and professional knowledge (-11). We are ahead of Omni-7B on MMAU emotion-flip detection ($+5$), music temporal reasoning ($+4$), emotion-state summaries ($+2$) and conversational fact retrieval ($+1$), and level on harmony and chord progressions.

Table 11: **MMAU by sub-category**, Qwen2.5-7B reader. Net = items right for us minus items right for Omni-7B.

sub-category	n	ours↑	Omni↑	blind↑	net↑
Acoustic Scene Reasoning	48	56.2	81.2	33.3	-12
Musical Texture Interpretation	34	58.8	82.4	44.1	-8
Event-Based Sound Reasoning	48	64.6	79.2	62.5	-7
Acoustic Source Inference	48	75.0	87.5	72.9	-6
Temporal Event Reasoning	48	58.3	70.8	25.0	-6
Melodic Structure Interpretation	33	48.5	66.7	42.4	-6
Ambient Sound Interpretation	47	76.6	89.4	59.6	-6
Instrumentation	35	71.4	85.7	48.6	-5
Emotional Tone Interpretation	33	72.7	87.9	45.5	-5
Phonological Sequence Decoding	48	83.3	89.6	62.5	-3
Eco-Acoustic Knowledge	46	80.4	87.0	63.0	-3
Sound-Based Event Recognition	46	84.8	91.3	45.7	-3
Counting	29	48.3	55.2	41.4	-2
Musical Genre Reasoning	34	76.5	82.4	58.8	-2
Event-Based Knowledge Retrieval	33	72.7	75.8	60.6	-1
Phonemic Stress Pattern Analysis	53	54.7	56.6	34.0	-1
Socio-cultural Interpretation	20	75.0	80.0	60.0	-1
Rhythm and Tempo Understanding	46	54.3	56.5	56.5	-1
Dissonant Emotion Interpretation	35	91.4	91.4	82.9	+0
Key highlight Extraction	21	90.5	90.5	61.9	+0
Multi Speaker Role Mapping	27	100.0	100.0	14.8	+0
Lyrical Reasoning	10	90.0	90.0	60.0	+0
Harmony and Chord Progressions	33	63.6	63.6	54.5	+0
Conversational Fact Retrieval	22	95.5	90.9	40.9	+1
Emotion State summarisation	44	59.1	54.5	31.8	+2
Temporal Reasoning	56	41.1	33.9	41.1	+4
Emotion Flip Detection	20	55.0	30.0	25.0	+5

Table 12: **MMAR by sub-category**, Qwen2.5-7B reader. Net = items right for us minus items right for Omni-7B.

sub-category	n	ours↑	Omni↑	blind↑	net↑
Environmental Perception and Reasoning	149	57.7	73.2	30.9	-23
Content Analysis	301	62.1	67.4	44.9	-16
Counting and Statistics	99	31.3	46.5	25.3	-15
Correlation Analysis	50	44.0	68.0	28.0	-12
Professional Knowledge and Reasoning	70	47.1	62.9	34.3	-11
Speaker Analysis	48	52.1	68.8	29.2	-8
Acoustic Quality Analysis	18	22.2	61.1	22.2	-7
Culture of Speaker	50	58.0	68.0	42.0	-5
Music Theory	63	42.9	50.8	42.9	-5
Emotion and Intention	60	55.0	61.7	30.0	-4
Aesthetic Evaluation	8	37.5	75.0	12.5	-3
Anomaly Detection	17	58.8	76.5	52.9	-3
Spatial Analysis	15	40.0	53.3	46.7	-2
Temporal Analysis	28	25.0	32.1	35.7	-2
Audio Difference Analysis	8	50.0	62.5	62.5	-1
Imagination	10	60.0	70.0	40.0	-1

I AUDIO LLMS AS JUDGES

Protocol. Items are sampled from clips outside the development set, with a fixed seed: 400 from MMAU (stratified by group and sub-category), 400 from MMAR (proportional to group) and 100

multi-hop items per SAKURA track. The judge hears the first 60 s of the clip and reads the served nodes as the reader receives them, without the transcript and without the question. It answers with one digit, and the score is the expected value of its probabilities over the digits 1–5, so no output needs parsing. The swap control uses the nodes of another sampled clip of the same benchmark from a different category (disjoint modalities, e.g. music for speech; Animal for the SAKURA speech tracks), chosen to share as few nodes as possible with the real ones (mean node overlap below 1%). Confidence intervals are 95% bootstrap intervals over items (2,000 resamples).

Prompt. The system message and the user message are given below; `{nodes}` is the node list of the item (or of the swapped clip), and the audio is attached to the user message.

System: You are an expert audio analyst auditing the output of an automatic audio-analysis system. Listen to the whole recording before answering, and base every judgement on what you hear.

User: An automatic system analysed this recording and matched it to nodes of an audio knowledge graph, written as "attribute: value" pairs; a value in parentheses is a more specific label. The system can be wrong.

Nodes:
{nodes}

Are these nodes related to what is actually audible in the recording?
5 = every node matches something audible in the recording
4 = nearly all nodes match; one is doubtful or too vague
3 = the main content is captured, but several nodes do not match
2 = most nodes do not match; the main source, speaker or music is wrong
1 = the nodes describe a different recording

Answer with a single digit from 1 to 5 and nothing else.

All three judges rank the served nodes above the swapped ones on most items (Figure 4). They differ in scale: Audio Flamingo 3 rates even the swapped nodes above 3, and Kimi-Audio rates the served nodes lowest, most of all on the SAKURA speech tracks (language and gender). Averaging the three judges after standardising each one's scores gives the same ordering on every benchmark.

J EXAMPLE CONTEXTS

Figure 5 shows three MMAU items with the complete context given to the reader. More examples, with the router's path through the tree chunk by chunk, are on the project page.

Correct: *What mechanical issue is likely occurring with the vehicle?*

Options: (a) Tire puncture (b) Engine misfire **(c) Engine knocking** (d) Brake failure Prediction: **(c)**

`sound scene: construction site; sound machine: engine (Engine), **engine (Engine knocking)**, mechanisms (Mechanisms); sound action material: impact (keys hitting), impact; sound proximity: close to the microphone.`

Correct: *Given the sounds, what type of weather event can be inferred?*

Options: **(a) Thunderstorm** (b) Clear skies (c) Heatwave (d) Snowstorm Prediction: **(a)**

`sound scene: park; sound nature: water (Rain), water (Water), **thunderstorm (Thunder)**, thunderstorm (Thunderstorm); sound action material: liquid flow (water rushing), air flow; sound proximity: close to the microphone.`

Wrong, and traceable: *Identify the source of the waterfall sound.*

Options: **(a) Waterfall** (b) Rain (c) Ocean waves (d) River Prediction: (b)

`sound scene: park; sound impact: liquid (Liquid), liquid (Boiling); sound household: frying food (Frying (food)), toilet flush; sound machine: mechanisms (Tick); sound nature: **water (Rain)**, water (Water); sound action material: liquid flow (water rushing).`

The router named `Rain` and no waterfall node; the reader followed it. The error is located in one node of the `water` class, which is where a fix (a `Waterfall` leaf with training clips) belongs.

Figure 5: **Interpretable by design.** Real MMAU items with the complete tree context given to the frozen Qwen2.5-7B (no transcript is given for these clips: no chunk heard speech or singing). Bold marks the node that decides the answer.

K COST COMPARISON

Table 13 gives the cost of LISTEN-TO-REASON and the audio LLMs discussed in §4.5. Qwen3-Omni needs the fewest MACs despite its 35.3B parameters, because it is a mixture-of-experts model that runs about 3B parameters per token, but its per-item time is still the highest (0.93 s).

Table 13: **Cost** of LISTEN-TO-REASON (Qwen2.5-7B reader) and audio LLMs. MACs and seconds are per MMAU item on one A100 GPU; MACs count executed operations, so a mixture-of-experts model counts only its active parameters.

	total params↓	trained params↓	audio training data↓	MACs / item↓	s / item↓
L2R (ours)	10.50B	7.7M	≈0.2M clips	5.3T	0.57
Qwen2.5-Omni-7B	10.73B	10.73B	300B tokens	2.7T	0.12
Audio Flamingo 3	8.27B	8.27B	9.6M QA + 13.2M captions	3.3T	0.11
Qwen3-Omni-30B-A3B	35.3B (3B active)	35.3B	20M h + 0.77T tokens	0.76T	0.93