FROM DECODABILITY TO UTILITY: SPEECH DEEPFAKE DETECTION IN AUDIO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio language models have emerged as general-purpose speech interfaces, yet how their speech deepfake detection capability should be characterized remains unclear. Existing evaluations often rely on a single model endpoint, obscuring whether observed behavior reflects information supported by the underlying audio-language system or its effective use. We present Representation Accessibility Versus Exploitation by the Language Pathway (RAVEL), a controlled evaluation framework that separates native behavior, supervised decodability of frozen representations, and downstream exploitation through task adaptation, together with encoder-depth analysis. We evaluate thirteen Audio-LLMs across five datasets, with supervised comparisons on ten open-weight models and source-protocol and depth analyses. Native outputs are benchmark-dependent and often fail in decision validity or correctness, whereas frozen audio- and language-side representations frequently support strong supervised detection. This coexistence shows that weak native behavior need not imply the absence of accessible authenticity information. Native adaptation consistently improves in-domain detection over language-side frozen readout. Depth-wise adaptation further shows that intermediate representations can outperform final blocks in cross-domain discrimination, while encoder regions favored by frozen readouts can reverse under adaptation and shift across target domains. Under 19LA→AT-ADD transfer, Native-SFT outperforms both frozen readouts in EER for 9/10 models but in Macro-F1 for only 1/10, revealing metric-dependent transfer gains. Overall, supervised decodability and adaptation utility are distinct: stronger frozen-readout performance does not determine which representations are most effective for encoder-to-language adaptation.

1 INTRODUCTION

Recent advances in speech synthesis have increased the need for reliable speech deepfake detection (SDD) Wang et al. (2024); Li et al. (2025a). Audio language models (Audio-LLMs) provide an instruction-based interface for this task, processing speech inputs and responding to natural-language queries Zhang et al. (2023); Tang et al. (2024); Chu et al. (2024). However, their detection capability and cross-domain generalization remain insufficiently understood. An Audio-LLM can be queried directly, used as a frozen feature extractor, or adapted to predict authenticity through its native language pathway. These settings motivate a more fundamental question: *How does native detection relate to supervised decodability from frozen representations, and do representations preferred by frozen readouts also provide effective inputs for encoder-to-language adaptation?*

Task-specific optimization has advanced Audio-LLM-based SDD through supervised adaptation, audio representation refinement, additional time–frequency inputs, and reasoning-oriented supervision Gu et al. (2025); Li et al. (2025b); Guo et al. (2026b); Xie et al. (2026b); Xu et al. (2026); Dvirniak et al. (2026); Chen et al. (2026). However, these approaches do not reveal how much detection information is already accessible from frozen representations. It therefore remains unclear how frozen audio-side and language-side readouts relate to native task adaptation, whether adaptation gains in score discrimination translate into reliable fixed-rule decisions under domain shift, and whether representations that are easily decoded when frozen are also those most effectively exploited after adaptation.

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.

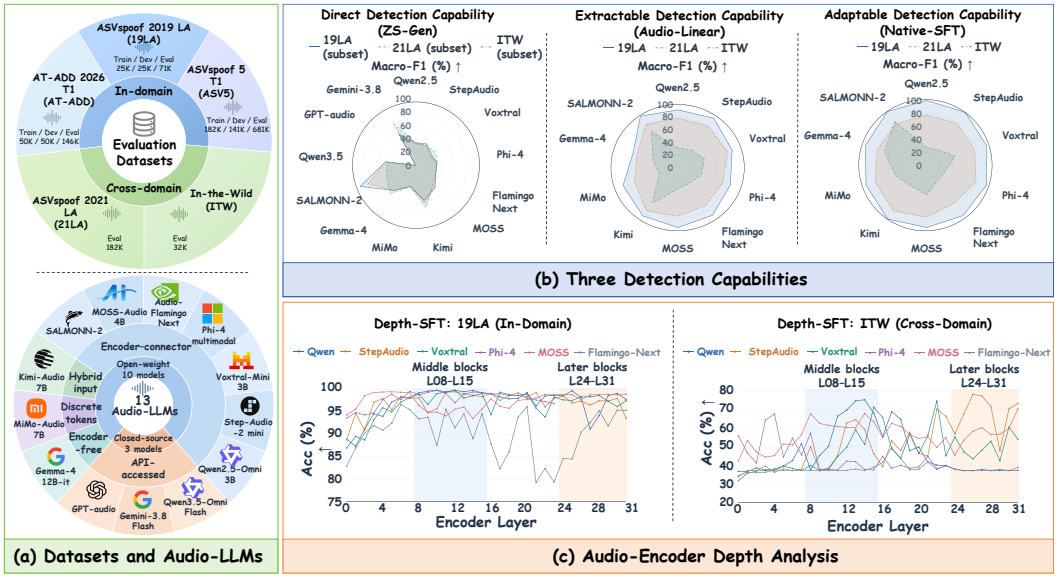


Figure 1: Overview of RAVEL and the evaluation scope. (a) Five benchmarks and thirteen Audio-LLMs. (b) Three complementary settings: direct detection from native outputs, extractable detection from frozen representations, and adaptable detection after task-specific adaptation. (c) Depth-SFT accuracy across encoder depths for six Audio-LLMs on in-domain 19LA and cross-domain ITW.

To investigate these questions, we introduce Representation Accessibility Versus Exploitation by the Language Pathway (RAVEL), a controlled evaluation framework with three complementary settings that separate native behavior, representation accessibility, and downstream exploitation. Direct detection evaluates zero-shot generation (ZS-Gen) and answer-token scoring (ZS-Score); extractable detection measures supervised decodability from frozen audio- and language-side representations; and adaptable detection uses native supervised fine-tuning (Native-SFT) to adapt the language pathway for authenticity prediction.

Beyond these fixed-interface settings, intermediate or multi-layer audio-encoder representations have been shown to improve Audio-LLM audio understanding and paralinguistic modeling through layer probing or feature fusion Pang et al. (2026); Yang et al. (2026b). Motivated by these findings, we introduce depth-wise supervised fine-tuning (Depth-SFT), which adapts selected encoder representations into the language pathway, and compare it with frozen readouts across depths and target domains. Figure 1 summarizes the evaluation design. We evaluate ten open-weight and three closed-source Audio-LLMs on three source benchmarks and two held-out test sets. Supervised comparisons primarily use ASVspoof 2019 LA (19LA) Wang et al. (2020), complemented by AT-ADD Xie et al. (2026a) and ASVspoof 5 (ASV5) Wang et al. (2024); encoder-depth analyses use 19LA-source supervision on representative models. Our main contributions are summarized as follows:

- To characterize Audio-LLM deepfake detection beyond a single endpoint, we introduce **RAVEL**, an evaluation framework separating **direct**, **extractable**, and **adaptable** settings, enabling comparison of native behavior, frozen decodability, and task adaptation.
- To characterize relationships between native behavior, supervised decodability, and adaptation, we compare native outputs, frozen readouts, and Native-SFT. Native-SFT improves in-domain detection over language-side frozen readout, yet under 19LA→AT-ADD transfer it beats both frozen readouts in EER for 9/10 models but in Macro-F1 for only 1/10, exposing a gap between discrimination and fixed-rule transfer.
- To study whether encoder-depth preferences persist from decodability to adaptation, we introduce **Depth-SFT**. Intermediate representations can outperform final-block inputs in cross-domain discrimination, while regions favored by frozen readouts need not remain favorable under adaptation and can shift across target domains.

2 Related Work

Audio-LLMs for speech deepfake detection. Existing work reports weak direct prompting but substantial gains from supervised adaptation, encoder–backbone optimization, and additional time–frequency inputs Gu et al. (2025); Chuchra et al. (2025); Li et al. (2025b); Guo et al. (2026b). Recent methods further introduce artifact-grounded reasoning, attack characterization, or acoustic descriptions Xie et al. (2026b); Xu et al. (2026); Dvirniak et al. (2026); Chen et al. (2026). These studies establish task adaptability, but leave the relationship between native behavior, frozen decodability, and post-adaptation transfer less explored.

Native outputs and internal representations. Language-model studies show that internal states can contain information recoverable by supervised probes before it is expressed in model outputs Hewitt & Liang (2019); Burns et al. (2022); Orgad et al. (2025); Belrose et al. (2023). In Audio-LLMs, Fan et al. (2026) trace audio-derived concepts through the language backbone, while Pang et al. (2026) show that paralinguistic cues can remain probe-accessible despite weak native use and improve their use through multi-layer mixing and preference optimization. Our focus is complementary: we test whether encoder regions favored by frozen linear and nonlinear readouts remain favorable under independent single-layer adaptation through the native language pathway, particularly under domain shift.

Encoder depth and cross-domain detection. Layer-aware speech deepfake detectors exploit intermediate representations through layer weighting, fusion, or selection Zhang et al. (2024); Guo et al. (2024; 2026a); Huang et al. (2026). Prior layer-wise analyses further show that discriminative depths vary across encoders and domains, with intermediate layers often outperforming final outputs El Kheir et al. (2025); Serrano et al. (2025); Beheshti et al. (2026). Rather than seeking a better frozen representation alone, we compare linear and nonlinear frozen readouts with Depth-SFT across encoder depths and target domains to test whether regional preferences transfer from decodability to adaptation.

3 RAVEL: A Controlled Evaluation Framework

As illustrated in Figure 2, the RAVEL framework characterizes Audio-LLM speech deepfake detection through three complementary settings that separate native behavior, representation accessibility, and downstream exploitation. *Direct* detection evaluates released models without task-specific training. *Extractable* detection measures representation accessibility through supervised decodability from frozen representations. *Adaptable* detection evaluates downstream exploitation through task-specific adaptation of the native language pathway. We further extend the comparison across audio-encoder depth to examine whether representations favored by frozen readouts remain effective for adaptation across target domains.

3.1 Model Interfaces and Detection Settings

Given a speech utterance x and authenticity label $y \in \{0, 1\}$, supervised components are trained on the labeled source split $\mathcal{D}_s^{\mathrm{tr}}$ and applied unchanged to held-out source and target evaluation data.

We consider three representation interfaces. Audio-Linear operates on a model-specific audio-side sequence $h_{\mathrm{aud}} = P(A(x))$, where $A(\cdot)$ denotes the audio encoder and $P(\cdot)$ comprises native transformations up to the selected audio-side interface. LM-Linear uses $h_{\mathrm{lm}} = H_{\mathrm{LM}}(x, p)_{N-1}$, the final language-side vector supplied to the output head at the first answer position for instruction p. For depth-wise analysis, $h_l = A_l(x)$ denotes the raw output of encoder block l. Since the audio-side interface h_{aud} may involve additional model-specific processing, raw encoder outputs are analyzed separately from Audio-Linear features.

Direct detection. ZS-Gen invokes each model's native autoregressive interface and evaluates the resulting authenticity response. ZS-Score instead reads the native `real` and `fake` logits at the first answer position without task-specific fitting:

$$s_{\mathrm{nat}}(x, p) = \big[\,\mathrm{LMHead}(h_{\mathrm{lm}})\big]_r - \big[\,\mathrm{LMHead}(h_{\mathrm{lm}})\big]_f, \tag{1}$$

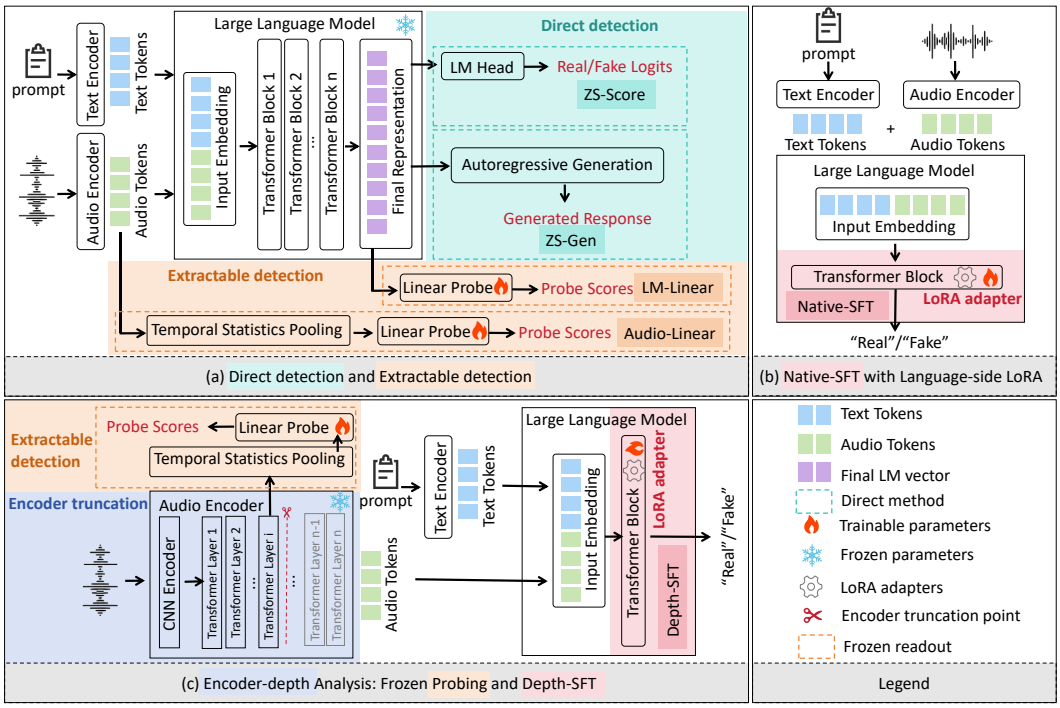


Figure 2: Overview of RAVEL, a controlled evaluation framework for Audio-LLM speech deepfake detection. Direct detection evaluates native responses and answer-token scores; extractable detection probes frozen audio-side and language-side representations; adaptable detection performs task adaptation. Encoder-depth analysis compares frozen readouts and Depth-SFT over encoder regions.

where larger values indicate a stronger preference for bona fide speech. Generated responses are adjudicated as `real`, `fake`, or refusal; detailed prompting and adjudication procedures are provided in Appendices D.1 and D.2.

Extractable detection. Audio-Linear and LM-Linear fit independent supervised linear readouts to h_{aud} and h_{lm}, respectively, while keeping the Audio-LLM frozen. Audio-side sequences are summarized by temporal mean–standard-deviation pooling, whereas LM-Linear directly uses the single language-side vector. Each readout is trained only on source-domain labels and transferred unchanged to target datasets. This setting measures supervised linear decodability rather than native zero-shot behavior.

Adaptable detection. Adaptable detection uses Native-SFT to adapt the native language pathway with single-token authenticity supervision while keeping the audio pathway and output head frozen. Evaluation uses the adapted answer-token scores without target-domain fitting. The exact optimization objective, trainable modules, and model-specific exceptions are described in Appendix C.3.

3.2 Depth-Wise Analysis of Decodability and Adaptation Utility

The fixed-interface comparisons above do not establish whether encoder representations that support stronger frozen readout are also preferable for task adaptation. We therefore extend the analysis across encoder depth and evaluate each encoder block independently through three branches.

Frozen readouts. Layer-wise linear probing applies the same supervised readout procedure to each raw encoder output h_l. We additionally train layer-specific AASIST Jung et al. (2022) detectors on the frozen hidden sequences, preserving temporal structure and providing a complementary nonlinear readout without the language backbone.

Depth-SFT. For each depth l, Depth-SFT truncates the encoder after block l and supplies h_l to the native language pathway through the model-specific terminal operations and audio-to-language connection. The selected connection and language-side adapters are trained with the same authenticity-

answer supervision, while the encoder blocks and native output head remain frozen. Each depth is trained and evaluated independently.

Depth-SFT differs from Native-SFT because it additionally adapts the connection from the selected encoder representation to the language backbone. Its final-block configuration is therefore used as the within-branch reference rather than being treated as equivalent to Native-SFT. Likewise, raw encoder-block probes are distinct from Audio-Linear, whose interface may include additional native transformations. Complete training and optimization details are provided in Appendix C.4, with full encoder-depth results reported in Appendix F.

4 Experimental Setup

Models and benchmarks. We evaluate ten open-weight Audio-LLMs and three closed-source models. Open-weight models participate in ZS-Gen, ZS-Score, and 19LA-source comparisons of Audio-Linear, LM-Linear, and Native-SFT. The closed-source models, Qwen3.5-Omni-Flash Team (2026), GPT-audio OpenAI (2025), and Gemini-3.8-Flash Google DeepMind (2026), are evaluated with ZS-Gen only. Qwen2.5-Omni-3B (Qwen) Xu et al. (2025), Kimi-Audio-7B (Kimi) Ding et al. (2025), and MOSS-Audio-4B (MOSS) Yang et al. (2026a) additionally support AT-ADD and ASV5 source protocols. Depth-wise analyses cover six representative models. SALMONN-2 is marked for its documented spoof-oriented training exposure. Complete model details are provided in Appendix B.2. We use three source benchmarks: 19LA, AT-ADD, and ASV5, together with two held-out test sets: ASVspoof 2021 LA (21LA) Liu et al. (2022) and In-the-Wild (ITW) Müller et al. (2022). The main supervised protocol uses 19LA as the source domain, while additional AT-ADD and ASV5 source evaluations are conducted for Qwen, Kimi, and MOSS. Dataset statistics and protocol details are provided in Appendix A.1.

Implementation and metrics. ZS-Gen, ZS-Score, LM-Linear, and Native-SFT retain native audio pipelines, whereas audio-side linear readouts standardize inputs to 64,600 samples at 16 kHz. For ZS-Gen, a blind Qwen3-Omni-30B-A3B-Thinking judge maps generated responses to `real`, `fake`, or `refuse`. Linear probes train logistic regression on frozen pooled audio features or the final language-side vector. Native-SFT applies language-side LoRA Hu et al. (2021) while freezing the audio pathway; Depth-SFT additionally adapts the specified encoder-to-language connection. We report EER (continuous scores), alongside Macro-F1 and accuracy (fixed real–fake threshold). ZS-Gen adds Valid Rate, treating non-decisions as abstentions. Full implementation details, prompting, and metric definitions are deferred to Appendices C and D.1–D.2.

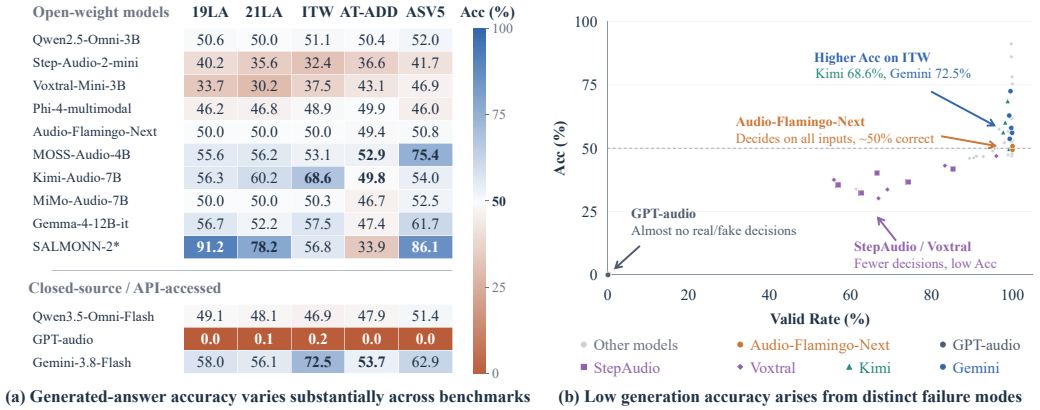
5 Results and Analysis

We examine how native behavior, supervised decodability, and adaptation utility relate to one another. We first characterize native detection, then evaluate supervised readouts of frozen representations, assess language-side adaptation under the same source protocols, and finally examine how encoder depth changes frozen-readout and adaptation preferences. Unless otherwise stated, supervised methods are trained under the 19LA-source protocol and evaluated unchanged across all five evaluation sets, with 19LA, 21LA, and ITW forming the main in-domain and transfer comparisons. Additional AT-ADD-source and ASV5-source experiments cover Qwen, Kimi, and MOSS, while all depth analyses use 19LA-source supervision.

5.1 RQ1: How effective are Audio-LLMs at native deepfake detection?

We evaluate released models through ZS-Gen and ZS-Score without additional task-specific training. Figure 3 reports generated-answer results for ten open-weight and three closed-source models on 1,000 class-balanced utterances per dataset under common semantic adjudication. Table 1 reports native-score results for the ten open-weight models on the full evaluation sets of the same five datasets.

Native generated-answer performance is strongly model- and benchmark-dependent. Gemini-3.8-Flash reaches 72.5% Acc on ITW, while MOSS achieves 75.4% on ASV5 but only 52.9% on AT-ADD; no evaluated model exceeds 53.7% Acc on AT-ADD. SALMONN-2, which has documented



(a) Generated-answer accuracy varies substantially across benchmarks **(b) Low generation accuracy arises from distinct failure modes**

Figure 3: Native generated-answer behavior. (a) Generated-answer accuracy (Acc) varies substantially across model–benchmark pairs. (b) Validity separates incorrect judgments from refusals or non-decisions: Audio-Flamingo-Next remains near chance despite high validity, whereas GPT-audio rarely provides a valid decision. Full results are in Appendix D.

HoliAntiSpoof training exposure, performs strongly on several benchmarks but does not retain the same advantage across all datasets. Figure 3 further shows that low accuracy arises from qualitatively different failures. Audio-Flamingo-Next makes valid decisions on all evaluated 19LA, 21LA, and ITW utterances but remains near chance, whereas GPT-audio rarely produces an identifiable real/fake decision. Other models occasionally abstain or generate off-task content. Native generation therefore fails through at least three modes: incorrect decisions, non-decisions, and off-task responses.

Native answer-token scores provide complementary evidence about discrimination without generation. Excluding SALMONN-2, 19LA EER ranges from 22.18% to 55.17%; on the transfer sets, the lowest EERs across the ten open-weight models are 24.46% on 21LA, 22.62% on ITW, and 40.80% on AT-ADD. Native scoring can therefore provide useful discrimination in some settings, but remains uneven across models and datasets. Overall, native detection alone gives an incomplete picture of Audio-LLM deepfake detection: released interfaces differ both in decision quality and in whether they produce a valid authenticity judgment at all. This motivates examining what frozen representations can support once supervised readouts are introduced.

5.2 RQ2: How Effective and Transferable Are Frozen Audio-LLM Readouts?

We next evaluate what frozen Audio-LLM representations can support once a supervised readout is fitted. Audio-Linear and LM-Linear use the same 19LA source labels while keeping the pretrained Audio-LLM frozen. On held-out 19LA evaluation, each readout achieves EER below 10% for eight of the ten models (Table 1), establishing strong supervised decodability at both audio-side and final language-side interfaces for most evaluated models. This strong supervised decodability can coexist with substantially weaker native scoring; the two settings, however, differ in task supervision and should therefore be interpreted as complementary evaluations rather than a controlled attribution of native failure.

We then test whether this supervised decodability transfers by applying each source-trained readout unchanged to 21LA and ITW. Across all twenty model–readout combinations, EER increases from 19LA to 21LA and further to ITW. For example, Voxtral's Audio-Linear increases from 3.09% EER on 19LA to 12.58% on 21LA and 21.47% on ITW. Thus, strong source-domain decodability does not imply equally strong cross-domain transfer.

The relative effectiveness of the two frozen-readout pipelines is also not stable across evaluation domains. Kimi and Phi-4 favor Audio-Linear on 19LA and 21LA but LM-Linear on ITW; for Kimi, the ordering changes from 12.37% versus 22.15% EER on 21LA to 33.02% versus 24.23% on ITW. Additional AT-ADD- and ASV5-source experiments show that the ordering can also change with the

Table 1: Results for ten open-weight Audio-LLMs. Audio-Linear, LM-Linear, and Native-SFT use 19LA training data; ZS-Score requires no additional task-specific training. Evaluation uses the full evaluation sets; Acc and F1 use the fixed real–fake score comparison. All metrics are percentages; F1 denotes Macro-F1.

		19LA			21LA			ITW			AT-ADD			ASV5		
Model	Method	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
Qwen2.5-Omni-3B	ZS-Score	40.07	20.72	20.73	40.03	22.40	22.49	42.87	**56.57**	**60.43**	48.40	35.13	36.99	33.45	**60.66**	67.11
	Audio-Linear	1.70	87.29	94.18	17.69	**79.52**	**93.63**	**33.79**	30.04	38.82	42.59	52.03	80.34	**28.40**	54.72	**81.36**
	LM-Linear	6.06	88.53	95.19	19.40	75.91	92.29	40.75	30.44	38.72	48.85	**52.85**	80.68	30.00	47.92	80.08
	Native-SFT	**1.57**	**96.88**	**98.81**	**17.30**	75.66	93.35	44.14	27.17	37.22	**40.01**	49.64	**83.82**	29.22	45.08	79.77
Step-Audio-2-mini	ZS-Score	47.65	12.85	13.27	47.69	11.98	12.46	32.46	**55.87**	**68.15**	47.38	24.99	24.99	40.32	37.75	37.79
	Audio-Linear	2.68	91.67	96.52	**15.01**	**85.38**	**95.04**	28.40	42.84	46.75	45.53	51.97	70.53	**28.34**	**51.81**	**81.01**
	LM-Linear	3.48	93.50	97.44	23.96	78.37	93.48	38.91	34.21	40.51	49.03	**52.82**	76.13	41.47	45.91	79.66
	Native-SFT	**1.51**	**97.73**	**99.15**	26.87	77.17	93.66	**20.24**	28.69	38.07	**45.04**	49.44	**85.15**	34.41	44.49	79.66
Voxtral-Mini-3B	ZS-Score	55.17	47.56	85.00	53.21	47.71	83.74	40.68	38.89	43.30	46.13	49.66	**83.63**	36.24	44.94	79.31
	Audio-Linear	**3.09**	88.04	94.64	**12.58**	**85.97**	**94.83**	**21.47**	**67.43**	**67.46**	41.44	**53.03**	68.63	24.13	60.40	82.74
	LM-Linear	6.77	87.20	94.61	16.27	79.69	93.27	47.48	42.74	44.99	48.13	51.91	76.18	**22.22**	**65.16**	**83.45**
	Native-SFT	3.15	**94.26**	**97.79**	15.91	75.23	93.30	21.92	45.06	48.63	**41.05**	51.59	83.08	25.56	45.39	79.83
Phi-4-multimodal	ZS-Score	37.27	9.40	10.36	36.14	9.28	10.21	43.30	**38.62**	**62.82**	46.10	12.04	13.68	48.71	16.99	20.41
	Audio-Linear	6.77	88.82	95.52	15.54	**81.66**	**94.10**	39.13	33.33	40.49	48.82	50.07	80.87	**29.43**	**49.21**	**80.10**
	LM-Linear	10.28	84.46	93.90	20.95	72.62	91.10	29.67	37.28	42.84	50.05	**50.89**	82.89	38.18	49.08	79.96
	Native-SFT	**3.78**	**93.92**	**97.71**	**15.30**	77.22	93.67	**26.43**	28.94	38.20	**42.79**	46.88	**85.19**	30.16	44.45	79.64
Audio-Flamingo-Next	ZS-Score	42.12	47.28	89.68	42.89	47.32	89.84	61.61	27.10	37.18	50.74	46.33	**86.33**	41.45	44.33	79.63
	Audio-Linear	**1.72**	86.80	93.90	**20.11**	**80.25**	**92.81**	**30.50**	**44.02**	**47.74**	42.04	**54.91**	71.76	**26.68**	**60.37**	**82.66**
	LM-Linear	2.60	89.87	95.60	26.77	71.66	91.08	36.21	29.24	38.31	48.19	50.84	67.59	37.29	49.71	80.24
	Native-SFT	2.07	**92.24**	**96.77**	28.39	69.55	91.36	45.06	27.64	37.47	**41.07**	52.60	81.24	33.99	49.39	80.37
MOSS-Audio-4B	ZS-Score	22.18	20.49	20.49	24.46	21.71	21.73	34.24	43.08	**64.22**	**42.69**	19.45	19.87	**19.62**	**60.50**	62.85
	Audio-Linear	4.32	84.78	92.81	21.50	76.60	92.56	34.33	35.46	42.05	47.08	49.73	**76.73**	26.14	55.08	80.86
	LM-Linear	3.58	91.37	96.42	22.12	74.30	92.88	29.04	32.52	40.30	44.47	**51.86**	71.90	23.17	51.70	80.75
	Native-SFT	**1.77**	**92.38**	**96.81**	**19.99**	**76.78**	**93.32**	**28.33**	**43.88**	47.61	44.76	51.59	74.74	27.83	55.53	**81.40**
Kimi-Audio-7B	ZS-Score	38.78	40.43	46.04	38.05	40.59	46.38	22.62	**71.94**	**76.27**	47.32	34.44	36.24	42.42	46.20	48.59
	Audio-Linear	2.26	91.25	96.29	**12.37**	**88.01**	**95.66**	33.02	53.72	55.02	43.13	52.84	70.72	26.37	**62.77**	**83.34**
	LM-Linear	4.58	91.05	96.34	22.15	81.24	93.91	24.23	66.39	66.40	44.29	**55.04**	78.23	**24.31**	55.60	81.74
	Native-SFT	**0.65**	**98.70**	**99.51**	14.70	83.87	95.30	**13.02**	38.08	43.83	**42.89**	50.12	**84.10**	38.22	47.23	80.20
MiMo-Audio-7B	ZS-Score	44.66	9.55	10.48	39.61	9.46	10.36	**30.32**	39.24	**63.04**	48.08	14.45	15.55	25.11	23.38	25.21
	Audio-Linear	13.73	56.18	64.90	25.94	57.76	70.68	45.40	27.10	37.18	43.14	50.79	**84.77**	45.67	**50.36**	74.90
	LM-Linear	6.13	88.19	95.04	21.11	69.68	91.60	43.73	30.08	38.65	49.04	47.85	82.43	**24.78**	48.59	80.34
	Native-SFT	**2.03**	**94.92**	**97.99**	**16.25**	**80.30**	**93.99**	31.05	**48.89**	51.23	**41.72**	**51.40**	83.47	39.80	49.47	**80.65**
Gemma-4-12B-it	ZS-Score	46.04	9.56	10.49	47.65	9.43	10.33	48.23	38.68	62.77	50.79	12.24	13.83	57.72	17.09	20.47
	Audio-Linear	21.85	67.53	82.84	32.98	61.58	82.61	57.46	31.88	39.74	48.91	**52.32**	77.23	44.20	52.43	76.05
	LM-Linear	18.29	72.11	86.72	27.08	66.00	86.22	36.56	40.13	44.25	49.94	49.46	**79.99**	34.98	49.11	79.51
	Native-SFT	**0.56**	**95.99**	**98.42**	**12.83**	**77.15**	**88.73**	**31.08**	**66.28**	**66.46**	**43.83**	49.90	66.96	**18.63**	**70.89**	**84.86**
SALMONN-2*	ZS-Score	6.26	76.41	86.90	27.21	61.94	78.52	25.81	63.78	71.55	40.80	51.16	62.10	14.24	**81.57**	**89.64**
	Audio-Linear	1.52	93.45	97.32	13.06	**84.77**	**95.26**	25.65	58.99	59.76	42.37	51.79	75.53	23.98	73.55	86.70
	LM-Linear	1.30	97.40	99.01	19.59	82.20	94.59	27.71	68.84	68.85	40.45	**52.01**	64.00	26.32	69.54	85.54
	Native-SFT	**0.49**	**99.23**	**99.72**	**11.04**	76.71	93.83	**18.08**	**80.96**	**81.69**	**33.39**	50.22	**84.92**	**11.98**	55.73	82.08

* SALMONN-2 has documented HoliAntiSpoof training exposure.

training source. Frozen decodability is therefore substantial, but the relative advantage of audio-side and language-side readouts depends on the source–target protocol.

5.3 RQ3: How Does Language-Side Adaptation Change Detection and Transfer?

Having established strong supervised decodability from frozen representations, we next examine language-side adaptation. Native-SFT uses the same source-domain labels as the frozen readouts while keeping the audio pathway, encoder-to-language connection, and native output head frozen.

Native-SFT consistently improves in-domain detection over LM-Linear. On 19LA, it achieves lower EER, higher Macro-F1, and higher accuracy for all ten models; the same holds for Qwen, Kimi, and MOSS under the additional AT-ADD- and ASV5-source protocols. Its advantage over Audio-Linear

Table 2: In-domain and cross-domain results under AT-ADD and ASV5 source protocols. Audio-Linear, LM-Linear, and Native-SFT use the indicated source training split and are evaluated on the corresponding evaluation set and ITW without target-domain fitting.

Training source		AT-ADD						ASV5					
Evaluation set		AT-ADD			ITW			ASV5			ITW		
Model	Method	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
Qwen2.5-Omni-3B	ZS-Score	48.40	35.13	36.99	42.87	**56.57**	**60.43**	33.45	60.66	67.11	42.87	56.57	60.43
	Audio-Linear	**30.39**	64.94	79.19	**17.86**	50.48	52.61	**11.70**	**73.42**	77.30	33.57	55.44	**68.53**
	LM-Linear	35.81	62.78	77.35	39.30	34.28	40.67	28.21	62.47	80.27	31.71	59.55	59.63
	Native-SFT	34.06	**67.63**	**82.74**	45.10	27.24	37.25	22.96	73.35	**83.59**	**26.59**	**61.94**	61.95
Kimi-Audio-7B	ZS-Score	47.32	34.44	36.24	22.62	**71.94**	**76.27**	42.42	46.20	48.59	22.62	71.94	76.27
	Audio-Linear	36.40	58.11	72.74	19.79	71.62	71.63	11.92	83.87	88.35	19.34	67.86	75.08
	LM-Linear	38.40	56.16	69.65	28.86	64.17	64.22	21.35	77.72	86.18	18.82	80.51	81.12
	Native-SFT	**30.53**	**64.77**	**78.58**	**16.25**	37.82	43.60	**8.79**	**89.73**	**93.15**	**11.54**	**86.06**	**87.52**
MOSS-Audio-4B	ZS-Score	42.69	19.45	19.87	34.24	43.08	**64.22**	19.62	60.50	62.85	34.24	43.08	64.22
	Audio-Linear	32.99	60.87	74.41	**21.19**	**62.93**	63.12	15.64	72.21	76.25	**26.08**	52.59	67.93
	LM-Linear	34.67	60.11	73.64	28.30	58.51	58.85	28.02	69.27	81.61	49.11	45.78	46.28
	Native-SFT	**32.25**	**65.44**	**79.44**	32.76	39.94	44.44	**13.16**	**85.59**	**90.37**	34.63	**65.07**	**69.50**

ZS-Score requires no additional task-specific training; its ITW results are reproduced from Table 1.

is less uniform, indicating that language-side adaptation improves over frozen language-side readout without universally dominating the audio-side interface. Because LM-Linear and Native-SFT differ in trainable capacity, this comparison reflects the practical gain of adaptation rather than a capacity-matched control.

Cross-domain behavior is less uniform. Under 19LA-source training, Native-SFT achieves lower EER than both frozen readouts for six models on 21LA and seven on ITW. On AT-ADD, Native-SFT outperforms both frozen readouts in EER for nine of ten models and in accuracy for seven, but in Macro-F1 for only one; no model improves over both readouts in both fixed-rule metrics simultaneously. Thus, adaptation gains in score discrimination do not translate uniformly across fixed-rule evaluation metrics. For Kimi on ITW, EER improves from 24.23% to 13.02% over LM-Linear, while Macro-F1 drops from 66.39% to 38.08%; under ASV5-source training, Native-SFT instead improves both metrics. Its transfer benefit is therefore source-, target-, and metric-dependent.

Native-SFT preserves the released encoder-to-language interface, leaving open whether another encoder depth would provide a better adaptation input and whether frozen-readout preferences persist after adaptation. We examine these questions next with Depth-SFT.

5.4 RQ4: How Does Encoder Depth Affect Decodability and Adaptation Utility?

We perform complete encoder-depth sweeps across six Audio-LLMs and three downstream procedures: Linear, AASIST, and Depth-SFT. Figure 4 highlights three representative patterns from these sweeps using EER and Macro-F1, while complete results for all six models, evaluation sets, and metrics are provided in Appendix F. All branches use 19LA-source supervision. Within each model and branch, preprocessing and training settings are held fixed across depths, and Depth-SFT is compared with its own final-block configuration. Because Linear, AASIST, and Depth-SFT differ in feature processing and trainable capacity, we focus on within-procedure depth preferences rather than absolute performance differences across procedures.

Intermediate encoder representations can improve adaptation across both discrimination and decision metrics. For Qwen, every Depth-SFT connection from L08 through L23 achieves lower EER than the final L31 connection on both 21LA and ITW. On 21LA, this region yields 6.38–12.96% EER versus 21.78% at L31; on ITW, it yields 12.24–39.42% versus 46.15%. Every depth in the same region also achieves higher Macro-F1 than L31 on both target datasets. The advantage therefore extends across a contiguous intermediate region and is not confined to score discrimination.

Adaptation preferences can change across target domains. For StepAudio Depth-SFT, L12–L19 is favored over L27–L31 on 21LA in EER, with ranges of 9.78–11.99% and 19.95–22.26%,

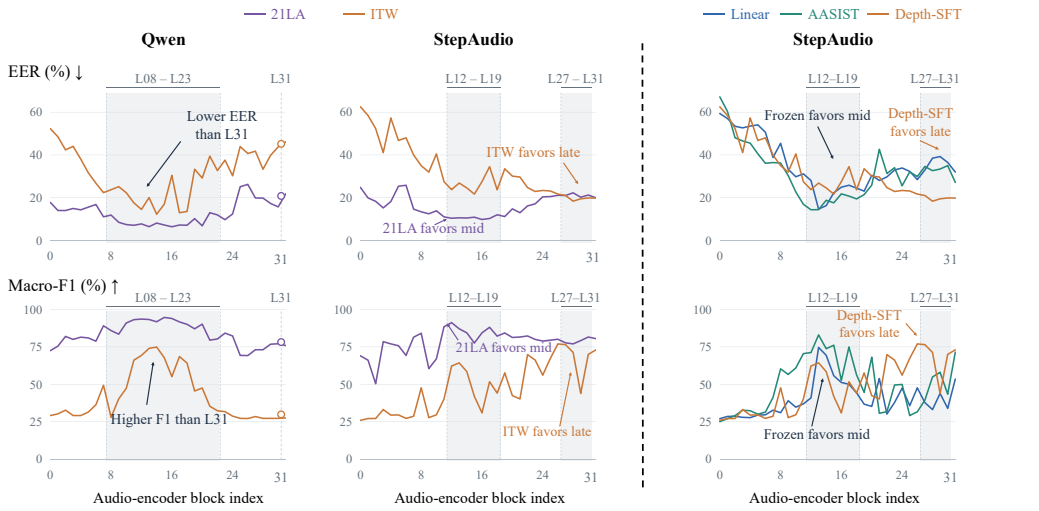


Figure 4: Selected encoder-depth findings under 19LA-source supervision. Top and bottom rows show EER and Macro-F1, respectively. (a) Qwen Depth-SFT across 21LA and ITW. (b) Target-dependent StepAudio Depth-SFT. (c) Frozen-readout versus Depth-SFT preferences for StepAudio on ITW. Shaded spans mark the regions discussed in the text; complete six-model profiles are in Appendix F.

respectively. On ITW, this ordering reverses: L27–L31 achieves 18.35–21.01% EER, below the 21.66–34.49% range at L12–L19. Macro-F1 exhibits the same overall regional shift, although the separation is not uniform at every depth. Thus, the preferred adaptation region depends on the target domain rather than being a fixed property of encoder depth.

Frozen-readout preferences need not transfer to adaptation. For StepAudio on ITW, both Linear and AASIST favor L12–L19 over L27–L31 in EER. Depth-SFT reverses this regional ordering: L27–L31 yields 18.35–21.01% EER, uniformly below the 21.66–34.49% range at L12–L19. Macro-F1 shows the same broad contrast in regional preference, although not with uniform separation at every depth. Across all six models, rank correlations and Top-5 overlaps show heterogeneous frozen-readout–Depth-SFT agreement across models and targets (Appendix F).

Together, these analyses show that encoder depth affects both score discrimination and fixed-rule decision performance. Intermediate representations can improve adaptation over final-block inputs, while adaptation-preferred regions can shift across target domains and need not coincide with those favored by frozen readouts.

6 Conclusion

Using RAVEL, we examined Audio-LLM speech deepfake detection through native outputs, frozen readouts, and task adaptation, together with depth-wise analysis of audio-encoder representations. Our results show that the presence of authenticity-related information and its effective downstream use are not the same. Weak native detection can coexist with strong supervised decodability from frozen representations, indicating that limited zero-shot behavior need not imply the absence of discriminative internal information. Depth-wise adaptation further shows that intermediate encoder representations can provide substantially stronger cross-domain discrimination than the final block. However, encoder regions favored by linear and nonlinear frozen readouts need not be the most effective inputs for encoder-to-language adaptation, and their relative utility can change across target domains. Under 19LA→AT-ADD transfer, Native-SFT beats both frozen readouts in EER for 9/10 models but in Macro-F1 for only 1/10, revealing a gap between discrimination and fixed-threshold decisions. Together, these findings highlight representation utilization as a key consideration in Audio-LLM speech deepfake detection: what is readily decodable from a frozen representation does not determine how effectively it can be exploited by the language pathway after adaptation.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language polishing. All AI assisted text was reviewed by the authors, and the experimental results, analyses, and scientific claims were verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide an anonymous supplementary code package for the ten open-weight Audio-LLMs, including implementations for native evaluation (ZS-Gen and ZS-Score), frozen readouts (Audio-Linear and LM-Linear), Native-SFT, layer-wise Linear and AASIST readouts, and Depth-SFT. The package includes model-specific configuration templates, dependency and setup instructions, fixed sample identifiers for the ZS-Gen subsets, and utilities for semantic adjudication and metric computation. The README documents model and branch coverage, execution commands, and configuration provenance. Some depth-analysis configurations are supplied as implementation-derived templates rather than archived original-run configurations, as identified in the README. Metric implementations and output conventions are documented in `METRICS.md`.

Appendix A specifies the datasets, evaluation splits, source–target protocols, and zero-shot sampling procedure; Appendix B documents the evaluated models and representation interfaces; Appendix C provides preprocessing, optimization, trainable and frozen components, checkpoint selection, random seeds, and metric definitions; Appendix D describes prompting, generation, and semantic adjudication; and Appendix F reports the complete encoder-depth results.

Public datasets, pretrained model checkpoints, and upstream model source trees are obtained separately following the setup instructions in the README. API client implementations are not included in the supplementary package. Results involving API-accessed models depend on the continued availability of the corresponding provider model versions.

REFERENCES

Marjan Beheshti, Majid Rostami, and Bo Chen. Probing-guided layer selection from self-supervised speech models for generalizable audio deepfake detection. *arXiv preprint arXiv:2606.30791*, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Runkun Chen, Yixiong Fang, Pengyu Chang, Yuante Li, Massa Baali, and Bhiksha Raj. Audio language model for deepfake detection grounded in acoustic chain-of-thought. *arXiv preprint arXiv:2603.28021*, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Akanksha Chuchra, Shukesh Reddy, Sudeepta Mishra, Abhijit Das, and Abhinav Dhall. Investigating the viability of employing multi-modal large language models in the context of audio deepfake detection. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11. IEEE, 2025.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Artem Dvirniak, Evgeny Kushnir, Dmitrii Tarasov, Artem Iudin, Oleg Kiriukhin, Mikhail Pautov, Dmitrii Korzh, and Oleg Y Rogov. Towards robust speech deepfake detection via human-inspired reasoning. *arXiv preprint arXiv:2603.10725*, 2026.

Yassine El Kheir, Younes Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. Comprehensive layer-wise analysis of ssl models for audio deepfake detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4070–4082, 2025.

Jiajun Fan, Jingyuan Li, Prashanth Gurunath Shivakumar, Qi Luo, Jia-Hong Huang, M Maruf, Roger Ren, Yile Gu, Rahul Pandey, Ge Liu, et al. Can we read the mind of an audio llm? a verbalizable, multilingual middle-layer workspace. *arXiv preprint arXiv:2608.24958*, 2026.

Google DeepMind. Gemini 3.8 flash: Model card. Google DeepMind Model Card, September 2026. Accessed: 2026-09-18.

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11736–11745, 2025.

Xiaoxuan Guo, Hengyan Huang, Jiayi Zhou, Renhe Sun, Jian Liu, Haonan Cheng, Long Ye, and Qin Zhang. Envsslam-ffn: Lightweight layer-fused system for esdd 2026 challenge. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21769–21771. IEEE, 2026a.

Xiaoxuan Guo, Yuankun Xie, Haonan Cheng, Jiayi Zhou, Jian Liu, Hengyan Huang, Long Ye, and Qin Zhang. Towards explicit acoustic evidence perception in audio llms for speech deepfake detection. *arXiv preprint arXiv:2601.23066*, 2026b.

Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12702–12706. IEEE, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Hengyan Huang, Xiaoxuan Guo, Jiayi Zhou, Yuankun Xie, Jian Liu, Haonan Cheng, Long Ye, and Qin Zhang. Envtricascade: An environment-aware tri-stage cascaded framework for esdd2 2026 challenge. *arXiv preprint arXiv:2605.18409*, 2026.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38, 2025a.

Yupei Li, Lin Wang, Yifei Wang, Lijun Wang, Runze Cai, Jing Shi, Björn W. Schuller, and Zhizheng Wu. Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components. *arXiv preprint arXiv:2512.08403*, 2025b.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *arXiv preprint arXiv:2210.02437*, 2022.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

OpenAI. GPT-Audio Model. `https://developers.openai.com/api/docs/models/gpt-audio`, 2025. Model snapshot: gpt-audio-2025-08-28; accessed 2026-09-18.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. *arXiv preprint arXiv:2605.27772*, 2026.

Pierre Serrano, Raphaël Duroselle, Florian Angulo, Jean-François Bonastre, and Olivier Boeffard. Improving out-of-domain audio deepfake detection via layer selection and fusion of ssl-based countermeasures. *arXiv preprint arXiv:2509.12003*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*, 2024.

Yuankun Xie, Haonan Cheng, Jiayi Zhou, Xiaoxuan Guo, Tao Wang, Changhao Zhang, Jian Liu, Weiqiang Wang, Ruibo Fu, Xiaopeng Wang, et al. At-add: A benchmark and challenge for robust and all-type audio deepfake detection. *arXiv preprint arXiv:2608.23437*, 2026a.

Yuankun Xie, Xiaoxuan Guo, Jiayi Zhou, Tao Wang, Jian Liu, Ruibo Fu, Xiaopeng Wang, Haonan Cheng, and Long Ye. Interpretable all-type audio deepfake detection with audio llms via frequency-time reinforcement learning. *arXiv preprint arXiv:2601.02983*, 2026b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report, 2025. *URL https://arxiv.org/abs/2503.20215*, 2025.

Xuenan Xu, Yiming Ren, Liwei Liu, Wen Wu, Baoxiang Li, Chaochao Lu, Shuai Wang, and Chao Zhang. Holiantispoof: Audio llm for holistic speech anti-spoofing. *arXiv preprint arXiv:2602.04535*, 2026.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

Xiaoyu Yang, Xuenan Xu, Wenyi Yu, Siyin Wang, Changli Tang, Terumi Chiba, Siyuan Hou, Ziyang Zhang, Wen Wu, Baoxiang Li, et al. Salmonn-2: Advancing general-purpose hearing abilities with self-supervised representations. *arXiv preprint arXiv:2607.17079*, 2026b.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 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.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6765–6773, 2024.

Table 3: Datasets and split sizes in utterances. Counts refer to complete partitions. Dashes indicate training or development partitions not used in this study.

Dataset	Role	Train	Dev	Eval
ASVspoof 2019 LA (19LA)	Source benchmark	25,380	24,844	71,237
AT-ADD 2026 Track 1 (AT-ADD)	Source benchmark	49,575	49,734	146,346
ASVspoof 5 Track 1 (ASV5)	Source benchmark	182,357	140,950	680,774
ASVspoof 2021 LA (21LA)	Held-out test set	–	–	181,566
In-the-Wild (ITW)	Held-out test set	–	–	31,779

A Datasets and Evaluation Protocols

A.1 Dataset Statistics and Splits

Our evaluation suite comprises three source benchmarks—ASVspoof 2019 LA (19LA), AT-ADD 2026 Track 1 (AT-ADD), and ASVspoof 5 Track 1 (ASV5)—and two additional held-out test sets, ASVspoof 2021 LA (21LA) and In-the-Wild (ITW). Each source benchmark provides separate training, development, and evaluation partitions. We use 21LA and ITW exclusively for held-out evaluation. Table 3 summarizes the complete split sizes in utterances.

A.2 Supervised Source Protocols

We consider three supervised source protocols based on 19LA, AT-ADD, and ASV5. Within each protocol, Audio-Linear, LM-Linear, and Native-SFT use the same source-domain training split. Development-based checkpoint selection, where applicable, uses only the corresponding source development split. The trained components are then applied unchanged to the evaluation sets.

19LA-source protocol. The main supervised comparison covers all ten open-weight Audio-LLMs. Audio-Linear, LM-Linear, and Native-SFT are trained on the 19LA training split and evaluated on the held-out 19LA evaluation split. The same trained configurations are evaluated on 21LA, ITW, AT-ADD, and ASV5 to assess cross-domain transfer. Under this protocol, AT-ADD and ASV5 serve as target datasets, and their training and development partitions are not used.

AT-ADD-source and ASV5-source protocols. We additionally evaluate Qwen2.5-Omni-3B, Kimi-Audio-7B, and MOSS-Audio-4B under two alternative training sources. For each source, Audio-Linear, LM-Linear, and Native-SFT are fitted separately using its training partition and evaluated on its corresponding evaluation partition and the shared ITW test set. These experiments support comparisons of in-domain detection and transfer to the same target dataset across different source protocols.

Source protocol for encoder-depth analysis. All layer-wise linear probes, AASIST readouts, and Depth-SFT configurations use 19LA-source supervision. Each depth is trained independently using the 19LA training split, with source-development checkpoint selection where applicable, and evaluated on 19LA eval, 21LA, and ITW. The same trained configuration at each depth is used across all three evaluation sets.

A.3 Zero-Shot Evaluation Sets

ZS-Gen and ZS-Score evaluate released models without additional task-specific training on all five datasets: 19LA, 21LA, ITW, AT-ADD, and ASV5. The two settings use different evaluation sample sizes.

Generated-answer evaluation. For each dataset, ZS-Gen uses a fixed class-balanced subset of 1,000 utterances, comprising 500 bona fide and 500 spoofed samples drawn from its evaluation set. The same utterance identifiers are used across all ten open-weight and three closed-source models,

ensuring that their generated answers are evaluated on identical inputs. Each model is therefore evaluated on 5,000 utterances across the five datasets.

Native-score evaluation. ZS-Score covers the ten open-weight models and uses the complete evaluation sets listed in Table 3. These evaluations retain the original class distributions without class-balanced subsampling. All ten models are evaluated on the same complete set of utterances within each dataset.

B Models and Representation Interfaces

B.1 Evaluated Models and Experimental Coverage

Open-weight models. We evaluate ten open-weight Audio-LLMs: Qwen2.5-Omni-3B, Step-Audio-2-mini, Voxtral-Mini-3B, Phi-4-multimodal, Audio-Flamingo-Next, MOSS-Audio-4B, Kimi-Audio-7B, MiMo-Audio-7B, Gemma-4-12B-it, and SALMONN-2. All ten participate in ZS-Gen and ZS-Score evaluation and in the 19LA-source comparison of Audio-Linear, LM-Linear, and Native-SFT. These experiments cover all five evaluation datasets.

Closed-source models. Three closed-source models accessed through APIs—Qwen3.5-Omni-Flash Team (2026), GPT-audio OpenAI (2025), and Gemini-3.8-Flash Google DeepMind (2026)—participate in ZS-Gen evaluation. Their generated responses are evaluated on the same fixed utterance subsets and under the same semantic adjudication procedure as those of the open-weight models.

Additional source protocols. Qwen2.5-Omni-3B, Kimi-Audio-7B, and MOSS-Audio-4B additionally participate in the AT-ADD-source and ASV5-source experiments. Each model is evaluated with Audio-Linear, LM-Linear, and Native-SFT under both protocols, covering in-domain evaluation and transfer to ITW.

Encoder-depth coverage. Layer-wise linear probing, AASIST probing, and Depth-SFT cover the same six models: Qwen2.5-Omni-3B, Step-Audio-2-mini, Voxtral-Mini-3B, Phi-4-multimodal, Audio-Flamingo-Next, and MOSS-Audio-4B. These sweeps evaluate 24 encoder blocks for Phi-4-multimodal and 32 blocks for each of the other participating models, using 19LA-source supervision and evaluation on 19LA eval, 21LA, and ITW.

Documented task-specific training exposure. The released SALMONN-2 checkpoint Yang et al. (2026b) has documented spoof-oriented instruction-tuning exposure associated with HoliAntiSpoof Xu et al. (2026). We retain SALMONN-2 in the evaluation and mark its results with an asterisk to identify this prior exposure.

B.2 Audio-Side and Language-Side Interfaces

We define representation interfaces by their functional roles in each model's audio-to-language pipeline. The fixed-interface experiments use an audio-side sequence and a language-side vector, while the depth-wise experiments access individual audio-encoder block outputs.

Audio-side interface. For an input utterance x, the audio-side representation is

$$h_{\mathrm{aud}}(x) = P(A(x)) \in \mathbb{R}^{T_{\mathrm{aud}} \times d_{\mathrm{aud}}}, \tag{2}$$

where A denotes the audio encoder and P comprises the model-specific operations up to the selected extraction point. These operations may include terminal normalization, temporal reduction, projection, adapter transformation, or embedding fusion. Here, T_{aud} is the sequence length and d_{aud} is the channel dimension. Audio-Linear summarizes this sequence by concatenating its channel-wise temporal mean and standard deviation, producing a feature vector of dimension $2d_{\mathrm{aud}}$.

Table 4: Model-specific representation interfaces. Audio-side taps identify the sequence used by Audio-Linear; language-side dimensions correspond to the final vector supplied to the native output head. RVQ and MLP denote residual vector quantization and multilayer perceptron, respectively.

Model	Audio-side representation tap	d_{aud}	d_{lm}
Qwen2.5-Omni-3B	Audio projector output	2048	2048
Step-Audio-2-mini	Audio adapter output	3584	3584
Voxtral-Mini-3B-2507	Multimodal projector output	3072	3072
Phi-4-multimodal-instruct	Speech projection output	3072	3072
Audio-Flamingo-Next	Multimodal projector output	3584	3584
MOSS-Audio-4B-Instruct	Audio adapter output	2560	2560
Kimi-Audio-7B-Instruct	Post-normalization continuous–discrete fusion	3584	3584
MiMo-Audio-7B-Instruct	Final continuous pre-RVQ representation	768	4096
Gemma-4-12B-it	Audio embedding projection output	3840	3840
SALMONN-2-8B*	Audio connector MLP output	4096	4096

* SALMONN-2 has documented HoliAntiSpoof training exposure.

Language-side interface. Let $H_{\mathrm{LM}}(x, p) \in \mathbb{R}^{N \times d_{\mathrm{lm}}}$ denote the sequence supplied to the native language-model output head after final normalization, given utterance x and instruction p. We extract

$$h_{\mathrm{lm}}(x, p) = H_{\mathrm{LM}}(x, p)_{N-1} \in \mathbb{R}^{d_{\mathrm{lm}}}, \tag{3}$$

using zero-based indexing. This vector is obtained at the final context position and predicts the first answer token. The context contains the audio, instruction, and native template tokens, without an appended reference answer. LM-Linear operates directly on this vector without temporal pooling. ZS-Score uses the same functional location but obtains authenticity scores through the pretrained output head.

Raw encoder-block interfaces. For depth-wise analysis, $h_l(x) = A_l(x)$ denotes the raw output of encoder block l, captured before subsequent encoder-terminal operations and audio-to-language transformations. Layer-wise linear and AASIST readouts operate on these raw block representations. Because the audio-side interface may include additional processing, the final raw encoder-block output h_{L-1} need not coincide with h_{aud}.

B.3 Model-Specific Interface Details

Table 4 lists the audio-side extraction point and representation dimensions for each open-weight model. The audio-side dimension d_{aud} refers to sequence channels before temporal pooling, whereas d_{lm} is the dimension of the single language-side vector. Audio-Linear therefore uses $2d_{\mathrm{aud}}$ pooled features, while LM-Linear uses d_{lm} features. These interfaces reflect the model families summarized in Figure 1; models with nonstandard audio-to-language pathways require additional treatment as described below.

Kimi's hybrid audio representation. Kimi-Audio-7B combines continuous and discrete audio representations before language-side processing. Audio-Linear uses the post-normalization continuous–discrete fusion sequence, which has 3584 channels and yields a 7168-dimensional feature after temporal mean–standard-deviation pooling.

MiMo's discrete-token pathway. MiMo-Audio-7B ultimately supplies discrete audio tokens to the language backbone. To retain a continuous sequence for Audio-Linear, we tap the final continuous representation immediately before RVQ. Its 768 channels yield a 1536-dimensional pooled feature, while LM-Linear uses the 4096-dimensional language-side vector. The two interfaces therefore differ in both processing stage and dimensionality.

Gemma-4's encoder-free pathway. Gemma-4-12B-it follows the encoder-free architecture category shown in Figure 1. Its audio representation does not pass through a standalone stack of audio-

encoder blocks comparable to those used in our layer-wise analysis. We therefore include Gemma-4 in the fixed-interface experiments but not in the encoder-depth sweeps.

SALMONN-2's multi-layer encoder interface. SALMONN-2 uses an audio encoder but constructs its audio-to-language representation through multi-layer feature fusion rather than a single selected encoder block. Consequently, replacing the released fused interface with one individual block would change the representation pathway rather than isolate depth under the same interface. We therefore retain SALMONN-2 in the fixed-interface evaluation but exclude it from the encoder-depth analysis.

C Implementation Details and Evaluation Metrics

C.1 Audio Preprocessing and Feature Extraction

Waveform preparation for audio-side linear readouts. Audio-Linear and layer-wise linear probing use waveforms standardized to 64,600 samples at 16 kHz, corresponding to approximately 4.04 seconds. Longer recordings are truncated, while shorter recordings are repeated and trimmed to the required length. The resulting waveform is processed by the model-specific audio front end. This waveform preparation is applied consistently to the source training and development splits and all evaluation sets.

Native audio-processing pathways. ZS-Gen, ZS-Score, LM-Linear, and Native-SFT retain each model's released audio-processing and duration-handling pipeline. Resampling, acoustic feature computation, padding, and duration handling follow the corresponding model processor or inference wrapper. These pathways therefore preserve model-specific input handling rather than imposing the common waveform length used for audio-side linear probing.

Frozen audio-side feature extraction. Audio-Linear extracts the model-specific sequence h_{aud} listed in Table 4, with all pretrained components frozen. Layer-wise linear probing extracts raw encoder-block outputs h_l; the block representations for an utterance are collected from the same frozen encoder forward pass. Each sequence is summarized by concatenating its channel-wise temporal mean and standard deviation:

$$\mathrm{Pool}(h) = [\mu_t(h); \sigma_t(h)] \in \mathbb{R}^{2d}, \tag{4}$$

where d is the sequence channel dimension. The native feature dimensionality is retained without additional dimensionality reduction.

Language-side feature extraction. LM-Linear processes the audio–instruction context through the frozen model and extracts the post-normalization vector supplied to the native output head at the position predicting the first answer token. Feature extraction uses a prefill forward pass without appending a reference answer or generating an answer sequence. The resulting single vector is used directly, without temporal pooling. ZS-Score reads the native answer-token logits at the same functional position; its extraction uses batch size one and introduces no batch-padding positions.

Sequence inputs for depth-wise experiments. AASIST readouts receive frozen raw encoder hidden sequences, preserving the temporal dimension without applying the mean–standard-deviation pooling used by linear probes. Depth-SFT passes the selected block's hidden sequence through the model-specific terminal operations and audio-to-language connection before language-side processing. Within each model and depth-analysis branch, preprocessing and feature-handling settings are held fixed across encoder depths.

C.2 Native Scoring and Frozen Linear Readouts

Native answer-token scoring. ZS-Score uses the released language-model output head without task-specific fitting. Let z_r and z_f denote the logits of the designated single-token answers `real` and `fake` at the first answer position. The native authenticity score is

$$s_{\mathrm{nat}}(x, p) = z_r - z_f, \tag{5}$$

where larger values indicate a stronger preference for bona fide speech. Normalization over the two answer tokens gives

$$P_{\mathrm{nat}}(\mathrm{real} \mid x, p) = \frac{\exp z_r}{\exp z_r + \exp z_f} = \sigma(s_{\mathrm{nat}}(x, p)) \,. \tag{6}$$

This computation uses only the audio–instruction context and requires no autoregressive answer generation.

Feature construction and standardization. Audio-Linear uses $u_{\mathrm{aud}} = \mathrm{Pool}(h_{\mathrm{aud}})$, with pooling defined in Eq. 4, whereas LM-Linear uses $u_{\mathrm{lm}} = h_{\mathrm{lm}}$. All pretrained Audio-LLM parameters remain frozen. For each model and representation interface, we fit a separate feature-wise standardizer using only the source training split. The resulting source-training means and scales are then used to transform development and evaluation features. We denote the standardized feature vector by $\widetilde{u}$.

Linear classifier fitting. Each readout is an independently fitted L_2-regularized logistic-regression classifier with inverse regularization strength $C = 1$ and balanced class weights. Its bona fide decision score and probability are

$$s_{\mathrm{lin}}(u) = w^\top \widetilde{u} + b, \qquad P_{\mathrm{lin}}(\mathrm{real} \mid u) = \sigma(s_{\mathrm{lin}}(u)) \,. \tag{7}$$

Audio-Linear and LM-Linear use the same source-domain authenticity labels but have separate standardizers and classifier parameters. Layer-wise linear probing follows the same procedure, fitting a separate standardizer and classifier for each raw encoder-block representation.

Inference and transfer. For all score-based detectors, predictions are obtained by comparing the real and fake decision scores. Equivalently, for methods represented as a two-class probability, this corresponds to a probability threshold of 0.5. The same decision rule is applied to source and target evaluation sets. EER is computed from continuous detection scores, while accuracy and Macro-F1 use the fixed classification rule. Once fitted, each linear readout and its source-domain standardizer are applied unchanged to the source evaluation split and all target datasets.

C.3 NATIVE-SFT

Trainable components. Native-SFT adapts the native language pathway using source-domain authenticity labels. The audio encoder, any model-specific audio tokenizer or codec, the audio-to-language connection, and the native output head remain frozen. For nine of the ten open-weight models, we introduce language-side LoRA modules into the query, key, value, and output projections, using rank $r = 8$, scaling factor $\alpha = 32$, and dropout 0.05. Phi-4-multimodal instead updates its released language-side speech and vision LoRA modules, which cover attention and MLP transformations. All remaining pretrained parameters remain frozen.

Single-token training objective. Each training utterance is paired with the same authenticity instruction used for native evaluation and one target answer token. Let $y = 0$ denote bona fide speech and $y = 1$ denote spoofed speech, with target token IDs $a(0) = r$ and $a(1) = f$ corresponding to `real` and `fake`. Given frozen parameters θ and trainable adaptation parameters ϕ, Native-SFT minimizes

$$\mathcal{L}_{\mathrm{SFT}}(\phi) = -\frac{1}{|\mathcal{D}_s^{\mathrm{tr}}|} \sum_{(x,y)\in\mathcal{D}_s^{\mathrm{tr}}} \log \frac{\exp z_{a(y)}(x, p; \theta, \phi)}{\sum_{v\in\mathcal{V}} \exp z_v(x, p; \theta, \phi)}, \tag{8}$$

where $\mathcal{V}$ is the complete language-model vocabulary and z_v is the logit for token v at the first answer position. Each example supervises exactly one authenticity token. Prompt, audio, padding, and end-of-sequence positions do not contribute to the loss.

Model-specific loss implementation. For implementations supporting appended-answer training, we append the target authenticity token and mask all other target positions, using the causal prediction alignment to compute its loss. For prompt-only implementations, the target answer is kept outside the input sequence and the loss is computed directly from the next-token logits at the final audio–instruction position. Both implementations optimize the full-vocabulary objective in Eq. 8.

Optimization and checkpoint selection. Training uses five epochs, a learning rate of 5×10^{-5}, cosine learning-rate decay, a warm-up ratio of 0.01, weight decay of 0.1, and random seed 42. Implementations with periodic source-development validation select the checkpoint with the lowest development loss. The remaining implementations use the final checkpoint after the prescribed training schedule. Within each source protocol, Native-SFT uses the same training split as Audio-Linear and LM-Linear.

Evaluation. At inference time, the adapted model receives only the audio–instruction context. Let $z_r^\star$ and $z_f^\star$ denote the resulting authenticity-token logits. The bona fide probability is

$$P_{\mathrm{SFT}}(\mathrm{real} \mid x, p) = \frac{\exp z_r^\star}{\exp z_r^\star + \exp z_f^\star}. \tag{9}$$

We compute EER from the continuous authenticity scores and obtain accuracy and Macro-F1 from the fixed real–fake score comparison. The selected adaptation parameters are applied unchanged to the source evaluation split and all target datasets.

C.4 Depth-Wise Training Procedures

We evaluate each audio-encoder depth independently through linear readout, nonlinear AASIST readout, and Depth-SFT. All branches use 19LA training data. Within each model and branch, preprocessing, supervision, and training settings are held fixed across depths.

Layer-wise linear readouts. For each encoder block l, we apply temporal mean–standard-deviation pooling to its frozen hidden sequence h_l. A separate feature standardizer and logistic-regression classifier are fitted using the source training split, following Appendix C.2. Each classifier uses L_2 regularization with $C = 1$ and balanced class weights. Standardization statistics and classifier parameters are learned independently at each depth, while all pretrained Audio-LLM parameters remain frozen.

Layer-wise AASIST readouts. For each depth, we train a separate AASIST detector on the frozen raw encoder hidden sequence. A trainable input projection maps the encoder channels to the detector's input space, and the AASIST readout preserves temporal structure through its nonlinear sequence processing. The input projection and detector are trained using source-domain authenticity labels; the audio encoder remains frozen and the language backbone is bypassed. The detector outputs a bona fide probability for each utterance. This branch complements pooled linear probing by evaluating the same encoder depths through a nonlinear sequence-level readout.

Depth-SFT. At depth l, the encoder is truncated after block l. Its output is passed through the model-specific terminal operations and audio-to-language connection before entering the language backbone. Each configuration uses a single encoder block as its connection point. We train the selected projection or adapter, designated terminal-normalization parameters where applicable, and language-side LoRA parameters. Encoder blocks, the native output head, and all remaining pretrained parameters stay frozen. Each depth is trained independently with the full-vocabulary, single-answer-token objective in Eq. 8. Evaluation uses the two-token normalization in Eq. 9.

Model-specific Depth-SFT configurations. For Qwen2.5-Omni-3B, the trainable connection includes the terminal `ln_post` normalization and audio projector. For Step-Audio-2-mini, the audio adapter is trainable. Both models use language-side LoRA on attention and MLP linear transformations, including the query, key, value, output, gate, up, and down projections. Their sweeps use five training epochs, learning rate 5×10^{-5}, cosine learning-rate decay, LoRA rank 8, and scaling factor $\alpha = 32$. Voxtral-Mini-3B, Phi-4-multimodal, MOSS-Audio-4B, and Audio-Flamingo-Next use five-epoch Depth-SFT configurations that train the model-specific audio-to-language connection and language-side attention LoRA; for Audio-Flamingo-Next, the trainable connection is its multimodal projector. Because Depth-SFT additionally adapts the connection, and its LoRA scope can also differ from Native-SFT, depth comparisons use its own final-block configuration as the reference.

Evaluation across depths. Each sweep uses random seed 42. Checkpoint selection, where applicable, uses only the 19LA development split. The trained detector at each depth is applied unchanged to 19LA eval, 21LA, and ITW. We retain complete depth profiles, reporting EER from continuous scores and accuracy and Macro-F1 from the fixed real–fake score comparison. Target-domain minima describe the observed profiles; they are not used to select a layer for deployment.

C.5 Score-Based Evaluation Metrics

We report equal error rate (EER), Macro-F1, and accuracy for ZS-Score, Audio-Linear, LM-Linear, Native-SFT, and the depth-wise detectors. All metrics are expressed as percentages. Labels are encoded as $y = 0$ for bona fide speech and $y = 1$ for spoofed speech. Detection scores are oriented such that larger values indicate stronger evidence for bona fide speech.

Equal error rate. Let s_i denote the continuous detection score for utterance i, and let N_0 and N_1 be the numbers of bona fide and spoofed utterances, respectively. At threshold τ, the false acceptance rate for spoofed speech and the false rejection rate for bona fide speech are

$$\begin{aligned} \mathrm{FAR}(\tau) &= \frac{1}{N_1} \sum_{i:y_i=1} \mathbb{I}[s_i \geq \tau], \\ \mathrm{FRR}(\tau) &= \frac{1}{N_0} \sum_{i:y_i=0} \mathbb{I}[s_i < \tau], \end{aligned} \tag{10}$$

where $\mathbb{I}[\cdot]$ is the indicator function. EER is the common error rate at the equal-error operating point of the threshold-swept curve, multiplied by 100. Score orientation remains fixed across datasets. The EER operating point is used only to calculate EER; it does not determine the threshold used for accuracy or Macro-F1.

Fixed score comparison. Let s_i denote the continuous detection score, where larger values indicate stronger evidence for bona fide speech. All score-based methods use the fixed decision rule

$$\hat{y}_i = \begin{cases} 0, & s_i \geq 0, \\ 1, & s_i < 0. \end{cases} \tag{11}$$

This rule compares the real and fake scores directly. For methods represented through normalized two-class probabilities, it is equivalent to assigning the real class when $P(\mathrm{real}) \geq 0.5$. The same decision rule is used for source-domain and target-domain evaluation.

Accuracy and Macro-F1. For an evaluation set containing N utterances, accuracy is

$$\mathrm{Acc} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}[\widehat{y}_i = y_i]. \tag{12}$$

Let TP_c, FP_c, and FN_c denote the true-positive, false-positive, and false-negative counts when class c is treated as the positive class. Macro-F1 is

$$\text{Macro-F1} = \frac{100}{2} \sum_{c \in \{0,1\}} \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{13}$$

Macro-F1 gives equal weight to the two class-wise F1 scores, while accuracy measures the fraction of correctly classified utterances under the evaluation set's class distribution. Both metrics are computed from the predictions in Eq. 11.

C.6 Reproducibility Settings

Random seed and configuration control. Native-SFT training and encoder-depth sweeps use random seed 42. Each reported depth profile contains one run per depth for the corresponding model and analysis branch. Within a depth sweep, preprocessing, detector architecture, optimization settings, and evaluation procedures are held fixed. Model-specific trainable components follow the configurations described in Appendices C.3 and C.4.

Common Zero-Shot Prompt

User Prompt
Is this audio real or fake? Answer with 'real' or 'fake'.

Figure 5: Common zero-shot prompt used for generated-answer and answer-token evaluation across all evaluated Audio-LLMs.

Source-specific fitting and evaluation. Each fitted component is associated with its source protocol. Feature standardizers and linear classifiers are estimated from the source training split, and development-based checkpoint selection uses the corresponding source development split. Implementations without periodic development evaluation retain the final checkpoint after the prescribed training schedule. The same fitted standardizer, classifier, or adaptation checkpoint is reused across the source evaluation split and all target datasets. Score orientation and the fixed classification threshold remain unchanged across these evaluations.

Metric recomputation and numerical precision. Where per-utterance outputs are available, score-based metrics can be recomputed offline from the saved scores and reference labels using the definitions in Section C.5. Score-based summary tables report EER, Macro-F1, and accuracy to two decimal places. Encoder-depth tables report EER and accuracy to two decimal places and Macro-F1 to one decimal place, matching the available result precision. Generated-answer tables report metrics to one decimal place. Best and second-best values are identified according to each table's comparison scope, including ties at the displayed precision.

D Generated-Answer Evaluation and Diagnostics

D.1 Prompting and Generation Settings

Common instruction. All thirteen models receive the same authenticity instruction shown in Figure 5 together with the input audio. ZS-Gen evaluates the released models without additional task-specific training, using the fixed class-balanced subsets described in Section A.3. The utterance identifiers and instruction are shared across models.

Model interfaces. For open-weight models, audio and text inputs are formatted using the released chat templates and processed through the native autoregressive generation interfaces. Closed-source models receive the same instruction and audio through their API interfaces. Audio preprocessing and duration handling follow the model-specific native pipelines described in Section C.1.

Decoding settings. The open-weight models use greedy token selection. Eight models use `do_sample=False`; Kimi-Audio-7B uses `text_temperature=0.0`, and MiMo-Audio-7B uses a model-specific generation wrapper that enforces greedy decoding. Closed-source generation is executed through the provider APIs. These settings govern the evaluated models' response generation; the semantic judge is run separately with its own decoding configuration.

Response extraction. Only newly generated answer content is retained for evaluation. For interfaces returning both input and generated tokens, the input-token prefix is removed before decoding the continuation. If a model restates the question within its generated answer, that text remains part of the response. Thus, removing the original input prefix does not remove question echoes produced by the model itself. Semantic adjudication and the diagnostic parsers operate on the same saved response for each model–utterance pair.

D.2 Semantic Adjudication and Generated-Answer Metrics

Common semantic judge. We use Qwen3-Omni-30B-A3B-Thinking to interpret the saved responses of all thirteen evaluated models. The judge receives only the generated text and sample identifier, without access to the input audio or reference authenticity label. It assigns one of three

outcomes: `real`, `fake`, or `refuse`. The first two indicate an expressed authenticity decision. The `refuse` category includes explicit refusals and responses without an identifiable authenticity decision, including uncertainty, irrelevant text, or question echoes. Merely mentioning `real` or `fake` within a refusal or an echoed question does not establish a decision.

Adjudication settings. The same adjudication procedure is applied across models and datasets. The judge uses greedy decoding with `do_sample=False`, thinking disabled through `enable_thinking=False`, and a maximum of eight new tokens. Adjudicated `refuse` outcomes are treated as abstentions when computing generated-answer metrics.

Generated-answer accuracy and validity. Let N be the number of evaluated utterances, with reference labels $y_i \in \{0, 1\}$ denoting bona fide and spoofed speech, respectively. The adjudicated prediction is $\hat{y}_i \in \{0, 1, \perp\}$, where $\perp$ denotes an abstention. Generated-answer accuracy (Acc) and Valid Rate are

$$\text{Acc} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i], \qquad \text{Valid} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i \in \{0, 1\}]. \tag{14}$$

Acc counts refusals and non-decisions as errors. Valid Rate measures the fraction of responses containing an adjudicated authenticity decision, irrespective of whether that decision is correct. Both metrics use all evaluated utterances as the denominator.

Macro-F1 with abstentions. We compute Macro-F1 over the bona fide and spoofed classes:

$$\text{Macro-F1}_{\text{Gen}} = \frac{100}{2} \sum_{c \in \{0,1\}} \frac{2\,\text{TP}_c}{2\,\text{TP}_c + \text{FP}_c + \text{FN}_c}. \tag{15}$$

The counts are computed from all N utterances. An abstention contributes a false negative to its reference class and no false positive to either class. The macro average includes only the two authenticity classes; abstention is not treated as a third class. Thus, abstained examples remain included in both Acc and Macro-F1 rather than being removed before evaluation.

D.3 Complete ZS-Gen Results

Table 5 reports the complete generated-answer results for ten open-weight and three closed-source models across all five datasets. Each model is evaluated on the same 1,000 class-balanced utterances per dataset using the common semantic adjudication procedure in Section D.2. We report generated-answer accuracy, Macro-F1, and Valid Rate jointly to distinguish incorrect authenticity decisions from refusals and non-decisions.

D.4 Sensitivity to Answer Interpretation

We assess sensitivity to answer interpretation by applying naive first-match parsing and semantic adjudication to the same saved responses. The evaluated utterances, generated text, reference labels, and metric definitions remain unchanged. Table 6 reports the resulting metrics for all thirteen models.

Interpretation procedures. The Naive parser lowercases each response and selects the first substring occurrence of `real` or `fake`; responses containing neither label are treated as abstentions. This procedure can extract labels appearing within refusals or repetitions of the question. Judge denotes the common semantic adjudication procedure in Section D.2, which identifies whether the response expresses an authenticity decision. The primary ZS-Gen results use Judge throughout.

Refusals and malformed responses. GPT-audio exhibits the largest sensitivity. Naive parsing reports Valid Rates of 92.3%–97.9% and Acc of 44.9%–48.7%, because label mentions inside refusal templates can be accepted as decisions. Under semantic adjudication, both metrics are at most 0.2%. Step-Audio-2-mini also changes substantially: Naive parsing yields 100.0% validity and 50.0% Acc on every dataset, whereas Judge yields Valid Rates of 57.0%–85.3% and Acc of 32.4%–41.7%. Thus, apparent label extractability can substantially exceed the frequency of identifiable authenticity decisions.

Table 5: Complete ZS-Gen results under common semantic adjudication. Acc, F1, and Valid denote generated-answer accuracy, Macro-F1, and Valid Rate (%), respectively. For Acc and F1, bold marks the best result within the open-weight and closed-source groups separately; blue and orange shading mark the best and second-best distinct values across all models for each dataset and metric, including ties at the displayed precision.

	19LA			21LA			ITW			AT-ADD			ASV5		
Model	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid
Open-weight models															
Qwen2.5-Omni-3B	50.6	34.7	100.0	50.0	33.3	100.0	51.1	35.7	100.0	50.4	34.4	100.0	52.0	37.6	100.0
Step-Audio-2-mini	40.2	37.3	66.5	35.6	34.9	57.0	32.4	34.6	62.7	36.6	36.2	74.3	41.7	40.2	85.3
Voxtral-Mini-3B	33.7	29.3	69.1	30.2	26.8	66.9	37.5	36.9	55.9	43.1	36.9	83.3	46.9	32.9	96.0
Phi-4-multimodal	46.2	32.9	90.4	46.8	33.2	91.1	48.9	34.0	95.1	49.9	36.3	95.6	46.0	37.1	89.4
Audio-Flamingo-Next	50.0	33.3	100.0	50.0	33.3	100.0	50.0	33.3	100.0	49.4	34.1	100.0	50.8	35.1	100.0
MOSS-Audio-4B	55.6	44.7	100.0	56.2	45.9	100.0	53.1	40.3	100.0	**52.9**	40.7	100.0	75.4	74.4	100.0
Kimi-Audio-7B	56.3	55.4	97.7	60.2	59.2	98.2	**68.6**	**67.0**	98.8	49.8	47.1	99.2	54.0	53.0	99.3
MiMo-Audio-7B	50.0	33.4	99.9	50.0	33.4	99.9	50.3	34.0	100.0	46.7	33.5	92.8	52.5	39.1	99.7
Gemma-4-12B-it	56.7	51.3	98.9	52.2	51.8	98.0	57.5	58.5	96.7	47.4	**47.7**	98.9	61.7	60.9	99.8
SALMONN-2*	**91.2**	**91.3**	99.7	**78.2**	**78.2**	99.9	56.8	47.9	99.7	33.9	42.1	61.4	**86.1**	**86.2**	99.7
Closed-source models															
Qwen3.5-Omni-Flash	49.1	46.6	99.9	48.1	45.1	99.9	46.9	44.3	99.8	47.9	45.7	99.9	51.4	49.6	99.9
GPT-audio	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.1
Gemini-3.8-Flash	**58.0**	**54.3**	99.7	**56.1**	**52.4**	99.9	**72.5**	**72.5**	99.5	**53.7**	**50.5**	99.3	**62.9**	**58.8**	99.2

Refusals and non-decisions count as errors in Acc and false negatives for the true class in F1. Valid measures the fraction of adjudicated real/fake decisions, irrespective of correctness. * SALMONN-2 has documented HoliAntiSpoof training exposure.

Model- and dataset-dependent sensitivity. Voxtral's ITW results change from 49.8% to 37.5% Acc, from 54.8% to 36.9% Macro-F1, and from 71.5% to 55.9% validity. Phi-4 also shows a substantial validity difference on ASV5, decreasing from 99.7% under Naive parsing to 89.4% under Judge. In contrast, Qwen2.5-Omni-3B, Audio-Flamingo-Next, MOSS-Audio-4B, SALMONN-2, and Qwen3.5-Omni-Flash have identical reported metrics under both procedures at the displayed precision. For Kimi, MiMo, Gemma, and Gemini, the maximum absolute difference across the reported datasets and metrics is 0.8 percentage points.

D.5 Qualitative Response Examples

Figures 6–8 present native responses to six selected utterances spanning the evaluation suite. Each panel compares ten open-weight and three closed-source models on the same utterance, with its identifier and ground-truth label shown above the responses. These examples illustrate characteristic response forms rather than their prevalence; aggregate behavior is reported in Table 5.

Direct decisions and benchmark-dependent bias. Figure 6(a) shows a fake 19LA utterance for which nearly all direct authenticity answers are real, while GPT-audio declines to make a decision. Conversely, the bona fide ASV5 utterance in Figure 6(b) receives several fake decisions. These examples illustrate that native answer tendencies can change substantially across benchmark conditions.

Cross-domain disagreement and non-decisions. For the fake 21LA utterance in Figure 7(c), most direct answers again favor real, whereas GPT-audio explicitly states that it lacks enough information to decide. Figure 7(d) shows stronger disagreement on an ITW spoof, with direct real and fake decisions both appearing across models. The AT-ADD example in Figure 8(e) further shows ASR-like content responses from MiMo and SALMONN-2 rather than authenticity decisions, alongside direct decisions and a GPT-audio refusal. These cases motivate semantic adjudication rather than interpreting every occurrence of *real* or *fake* as a valid decision.

Explanations accompanying incorrect decisions. For the bona fide 21LA utterance in Figure 8(f), Gemini predicts fake and attributes its decision to "robotic cadence, unnatural vowel tran-

Table 6: Sensitivity of generated-answer metrics to answer interpretation. Naive and Judge operate on the same saved responses for 1,000 class-balanced utterances per dataset. Acc, F1, and Valid denote Acc, Macro-F1, and Valid Rate (%), respectively. Both indicates identical reported metrics under the two procedures across all five datasets at the displayed precision.

		19LA			21LA			ITW			AT-ADD			ASV5		
Model	Parser	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid	Acc	F1	Valid
Open-weight models																
Qwen2.5-Omni-3B	Both	50.6	34.7	100.0	50.0	33.3	100.0	51.1	35.7	100.0	50.4	34.4	100.0	52.0	37.6	100.0
Step-Audio-2-mini	Naive	50.0	33.3	100.0	50.0	33.3	100.0	50.0	33.3	100.0	50.0	33.3	100.0	50.0	33.3	100.0
	Judge	40.2	37.3	66.5	35.6	34.9	57.0	32.4	34.6	62.7	36.6	36.2	74.3	41.7	40.2	85.3
Voxtral-Mini-3B	Naive	35.3	32.0	74.1	31.4	28.9	71.6	49.8	54.8	71.5	44.7	39.3	87.2	47.7	34.4	97.3
	Judge	33.7	29.3	69.1	30.2	26.8	66.9	37.5	36.9	55.9	43.1	36.9	83.3	46.9	32.9	96.0
Phi-4-multimodal	Naive	49.1	33.0	98.6	49.5	33.2	98.9	49.8	33.7	99.3	50.5	35.9	98.9	51.2	38.4	99.7
	Judge	46.2	32.9	90.4	46.8	33.2	91.1	48.9	34.0	95.1	49.9	36.3	95.6	46.0	37.1	89.4
Audio-Flamingo-Next	Both	50.0	33.3	100.0	50.0	33.3	100.0	50.0	33.3	100.0	49.4	34.1	100.0	50.8	35.1	100.0
MOSS-Audio-4B	Both	55.6	44.7	100.0	56.2	45.9	100.0	53.1	40.3	100.0	52.9	40.7	100.0	75.4	74.4	100.0
Kimi-Audio-7B	Naive	56.6	55.5	98.5	60.3	59.1	98.9	68.6	67.0	98.9	49.8	47.1	99.3	54.1	53.0	99.4
	Judge	56.3	55.4	97.7	60.2	59.2	98.2	68.6	67.0	98.8	49.8	47.1	99.2	54.0	53.0	99.3
MiMo-Audio-7B	Naive	50.0	33.3	100.0	50.0	33.3	100.0	50.3	34.0	100.0	46.6	33.3	92.7	52.8	39.3	100.0
	Judge	50.0	33.4	99.9	50.0	33.4	99.9	50.3	34.0	100.0	46.7	33.5	92.8	52.5	39.1	99.7
Gemma-4-12B-it	Naive	56.9	51.4	99.1	52.3	51.8	98.1	57.5	58.4	96.8	47.5	47.7	99.1	61.7	60.9	99.8
	Judge	56.7	51.3	98.9	52.2	51.8	98.0	57.5	58.5	96.7	47.4	47.7	98.9	61.7	60.9	99.8
SALMONN-2*	Both	91.2	91.3	99.7	78.2	78.2	99.9	56.8	47.9	99.7	33.9	42.1	61.4	86.1	86.2	99.7
Closed-source models																
Qwen3.5-Omni-Flash	Both	49.1	46.6	99.9	48.1	45.1	99.9	46.9	44.3	99.8	47.9	45.7	99.9	51.4	49.6	99.9
GPT-audio	Naive	44.9	31.6	92.3	47.4	33.0	94.4	47.6	33.0	95.8	48.7	33.2	96.8	48.7	32.9	97.9
	Judge	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.1
Gemini-3.8-Flash	Naive	58.0	54.3	99.7	56.1	52.4	99.9	72.6	72.5	99.7	53.8	50.7	99.4	62.9	58.8	99.2
	Judge	58.0	54.3	99.7	56.1	52.4	99.9	72.5	72.5	99.5	53.7	50.5	99.3	62.9	58.8	99.2

Under both procedures, abstentions count as errors in Acc and false negatives for the reference class in F1. Valid measures the fraction of responses assigned a real/fake decision. * SALMONN-2 has documented HoliAntiSpoof training exposure.

sitions, and synthetic artifacts." The explanation accompanies an incorrect authenticity decision and is shown only as part of the native response, not as independently verified acoustic evidence.

E Complete Score-Based Detection Results

E.1 Results under the 19LA-Source Protocol

Table 1 reports the complete score-based results for ten open-weight Audio-LLMs across five evaluation datasets. Audio-Linear, LM-Linear, and Native-SFT use the same 19LA training split, with source-development data used for checkpoint selection where applicable. Evaluation on 19LA measures held-out in-domain performance, while evaluation on 21LA, ITW, AT-ADD, and ASV5 measures transfer without target-domain fitting. ZS-Score provides a native-score baseline without additional task-specific training. All methods are evaluated on the complete evaluation sets; EER uses continuous detection scores, whereas Macro-F1 and accuracy use the fixed real–fake score comparison.

On 19LA, Audio-Linear and LM-Linear each achieve EER below 10% for eight of the ten models. Native-SFT obtains EERs ranging from 0.49% to 3.78% and improves EER, Macro-F1, and accuracy over LM-Linear for every model. Relative to Audio-Linear, Native-SFT achieves lower EER for eight models; Voxtral and Audio-Flamingo-Next retain slightly lower EER with the frozen audio-side readout. These results establish strong in-domain performance under source supervision while preserving differences between the two representation interfaces and native adaptation.

The complete target-domain results reveal changes in method ordering across both datasets and metrics. For example, on ASV5, Voxtral's LM-Linear outperforms its Audio-Linear and Native-SFT in all three metrics, despite Audio-Linear having lower EER than LM-Linear on 19LA. For

(a) 19LA — **Ground truth: fake**

Utterance ID: `LA_E_7268887`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realcontent<中文>Is this audio real or fake?
Voxtral-Mini-3B	fake
Phi-4-multimodal	real
Audio-Flamingo-Next	real
MOSS-Audio-4B	real
Kimi-Audio-7B	real
MiMo-Audio-7B	real
Gemma-4-12B-it	real
SALMONN-2*	real
Closed-source models	
GPT-audio	I'm sorry, but I can't determine whether an audio is real or fake.
Qwen3.5-Omni-Flash	real
Gemini-3.8-Flash	real

(b) ASV5 — **Ground truth: real**

Utterance ID: `E_0003308258`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realpath: code/1
Voxtral-Mini-3B	fake
Phi-4-multimodal	This audio is fake.
Audio-Flamingo-Next	real
MOSS-Audio-4B	fake
Kimi-Audio-7B	real
MiMo-Audio-7B	real
Gemma-4-12B-it	fake
SALMONN-2*	Fake
Closed-source models	
GPT-audio	I'm sorry, but I can't determine whether an audio sample is real or fake.
Qwen3.5-Omni-Flash	fake
Gemini-3.8-Flash	fake

Figure 6: Native responses to two selected utterances. Each panel shows the same utterance across ten open-weight and three closed-source models; the badge gives the ground-truth label. (a) Most direct answers label a fake 19LA utterance as real, while GPT-audio refuses to decide. (b) A bona fide ASV5 utterance receives several fake decisions, illustrating benchmark-dependent answer bias. Responses are reproduced from the saved native outputs. Label mentions inside refusals or echoed questions are not treated as semantic decisions in the primary evaluation. These selected examples illustrate response forms, not their prevalence. SALMONN-2 has documented HoliAntiSpoof training exposure.

(c) 21LA — **Ground truth: fake**

Utterance ID: `LA_E_4355851`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realcontent<中文>Is this audio real or fake?
Voxtral-Mini-3B	fake
Phi-4-multimodal	Real.
Audio-Flamingo-Next	real
MOSS-Audio-4B	real
Kimi-Audio-7B	real
MiMo-Audio-7B	real
Gemma-4-12B-it	real
SALMONN-2*	real
Closed-source models	
GPT-audio	I currently don't have enough information to determine if the audio is real or fake.
Qwen3.5-Omni-Flash	real
Gemini-3.8-Flash	real

(d) ITW — **Ground truth: fake**

Utterance ID: `26484`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realcontent:real
Voxtral-Mini-3B	fake
Phi-4-multimodal	This audio is real.
Audio-Flamingo-Next	real
MOSS-Audio-4B	real
Kimi-Audio-7B	fake
MiMo-Audio-7B	real
Gemma-4-12B-it	fake
SALMONN-2*	Fake
Closed-source models	
GPT-audio	I'm sorry, but I can't determine whether the audio is real or fake.
Qwen3.5-Omni-Flash	fake
Gemini-3.8-Flash	The audio is fake.

Figure 7: Native responses to two spoofed transfer utterances. Each panel compares ten open-weight and three closed-source models. (c) Most direct labels remain real for a fake 21LA utterance, while GPT-audio explicitly declines to decide. (d) The fake ITW utterance produces sharply divided real/fake decisions across models, again accompanied by a GPT-audio refusal. Responses are reproduced from the saved native outputs; refusal text is shown verbatim and is interpreted semantically rather than by first label occurrence. SALMONN-2 has documented HoliAntiSpoof training exposure.

(e) AT-ADD — **Ground truth: real**

Utterance ID: `ATADD_T1_Eval_076577`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realcontent<中文>这是真实的声音。<中文
Voxtral-Mini-3B	fake
Phi-4-multimodal	real
Audio-Flamingo-Next	real
MOSS-Audio-4B	real
Kimi-Audio-7B	real
MiMo-Audio-7B	星球大战的人物有什么？ [*off-task content response*]
Gemma-4-12B-it	real
SALMONN-2*	新仇大战的人物有什么 [*off-task content response*]
Closed-source models	
GPT-audio	I'm sorry, but I can't determine whether an audio is real or fake.
Qwen3.5-Omni-Flash	real
Gemini-3.8-Flash	fake

(f) 21LA — **Ground truth: real**

Utterance ID: `LA_E_8915700`

Qwen2.5-Omni-3B	real
Step-Audio-2-mini	realcontent<中文>这是真实的声音。<中文
Voxtral-Mini-3B	fake
Phi-4-multimodal	Real.
Audio-Flamingo-Next	real
MOSS-Audio-4B	real
Kimi-Audio-7B	real
MiMo-Audio-7B	real
Gemma-4-12B-it	fake
SALMONN-2*	Real
Closed-source models	
GPT-audio	I currently don't have enough information to determine if the audio is real or fake.
Qwen3.5-Omni-Flash	fake
Gemini-3.8-Flash	Based on the robotic cadence, unnatural vowel transitions, and synthetic artifacts, the audio is **fake**.

Figure 8: Native responses to two bona fide transfer utterances. Each panel compares ten open-weight and three closed-source models. (e) The AT-ADD example includes direct decisions, a GPT-audio refusal, and off-topic outputs from MiMo and SALMONN-2. (f) The 21LA example combines disagreement and refusal with a confident but incorrect Gemini fake judgment accompanied by an acoustic-sounding rationale. Responses are reproduced from the saved native outputs. These selected examples illustrate response forms, not their prevalence. SALMONN-2 has documented HoliAntiSpoof training exposure.



Figure 9: Complete audio-encoder depth profiles under 19LA-source supervision. Columns correspond to six Audio-LLMs and rows to EER, accuracy, and Macro-F1. Colors distinguish Linear, AASIST, and Depth-SFT, while solid and dashed lines denote 19LA and ITW evaluation, respectively. Gray regions indicate blocks beyond Phi-4's 24-block encoder. All metrics are percentages.

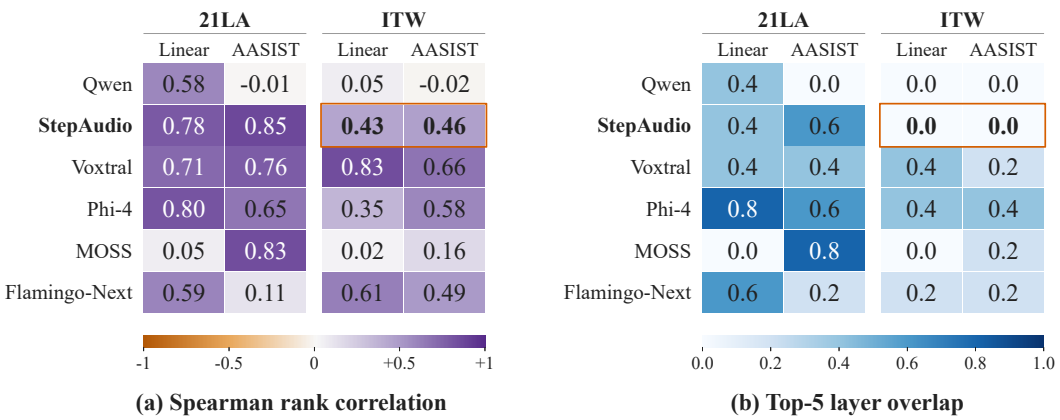
SALMONN-2 on ASV5, Native-SFT reduces EER from the ZS-Score baseline's 14.24% to 11.98%, while Macro-F1 decreases from 81.57% to 55.73% at the fixed threshold. Reporting all three metrics therefore preserves the distinction between score discrimination and classification at the prescribed operating point. SALMONN-2 is marked throughout for its documented HoliAntiSpoof training exposure.

E.2 Results under the AT-ADD and ASV5 Source Protocols

Table 2 reports the additional source-protocol results for Qwen2.5-Omni-3B, Kimi-Audio-7B, and MOSS-Audio-4B. Audio-Linear, LM-Linear, and Native-SFT are trained separately on the AT-ADD or ASV5 training split and evaluated on the corresponding in-domain evaluation set and the complete ITW set of 31,779 utterances. Source-development data are used for checkpoint selection where applicable, and the resulting detectors are transferred without target-domain fitting or threshold adjustment. ZS-Score provides the same training-free reference across protocols; its ITW results are therefore identical in both source groups.

Within the source domain, Audio-Linear achieves lower EER than LM-Linear for all three models under both protocols. Native-SFT improves EER, Macro-F1, and accuracy over LM-Linear in all six model–source combinations. Its comparison with Audio-Linear remains model-dependent: Kimi and MOSS achieve lower EER with Native-SFT under both sources, whereas Qwen retains lower EER with Audio-Linear. Thus, the additional protocols extend the in-domain advantage of native adaptation over frozen language-side readout while retaining exceptions relative to audio-side readout.

Evaluation on the shared ITW target isolates changes in observed performance across training sources. For Qwen, Audio-Linear has lower EER than LM-Linear under AT-ADD-source training (17.86% versus 39.30%), whereas their ordering reverses under ASV5-source training (33.57% versus 31.71%). The relationship between metrics also changes with the source. Kimi's AT-ADD-trained Native-SFT achieves lower ITW EER than both frozen readouts but lower Macro-F1 and accuracy. Under ASV5-source training, Native-SFT instead outperforms both readouts in all three metrics, reaching 11.54% EER, 86.06% Macro-F1, and 87.52% accuracy. These results show that both readout rankings and the fixed-threshold benefits of adaptation depend on the source protocol.



(a) Spearman rank correlation **(b) Top-5 layer overlap**

Figure 10: Agreement between frozen-readout decodability and Depth-SFT utility across encoder depth, measured using EER. (a) Spearman rank correlation between complete Linear or AASIST depth rankings and the corresponding Depth-SFT ranking. (b) Top-5 overlap between the lowest-EER layers selected by each frozen readout and by Depth-SFT. Higher values indicate stronger agreement. Outlined cells correspond to the StepAudio–ITW case discussed in Sec. 5.4, where positive global rank correlation coexists with zero Top-5 overlap. All available encoder blocks are included (32 per model; 24 for Phi-4).

F Complete Encoder-Depth Results

This section provides the complete encoder-depth analysis underlying Section 5.4. All experiments use 19LA-source supervision. For each model and detection procedure, preprocessing and training settings are held fixed across encoder depths, and each depth is trained independently. Linear and AASIST operate on frozen encoder representations, whereas Depth-SFT adapts the selected encoder-to-language connection together with language-side modules. Depth-SFT comparisons use its own final-block configuration as the within-branch reference.

Figure 9 shows the complete depth profiles for all six models across EER, accuracy, and Macro-F1 on 19LA and ITW. Unlike the selected cases highlighted in the main text, this figure retains the full model and metric coverage and therefore also shows cases in which depth preferences differ across metrics or do not exhibit the same regional pattern. The corresponding numerical results, including 21LA evaluation, are reported in the model-specific tables below.

Agreement between frozen-readout and adaptation depth preferences. Beyond the complete depth profiles in Fig. 9, we quantify how closely frozen-readout rankings align with Depth-SFT across all six models. For each model and target dataset, we compute the Spearman rank correlation between the EER-based depth rankings of Linear or AASIST and the corresponding Depth-SFT ranking, together with the Top-5 overlap between their lowest-EER layers. Figure 10 shows that agreement varies substantially across models, readouts, and target domains. Notably, moderate positive global correlation does not necessarily imply agreement among the highest-ranked layers: for StepAudio on ITW, Linear and AASIST have positive Spearman correlations with Depth-SFT but share no Top-5 layers with it. These results complement the regional comparisons in Sec. 5.4 and further indicate that frozen-readout depth preferences are not a stable proxy for adaptation utility.

The complete profiles reinforce three observations discussed in the main text. First, favorable depths are often distributed over contiguous encoder regions rather than concentrated at a single layer. Second, the depth regions favored by frozen Linear or AASIST readouts need not coincide with those favored by Depth-SFT. Third, depth rankings may vary across evaluation domains and metrics. These observations are descriptive of the reported sweeps; each depth uses a single random seed, and target-domain results are not used to select a deployment layer.

F.1 QWEN2.5-OMNI-3B

Tables 7–9 report the complete depth profiles for the three procedures. Every Depth-SFT connection from L08 through L23 has lower EER than L31 on 19LA, 21LA, and ITW. AASIST has its lowest observed ITW EER at L31 (38.07%), while L22 achieves higher fixed-threshold Macro-F1 (62.6% versus 35.5%).

Table 7: Qwen2.5-Omni-3B encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.20	**17.97**	**72.6**	**91.45**	15.01	77.7	90.02	29.97	27.1	37.18
L01	0.04	**18.89**	**73.7**	**91.53**	14.80	78.4	90.79	26.18	27.1	37.18
L02	**0.00**	19.86	70.2	90.70	14.46	80.9	92.17	**21.72**	27.1	37.18
L03	**0.00**	23.18	61.4	90.48	13.19	81.9	92.25	27.34	27.1	37.18
L04	**0.00**	27.63	60.7	90.10	13.12	82.8	92.72	25.97	27.1	37.18
L05	**0.00**	35.66	57.5	88.30	13.33	83.2	93.28	35.61	27.1	37.18
L06	**0.00**	31.56	57.6	89.39	12.69	85.4	94.39	**24.01**	27.1	37.18
L07	**0.00**	27.55	61.6	89.35	12.96	84.4	93.67	29.24	**29.3**	**38.43**
L08	**0.00**	29.79	58.3	89.33	12.48	85.3	94.06	28.46	27.1	37.21
L09	**0.00**	29.07	53.4	89.74	12.45	87.6	95.34	32.85	27.2	37.21
L10	**0.00**	24.47	54.5	90.16	12.29	89.0	95.97	30.15	27.1	37.18
L11	**0.00**	27.11	50.9	89.94	12.13	90.5	96.63	27.70	27.1	37.18
L12	**0.00**	29.52	49.8	89.85	12.01	91.5	96.94	39.45	27.1	37.18
L13	**0.00**	29.93	51.5	89.76	12.10	91.8	97.03	32.64	**27.5**	**37.36**
L14	**0.00**	36.68	49.4	89.73	11.93	**92.8**	**97.54**	41.47	27.2	37.22
L15	**0.00**	34.40	49.1	89.77	12.45	92.6	97.38	40.33	27.1	37.18
L16	**0.00**	36.44	48.6	89.71	11.93	**92.7**	**97.51**	37.33	27.1	37.18
L17	**0.00**	27.53	51.4	89.96	**11.53**	**92.7**	97.42	33.29	27.1	37.18
L18	**0.00**	33.17	48.0	89.70	11.77	91.6	97.01	33.62	27.1	37.18
L19	**0.00**	28.59	49.3	89.80	11.69	89.7	96.65	37.62	27.3	37.27
L20	**0.00**	38.11	47.7	89.31	11.97	89.4	96.53	37.84	27.1	37.18
L21	0.05	38.30	47.8	89.59	12.78	85.6	95.60	40.64	27.1	37.18
L22	0.04	31.99	47.6	89.67	14.20	81.6	94.78	46.73	27.1	37.18
L23	**0.00**	26.73	47.4	89.69	15.81	78.4	94.15	46.59	27.1	37.18
L24	**0.00**	26.93	47.3	89.68	14.65	82.7	95.05	67.97	27.1	37.18
L25	**0.00**	19.25	47.5	89.70	**11.64**	84.8	95.50	39.27	27.1	37.18
L26	**0.00**	22.01	47.3	89.68	12.85	84.0	95.31	54.50	27.1	37.18
L27	0.04	32.36	47.7	89.69	13.82	83.0	95.10	48.19	27.1	37.18
L28	**0.00**	32.01	47.9	89.69	14.57	82.5	94.98	50.52	27.1	37.18
L29	**0.00**	28.06	52.2	89.61	13.35	82.0	94.89	40.92	27.1	37.18
L30	**0.00**	29.89	50.8	89.64	14.02	79.3	94.32	39.23	27.1	37.18
L31	**0.03**	33.26	49.1	89.72	14.87	79.0	94.27	45.72	27.1	37.18

Table 8: Qwen2.5-Omni-3B encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	**1.18**	13.09	66.1	77.22	45.11	57.1	84.16	56.75	30.5	32.93
L01	**5.37**	12.72	77.8	89.60	37.72	58.5	84.60	55.06	33.7	35.56
L02	14.37	15.12	70.0	82.17	44.24	56.7	83.71	60.00	26.4	29.55
L03	12.28	12.40	77.0	92.79	39.04	58.2	86.47	57.11	27.8	30.94
L04	15.77	12.07	81.0	93.16	39.17	54.2	72.48	54.20	38.8	39.23
L05	13.98	11.55	81.8	93.94	40.17	49.8	63.52	63.19	31.4	31.68
L06	23.90	14.32	77.0	90.49	43.51	42.5	50.88	58.43	38.1	38.06
L07	21.26	15.91	72.4	90.38	41.92	53.8	73.08	55.52	34.7	36.17

Continued on next page

Table 8 (continued): Qwen2.5-Omni-3B, AASIST readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L08	15.94	13.71	71.9	83.89	36.32	52.6	66.59	60.10	35.4	35.60
L09	20.17	13.16	76.7	88.95	37.45	53.6	69.18	56.97	38.1	38.24
L10	19.47	6.95	77.2	87.62	40.19	55.9	77.91	50.68	48.0	48.63
L11	21.47	9.27	81.1	91.03	35.44	59.8	85.71	53.82	42.7	42.68
L12	20.08	8.50	85.8	94.40	43.28	54.4	85.21	41.51	51.6	51.84
L13	10.37	5.63	83.5	92.09	34.63	40.5	45.56	43.15	55.9	56.86
L14	8.00	**5.25**	85.3	95.50	27.36	**66.2**	85.09	42.18	49.0	49.35
L15	9.14	**5.21**	**91.1**	**96.51**	32.80	53.1	65.70	40.96	**59.1**	**61.86**
L16	10.16	7.34	86.3	94.16	36.90	43.5	50.71	44.39	55.6	57.58
L17	9.93	8.36	78.9	89.10	39.90	48.8	61.16	43.76	54.3	54.64
L18	11.26	8.54	77.4	87.97	37.36	42.9	50.04	44.61	54.7	58.25
L19	12.87	5.28	**90.8**	**96.67**	34.87	46.1	54.67	42.29	57.4	59.34
L20	16.21	8.36	50.0	89.96	36.54	60.9	85.25	41.10	29.8	38.61
L21	26.22	11.15	47.5	89.70	35.20	56.0	87.39	40.56	31.7	39.53
L22	24.18	10.59	77.4	93.55	**26.24**	54.8	65.12	**38.25**	**62.6**	**65.32**
L23	84.21	65.07	47.4	89.65	64.97	47.3	**89.83**	57.52	27.2	37.23
L24	7.57	20.59	67.2	89.60	34.81	25.0	25.19	45.62	53.4	56.03
L25	10.91	12.43	51.8	90.14	32.94	55.1	69.36	42.25	50.2	50.86
L26	10.99	37.58	47.3	89.68	28.55	**61.4**	88.15	39.18	35.6	41.96
L27	40.38	53.15	47.6	89.70	28.05	60.1	**89.77**	47.29	31.5	39.51
L28	18.80	20.49	50.4	89.99	26.92	52.6	62.59	40.19	57.6	58.03
L29	9.15	9.79	66.9	92.03	**22.30**	59.9	72.21	39.94	45.9	47.57
L30	32.89	8.29	47.3	89.68	48.11	51.4	86.33	41.13	27.6	37.42
L31	34.53	17.42	47.9	89.73	46.83	50.7	71.88	**38.07**	35.5	41.48

Table 9: Qwen2.5-Omni-3B encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.32	7.34	79.1	88.88	17.76	72.5	86.99	52.31	29.1	36.75
L01	0.27	6.26	77.2	87.37	14.00	75.5	87.70	48.48	30.2	36.66
L02	0.20	4.02	82.4	91.18	13.98	82.0	93.22	42.36	32.7	38.65
L03	0.35	2.31	88.0	94.56	14.94	80.0	93.43	44.03	29.1	37.55
L04	0.13	1.88	86.9	93.97	14.30	81.5	93.79	38.08	29.1	38.08
L05	0.23	1.31	93.2	97.18	15.56	81.0	94.41	31.66	31.4	39.56
L06	0.16	0.96	95.4	98.15	16.80	78.8	94.07	26.78	36.3	42.56
L07	0.20	0.79	95.2	98.10	11.04	89.1	96.43	22.36	49.3	51.55
L08	0.16	0.45	98.0	99.22	11.84	85.9	95.74	23.72	27.9	37.61
L09	0.08	0.47	98.9	99.58	8.46	83.5	95.23	25.14	40.2	45.20
L10	0.04	0.14	**99.5**	**99.80**	7.40	91.0	97.11	22.31	47.3	50.14
L11	0.04	0.30	98.4	99.38	7.07	93.2	97.67	17.56	66.1	66.21
L12	**0.00**	0.18	**99.3**	**99.76**	7.72	93.6	97.80	14.51	69.3	69.38
L13	**0.00**	**0.10**	98.6	99.46	**6.42**	93.3	97.53	20.02	**74.0**	**74.33**
L14	**0.00**	0.22	**99.3**	99.75	8.11	91.8	97.33	**12.24**	**74.9**	**74.96**
L15	**0.00**	0.14	97.9	99.20	7.21	**94.7**	**98.16**	16.94	67.8	67.82
L16	**0.00**	**0.11**	97.5	99.04	**6.38**	**94.0**	**97.85**	30.42	55.1	55.31
L17	**0.00**	0.27	95.7	98.31	7.19	91.7	97.21	**12.95**	68.5	68.57
L18	0.04	0.51	94.7	97.86	7.08	90.1	96.72	13.54	64.2	64.50
L19	**0.01**	0.31	95.0	97.98	10.18	87.1	95.87	33.28	45.5	47.84
L20	**0.00**	0.18	98.1	99.28	6.83	90.2	96.79	29.21	47.5	49.77
L21	**0.01**	0.90	91.4	96.31	12.96	79.5	93.67	39.42	35.3	41.80
L22	**0.01**	0.42	96.2	98.50	11.93	80.3	94.21	32.69	32.3	40.11
L23	0.03	0.30	96.6	98.66	9.69	84.2	95.06	37.55	31.7	39.82
L24	0.03	0.29	96.3	98.53	12.38	82.3	94.54	30.27	28.7	38.08

Continued on next page

Table 9 (continued): Qwen2.5-Omni-3B, Depth-SFT

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L25	0.09	0.35	99.1	99.68	25.12	69.3	92.37	43.90	27.2	37.21
L26	0.05	0.65	95.4	98.16	26.21	69.2	92.02	40.68	27.1	37.18
L27	0.08	0.88	82.4	91.17	19.85	73.1	91.03	41.75	28.4	37.79
L28	0.08	1.48	87.3	94.16	19.80	73.1	91.67	33.30	27.2	37.23
L29	0.09	0.71	96.4	98.60	17.29	76.8	93.50	39.83	27.2	37.22
L30	0.31	0.71	97.6	99.08	15.69	77.1	93.62	43.74	27.2	37.24
L31	0.32	0.95	94.4	97.73	21.78	75.9	93.17	46.15	27.5	37.38

F.2 STEP-AUDIO-2-MINI

Tables 10–12 report the complete depth profiles for the three procedures. Linear and AASIST favor L12–L19 over L27–L31 on ITW: their EER ranges are 14.64%–28.03% versus 31.99%–39.23%, and 14.34%–21.77% versus 27.16%–34.99%, respectively. Depth-SFT reverses this ordering, with 18.35%–21.01% at L27–L31 versus 21.66%–34.49% at L12–L19.

Table 10: Step-Audio-2-mini encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.31	6.13	63.8	74.20	19.55	69.7	84.12	59.37	27.0	35.52
L01	0.20	4.31	69.2	80.08	18.55	70.7	84.51	57.06	28.5	36.52
L02	0.12	4.68	70.2	81.09	18.46	70.2	83.87	53.41	28.6	36.43
L03	0.12	3.55	67.7	78.54	15.80	66.8	79.48	52.67	27.9	35.59
L04	0.16	3.07	70.4	81.30	14.32	73.0	85.53	53.49	27.7	36.57
L05	0.24	2.60	68.3	79.19	13.74	73.1	85.44	54.07	29.4	36.60
L06	0.16	3.19	68.4	79.23	13.73	74.3	86.55	50.60	29.6	37.37
L07	**0.08**	1.74	76.4	86.73	11.08	83.2	92.77	38.85	32.7	39.60
L08	0.20	2.39	71.5	82.37	12.02	71.1	83.24	45.43	31.0	38.70
L09	**0.04**	1.32	76.8	87.07	9.78	81.3	91.29	33.29	39.0	43.75
L10	0.12	1.41	81.5	90.55	10.41	86.7	94.85	29.79	34.7	41.51
L11	0.16	0.86	80.4	89.76	10.26	84.6	93.55	31.13	36.9	42.72
L12	0.12	0.63	86.8	93.87	9.70	87.3	95.07	28.03	40.9	45.43
L13	0.11	0.60	88.3	94.71	11.07	88.6	95.84	**14.64**	**74.6**	**74.67**
L14	**0.04**	**0.52**	88.4	94.75	10.55	88.6	95.89	**16.26**	**69.4**	**69.46**
L15	**0.08**	0.53	92.1	96.69	10.78	90.2	96.58	22.27	55.7	56.69
L16	**0.04**	**0.50**	92.9	97.05	**8.50**	**91.3**	**96.94**	24.90	51.2	52.73
L17	0.12	0.60	92.5	96.84	9.83	**90.9**	**96.82**	25.82	50.0	52.02
L18	0.12	0.73	93.1	97.15	9.64	**90.9**	**96.82**	24.45	44.3	47.86
L19	0.12	0.69	94.1	97.61	10.10	89.9	96.61	23.05	36.7	42.82
L20	0.16	1.09	87.4	94.22	10.87	88.2	95.95	30.40	35.2	41.58
L21	0.24	1.43	91.3	96.32	**9.02**	89.1	96.10	28.05	53.8	54.94
L22	**0.08**	1.05	92.1	96.67	10.90	86.6	95.51	29.73	30.1	38.51
L23	0.20	1.42	90.9	96.08	15.57	81.5	94.04	32.94	37.8	43.03
L24	0.20	1.94	89.0	95.11	14.54	82.3	94.02	33.89	47.4	49.35
L25	0.11	1.73	90.4	95.85	13.70	82.7	94.33	32.30	35.7	41.93
L26	0.35	1.77	93.4	97.29	13.85	84.3	94.98	28.44	47.6	50.11
L27	0.24	2.03	92.6	96.95	14.15	84.8	95.04	32.95	37.7	43.17
L28	0.39	2.16	90.8	96.06	16.80	83.3	94.57	38.49	33.1	40.02
L29	0.39	1.93	93.2	97.24	16.44	84.0	94.96	39.23	43.9	46.42
L30	0.47	2.19	**95.0**	**98.05**	16.77	83.5	94.96	36.48	33.9	40.86
L31	0.71	2.34	**94.9**	**97.98**	15.75	84.7	95.20	31.99	53.4	54.30

Table 11: Step-Audio-2-mini encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.59	9.40	68.6	79.61	32.28	58.8	85.89	67.14	25.0	32.81
L01	0.43	4.91	66.9	77.70	25.28	65.1	85.62	60.40	27.0	36.43
L02	0.16	4.76	71.9	82.71	25.33	65.4	83.88	48.05	29.0	37.50
L03	0.12	2.61	77.1	87.25	18.50	63.7	76.01	46.54	32.7	38.43
L04	0.16	2.89	82.9	91.53	18.64	71.8	86.24	45.51	32.3	39.75
L05	0.12	1.35	82.0	90.86	16.24	64.7	76.77	40.55	30.0	38.49
L06	0.16	1.40	92.7	96.97	16.45	74.7	87.94	36.10	31.4	39.59
L07	0.20	0.85	93.1	97.14	10.87	51.8	59.38	36.43	41.2	44.22
L08	0.39	0.92	66.3	77.08	11.44	31.4	32.40	36.20	60.3	60.43
L09	0.35	1.20	85.8	93.31	9.70	40.6	44.03	30.32	56.6	56.96
L10	0.08	0.71	92.7	96.95	9.49	53.7	61.80	22.71	60.9	61.19
L11	**0.07**	**0.37**	97.1	98.88	8.57	75.8	87.05	16.89	70.5	70.49
L12	0.12	0.45	**98.9**	**99.58**	**8.10**	80.2	90.32	**14.34**	71.2	71.21
L13	0.08	**0.31**	97.6	99.09	**8.30**	76.7	87.75	**14.39**	**83.0**	**83.45**
L14	0.08	0.41	**97.8**	**99.16**	9.28	70.4	82.36	18.77	73.8	73.80
L15	**0.04**	0.42	96.5	98.62	8.64	68.5	80.30	17.59	**76.1**	**76.20**
L16	0.27	1.85	95.8	98.36	9.47	50.5	57.36	21.77	52.7	54.45
L17	**0.04**	0.44	91.8	96.50	8.75	70.8	82.68	20.73	74.9	75.13
L18	0.16	0.67	94.7	97.87	9.45	60.1	70.68	19.42	56.3	57.21
L19	0.16	1.09	97.4	98.99	9.46	76.6	87.86	21.42	44.6	48.26
L20	0.24	3.63	77.1	87.26	12.01	**82.0**	**92.45**	25.89	68.1	68.13
L21	0.75	2.24	96.7	98.81	11.64	**82.7**	**92.95**	42.60	30.6	39.17
L22	0.34	2.05	87.5	94.33	13.55	79.1	90.87	31.32	31.7	39.69
L23	0.12	1.25	70.0	80.91	20.38	63.0	75.68	34.01	49.5	51.01
L24	0.20	2.01	91.8	96.53	19.65	72.8	87.32	25.58	49.9	51.74
L25	0.20	1.58	97.4	99.04	16.17	77.8	90.45	32.07	29.1	38.26
L26	0.31	2.93	95.1	98.21	18.75	70.2	84.09	29.88	31.6	39.73
L27	0.24	1.19	96.5	98.66	23.14	50.0	57.35	34.60	39.4	44.08
L28	0.21	1.51	92.8	97.00	26.84	42.7	47.27	32.63	54.9	55.37
L29	0.38	2.50	94.4	97.78	25.21	51.3	60.07	33.40	58.0	58.06
L30	0.67	1.90	95.8	98.37	27.09	50.1	58.34	34.99	43.4	46.72
L31	0.79	3.41	91.1	96.27	26.20	55.5	66.74	27.16	71.4	72.07

Table 12: Step-Audio-2-mini encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	2.70	9.79	78.2	88.79	24.77	69.0	86.60	62.51	25.9	33.93
L01	4.48	9.31	83.8	94.66	19.87	66.1	91.66	58.45	27.1	37.16
L02	5.42	8.62	58.0	90.84	18.29	50.4	90.14	52.34	27.1	37.18
L03	1.29	3.58	92.7	97.10	15.21	78.5	93.63	41.11	33.1	39.58
L04	1.22	3.10	94.7	97.96	18.18	77.0	93.47	57.29	29.4	36.23
L05	0.51	2.23	89.6	95.45	25.34	75.8	92.69	46.76	29.5	38.15
L06	0.59	2.28	95.9	98.43	25.81	69.3	92.26	48.00	27.2	37.22
L07	1.57	3.33	94.7	98.02	14.63	81.5	94.52	39.96	28.6	38.04
L08	2.67	3.34	94.9	98.08	13.38	84.1	95.09	34.98	47.6	49.92
L09	1.14	2.29	81.7	94.78	12.45	60.5	91.33	31.97	27.7	37.52
L10	1.69	3.94	85.5	95.66	13.84	67.0	92.23	40.45	29.6	38.55
L11	**0.35**	**0.83**	**98.3**	**99.35**	11.01	**88.5**	**96.35**	27.45	40.1	45.01
L12	0.39	**1.02**	**98.5**	**99.45**	10.37	**91.2**	**97.10**	23.77	62.2	62.54
L13	0.51	1.54	97.0	98.92	10.58	87.1	95.99	26.84	64.3	64.35
L14	0.78	1.70	96.3	98.68	10.48	84.4	95.39	24.57	58.5	59.12

Continued on next page

Table 12 (continued): Step-Audio-2-mini, Depth-SFT

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L15	**0.24**	1.03	95.3	98.36	10.86	77.6	94.02	21.66	41.9	46.31
L16	0.51	1.51	97.4	99.05	**9.78**	84.4	95.35	27.49	30.9	39.31
L17	0.71	2.58	93.8	97.52	**10.25**	88.1	96.20	34.49	51.7	52.96
L18	0.55	2.72	96.1	98.58	11.99	82.2	94.88	23.70	44.1	47.86
L19	0.50	1.39	97.8	99.21	11.15	84.3	95.38	33.52	57.6	57.60
L20	0.94	3.22	95.4	98.31	14.76	81.3	94.57	30.07	42.4	46.32
L21	0.55	1.39	98.1	99.30	12.96	81.6	94.71	29.66	40.2	44.99
L22	0.54	2.88	90.8	96.09	16.14	82.3	94.23	24.71	69.8	69.95
L23	0.98	2.68	94.9	98.02	17.10	80.0	94.06	22.90	66.2	66.23
L24	0.98	2.42	96.2	98.53	20.38	78.9	93.92	23.37	56.1	56.80
L25	1.30	2.29	92.6	96.96	20.51	79.5	93.59	23.13	67.3	67.35
L26	1.58	2.08	93.6	97.39	21.11	80.1	93.75	21.67	**77.0**	**77.81**
L27	1.22	2.91	91.6	96.50	21.01	77.7	93.23	21.01	**76.5**	**76.99**
L28	2.08	3.55	93.5	97.45	22.26	77.0	93.30	**18.35**	71.3	71.31
L29	1.21	2.92	93.5	97.40	20.22	79.2	93.75	**19.45**	43.9	47.79
L30	1.65	1.92	95.0	98.00	21.22	81.6	94.42	19.86	70.0	70.03
L31	1.62	2.26	86.1	93.51	19.95	80.5	93.04	19.81	73.0	73.03

F.3 VOXTRAL-MINI-3B

Tables 13–15 report the complete depth profiles for the three procedures. Their lowest observed ITW EERs occur at L30 for Linear (18.79%), L29 for AASIST (22.87%), and L28 for Depth-SFT (15.99%). The corresponding 19LA EER minima occur at L15, L10, and L12, respectively, illustrating different depth preferences between in-domain and transfer evaluation.

Table 13: Voxtral-Mini-3B encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	1.84	10.32	67.8	78.93	20.49	68.9	83.82	64.74	25.4	29.19
L01	2.91	6.96	72.3	83.33	18.15	74.3	88.84	64.95	25.8	30.65
L02	1.09	5.74	71.6	82.53	16.77	76.3	90.21	59.63	28.7	32.85
L03	0.99	3.97	76.2	86.60	14.63	76.7	88.88	70.95	22.5	26.03
L04	0.43	2.90	78.1	88.09	13.70	79.1	90.66	62.25	28.6	31.50
L05	0.59	3.32	73.1	83.90	14.19	77.6	89.58	57.85	34.0	36.55
L06	0.63	3.40	74.1	84.81	13.70	78.1	89.90	51.48	37.0	39.29
L07	0.52	3.11	75.5	86.02	11.91	80.8	91.44	57.09	29.4	35.28
L08	0.24	2.06	77.2	87.39	11.74	82.2	92.33	53.52	29.4	36.04
L09	0.35	2.69	73.6	84.31	11.58	71.7	83.79	47.76	42.6	43.04
L10	0.24	1.66	80.9	90.14	10.35	86.5	94.69	41.03	33.7	39.63
L11	0.20	1.02	81.1	90.25	11.60	86.3	94.82	37.19	50.0	50.98
L12	0.15	1.22	80.1	89.53	11.34	86.7	95.00	40.58	39.6	43.45
L13	0.16	1.06	86.3	93.58	10.81	88.7	96.06	31.63	45.6	48.49
L14	**0.12**	0.92	87.4	94.24	**9.36**	**90.7**	**96.72**	24.46	45.9	49.05
L15	**0.12**	**0.72**	87.9	94.49	**9.47**	**90.1**	**96.60**	23.24	53.0	54.61
L16	**0.12**	0.92	89.4	95.31	10.91	86.2	95.59	28.10	50.7	52.51
L17	**0.13**	0.80	89.1	95.18	13.26	85.2	95.37	31.27	55.4	56.03
L18	**0.12**	0.86	85.8	93.32	12.47	86.3	95.44	29.22	**56.3**	56.89
L19	**0.12**	**0.75**	91.5	96.36	11.18	87.0	95.86	28.98	43.1	47.05
L20	0.27	1.20	87.5	94.27	11.23	88.0	95.95	27.69	42.7	46.76
L21	0.24	1.36	92.3	96.78	10.72	87.2	95.88	30.53	33.0	40.58
L22	0.27	1.20	94.1	97.61	12.22	85.7	95.55	26.57	33.2	40.72
L23	0.23	1.22	94.8	97.90	11.79	86.0	95.67	26.01	32.3	40.17
L24	**0.12**	1.13	95.0	97.99	11.63	85.8	95.61	24.69	35.8	42.33

Continued on next page

Table 13 (continued): Voxtral-Mini-3B, Linear readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L25	0.16	1.03	**95.5**	**98.21**	11.40	87.6	96.06	29.60	31.5	39.69
L26	0.16	1.25	**96.1**	**98.47**	10.73	88.7	96.40	26.11	46.5	49.66
L27	0.28	1.45	94.0	97.57	10.67	89.1	96.45	26.48	40.5	45.45
L28	0.20	1.92	91.9	96.58	12.18	86.6	95.70	23.44	48.8	51.41
L29	0.31	1.72	94.8	97.91	11.84	88.1	96.16	21.73	55.8	**57.03**
L30	**0.13**	1.50	93.8	97.50	11.49	88.1	96.14	**18.79**	**59.1**	**59.96**
L31	0.21	1.56	95.3	98.16	11.88	87.2	95.96	**19.24**	42.1	46.55

Table 14: Voxtral-Mini-3B encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.94	7.26	78.2	88.41	26.99	65.1	85.05	51.71	27.1	30.62
L01	4.20	13.73	58.9	68.31	24.12	64.3	80.99	56.85	40.2	40.65
L02	2.24	4.76	89.1	95.35	17.98	66.1	78.99	68.51	29.8	29.93
L03	0.47	2.98	84.1	92.33	24.73	68.3	91.76	40.77	27.1	37.18
L04	0.55	3.02	85.1	92.87	15.95	69.1	81.97	68.80	27.7	29.31
L05	0.43	3.10	91.3	96.35	15.58	72.6	85.43	56.13	41.9	41.86
L06	0.27	2.01	88.4	94.82	15.41	70.6	83.38	57.51	41.8	42.52
L07	0.27	2.69	73.0	83.77	17.21	75.6	90.90	48.42	33.1	38.62
L08	0.20	1.66	85.6	93.17	13.80	75.5	87.62	45.55	51.7	52.02
L09	0.35	2.95	78.2	88.15	14.57	79.4	90.89	47.34	41.4	42.65
L10	**0.16**	**1.14**	78.4	88.32	12.72	**86.0**	**95.42**	36.01	34.4	41.06
L11	0.43	2.95	**93.7**	**97.51**	14.43	80.0	93.52	34.66	51.4	52.86
L12	0.24	2.30	59.4	68.84	12.99	42.1	46.04	33.19	43.6	62.96
L13	0.32	1.77	91.7	96.49	11.01	78.6	89.55	35.58	58.7	58.91
L14	0.35	4.54	92.6	97.30	13.88	69.1	91.88	44.52	43.8	45.65
L15	0.23	1.47	76.3	86.61	11.52	**85.5**	**95.02**	36.95	53.7	54.05
L16	0.24	2.50	79.9	89.43	12.38	82.2	93.24	35.58	61.9	62.45
L17	0.20	3.51	79.9	89.46	13.62	79.2	91.45	32.21	67.7	69.83
L18	0.31	1.96	79.7	89.24	12.58	77.9	89.55	33.57	61.2	61.25
L19	**0.15**	1.70	73.8	84.45	11.74	76.7	88.15	26.47	**70.0**	70.17
L20	0.27	**1.34**	91.4	96.35	**10.63**	77.4	88.50	32.55	59.1	59.17
L21	0.24	1.77	86.3	93.59	**10.29**	68.3	79.71	32.41	66.3	69.10
L22	0.24	1.48	87.2	94.15	13.59	67.2	78.97	36.10	55.7	66.05
L23	0.24	2.62	88.5	94.87	14.97	45.8	51.05	37.56	42.0	62.87
L24	0.27	1.69	85.3	93.00	14.54	54.3	62.80	30.14	61.6	70.44
L25	0.20	1.60	69.3	80.23	13.59	67.2	79.19	26.95	69.5	**74.09**
L26	0.43	2.11	92.6	96.93	14.38	68.8	81.04	**25.78**	**71.4**	**71.73**
L27	0.23	1.58	90.8	96.07	12.33	75.8	87.60	33.84	64.6	65.05
L28	0.31	2.53	92.0	96.66	14.60	69.5	81.67	29.03	62.7	69.90
L29	0.27	2.22	88.6	94.93	17.72	62.2	73.97	**22.87**	61.9	70.51
L30	0.51	3.03	81.8	90.79	14.12	63.5	74.85	29.76	55.1	67.76
L31	0.55	3.48	**93.8**	**97.78**	14.68	73.0	85.62	30.45	67.0	67.49

Table 15: Voxtral-Mini-3B encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.34	8.36	76.6	86.95	17.35	71.5	85.60	58.03	25.6	31.87
L01	0.67	6.67	79.9	89.56	18.78	74.4	88.76	57.58	26.8	35.75
L02	0.23	5.53	78.9	88.72	15.63	75.7	88.88	49.46	27.5	37.06

Continued on next page

Table 15 (continued): Voxtral-Mini-3B, Depth-SFT

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L03	0.27	3.33	85.5	93.17	14.43	81.6	93.29	41.99	27.3	37.18
L04	0.20	2.30	90.5	95.89	13.06	84.9	94.82	50.31	27.7	37.09
L05	0.08	2.24	88.8	95.01	13.59	85.3	95.21	39.67	32.1	39.67
L06	0.12	1.62	93.0	97.11	13.71	84.5	95.25	35.67	29.2	38.36
L07	0.12	1.00	95.5	98.20	14.33	82.4	94.85	45.81	27.3	37.28
L08	0.11	**0.37**	97.1	98.90	8.48	89.5	96.58	30.90	31.9	39.93
L09	0.28	1.13	92.1	96.65	12.13	79.3	93.97	30.71	27.5	37.43
L10	0.24	0.92	97.5	99.04	11.45	85.7	95.71	28.26	28.3	37.86
L11	0.11	0.43	**98.3**	**99.36**	10.31	87.1	96.05	31.43	39.5	44.69
L12	**0.07**	**0.22**	**99.5**	**99.82**	8.73	92.3	97.46	20.76	46.4	49.60
L13	**0.04**	0.56	95.1	98.04	9.52	**92.9**	**97.55**	23.50	58.3	58.99
L14	**0.04**	0.53	96.3	98.55	8.77	91.7	97.26	17.63	48.6	51.21
L15	0.08	0.56	97.5	99.06	**8.31**	**93.1**	**97.69**	18.33	**71.0**	**70.98**
L16	0.08	1.28	87.6	94.36	9.06	90.2	96.38	**16.96**	65.3	65.47
L17	0.08	0.39	98.2	99.33	10.96	83.6	95.20	19.21	45.3	48.81
L18	**0.04**	0.53	97.0	98.83	13.21	82.0	94.73	23.18	38.1	43.84
L19	0.11	0.69	96.5	98.61	10.60	90.0	96.62	24.31	55.4	56.58
L20	**0.04**	0.66	97.3	98.95	**8.47**	92.5	97.42	22.69	41.7	46.26
L21	0.17	1.71	90.3	95.82	9.51	88.7	95.68	25.37	54.6	55.67
L22	0.15	1.90	86.7	93.83	9.78	87.4	95.13	22.17	**74.0**	**74.28**
L23	0.20	1.82	95.6	98.26	10.50	88.4	96.25	24.12	40.2	45.06
L24	0.16	1.51	94.2	97.64	12.14	86.7	95.76	24.73	54.0	55.09
L25	0.19	2.09	93.9	97.52	12.84	83.1	94.81	21.16	43.4	47.25
L26	0.12	1.31	95.8	98.33	11.61	85.4	95.47	23.00	37.7	43.47
L27	0.19	1.70	95.7	98.31	12.29	84.3	95.21	21.36	44.4	48.03
L28	0.12	1.65	96.4	98.61	12.34	82.7	94.88	**15.99**	50.6	52.81
L29	0.20	1.85	93.5	97.36	13.88	83.8	94.91	19.23	36.1	42.54
L30	0.15	2.01	91.0	96.16	14.37	83.4	94.63	19.53	59.7	60.29
L31	0.19	1.62	93.7	97.45	12.52	84.0	95.02	19.46	52.0	53.91

F.4 PHI-4-MULTIMODAL

Tables 16–18 report the complete depth profiles for the three procedures. On ITW, Linear, AASIST, and Depth-SFT have their lowest observed EERs at L07 (22.65%), L08 (17.13%), and L04 (16.89%), respectively. Each improves over its own final-block configuration, whose ITW EERs are 31.12%, 33.05%, and 30.05%.

Table 16: Phi-4-multimodal encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.78	3.22	84.1	92.34	20.47	73.9	88.86	37.13	58.3	58.40
L01	0.51	2.74	83.4	91.83	13.17	75.5	87.52	38.54	55.3	55.36
L02	0.27	**1.07**	93.4	97.30	12.58	83.8	93.42	**22.94**	**66.8**	**66.79**
L03	**0.16**	1.51	92.4	96.81	11.01	86.2	94.76	27.93	62.9	62.98
L04	0.27	1.48	91.2	96.23	**9.50**	89.4	96.04	27.96	**68.4**	**68.57**
L05	0.27	**1.02**	**94.9**	**97.97**	**8.94**	89.2	95.73	24.36	65.0	65.05
L06	**0.08**	1.82	90.0	95.66	9.51	88.7	95.67	26.71	60.7	60.91
L07	0.43	1.58	93.6	97.39	10.28	**89.9**	**96.33**	**22.65**	57.8	58.53
L08	0.17	1.10	**96.7**	**98.74**	10.47	**90.5**	**96.58**	25.86	43.0	46.87
L09	0.35	1.40	94.3	97.70	10.70	87.9	95.70	28.01	42.5	46.33
L10	0.74	2.58	91.2	96.30	11.61	87.1	95.36	24.55	48.4	50.79
L11	0.78	3.07	92.5	96.92	13.90	84.9	95.07	30.35	46.0	48.69
L12	0.82	2.70	93.5	97.41	12.27	86.8	95.67	32.56	42.8	46.79

Continued on next page

Table 16 (continued): Phi-4-multimodal, Linear readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L13	0.98	2.86	93.8	97.53	12.35	86.5	95.49	31.92	45.7	48.72
L14	0.75	2.58	93.2	97.26	13.10	84.4	94.75	26.99	46.5	49.53
L15	0.94	2.96	93.1	97.24	12.33	87.3	95.68	29.55	53.3	54.69
L16	1.52	3.63	91.7	96.60	12.55	85.8	95.06	30.21	58.7	59.04
L17	2.39	4.73	90.8	96.28	14.78	82.8	94.44	32.34	45.3	48.49
L18	3.34	5.06	90.9	96.37	14.80	82.0	94.32	32.46	40.3	44.99
L19	4.19	5.89	89.1	95.60	14.33	81.0	93.73	27.81	41.5	45.89
L20	4.16	6.10	88.7	95.42	13.97	81.4	93.98	29.93	42.1	46.22
L21	5.10	6.28	88.7	95.47	14.33	80.6	93.82	30.52	42.5	46.50
L22	5.62	5.67	89.6	95.94	13.08	81.9	94.22	29.82	44.6	47.91
L23	6.47	6.20	89.0	95.75	13.40	80.9	94.03	31.12	39.5	44.52

Table 17: Phi-4-multimodal encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.11	2.62	82.0	90.89	37.83	15.8	15.93	39.07	41.9	63.99
L01	0.20	1.73	90.8	96.03	15.71	19.6	19.60	25.93	39.3	63.07
L02	0.08	1.35	91.7	96.49	28.77	45.8	51.49	29.87	72.4	75.10
L03	0.12	0.69	94.8	97.89	7.76	40.1	43.32	34.64	40.3	63.38
L04	**0.05**	0.65	96.8	98.77	9.06	89.7	**96.54**	25.14	53.8	55.23
L05	**0.05**	0.57	96.8	98.76	**7.13**	57.2	66.26	19.97	62.5	72.61
L06	**0.04**	0.82	90.9	96.08	9.58	61.6	72.06	19.61	52.8	68.04
L07	0.08	**0.51**	**97.2**	98.91	11.09	71.9	83.73	**18.03**	**82.1**	**83.85**
L08	**0.04**	**0.38**	96.5	98.62	**7.13**	87.3	94.67	**17.13**	**79.3**	**82.04**
L09	**0.05**	0.54	**97.2**	**98.93**	8.22	**93.6**	**97.73**	19.77	72.3	72.33
L10	0.08	0.90	93.2	97.18	13.19	62.6	73.32	22.57	78.0	80.71
L11	0.20	1.07	90.5	95.87	9.05	79.9	90.21	23.55	75.7	77.30
L12	0.36	1.40	94.4	97.75	8.28	87.1	94.69	21.82	76.8	77.50
L13	0.16	1.01	96.0	98.44	11.82	68.6	80.17	20.81	76.5	79.09
L14	0.31	1.20	92.7	96.97	**6.90**	**90.5**	96.45	23.25	74.1	74.40
L15	0.35	1.40	**97.5**	**99.05**	28.28	58.9	73.56	25.06	71.2	71.38
L16	0.82	2.18	93.7	97.43	20.99	64.2	77.62	27.44	68.7	68.97
L17	0.59	2.45	92.6	96.96	34.70	48.7	57.56	24.78	76.6	78.81
L18	1.06	2.74	93.1	97.19	11.72	86.0	95.60	29.62	40.4	45.03
L19	1.02	2.47	94.7	97.91	21.10	76.0	91.39	26.59	49.7	51.75
L20	1.03	3.55	91.6	96.53	17.56	76.5	90.69	22.04	76.0	76.63
L21	1.77	4.69	90.5	96.08	14.16	81.2	93.94	23.55	68.2	68.22
L22	2.31	4.32	88.8	95.16	15.40	79.4	92.02	29.72	51.6	53.09
L23	2.59	4.39	90.6	96.14	22.25	67.4	83.02	33.05	53.5	54.17

Table 18: Phi-4-multimodal encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.24	3.21	86.2	93.59	12.25	84.6	94.05	34.99	55.0	55.75
L01	0.27	3.09	88.4	94.82	14.92	80.3	93.73	30.88	42.5	46.63
L02	0.59	2.90	87.4	94.31	15.04	76.5	92.84	29.26	34.0	41.21
L03	0.43	1.69	88.3	94.73	9.65	86.3	94.36	19.37	**64.0**	**64.18**
L04	0.35	0.87	94.3	97.68	**8.43**	90.2	96.48	**16.89**	**67.3**	**67.34**
L05	**0.12**	0.46	97.2	98.92	8.58	90.9	96.86	21.27	39.2	44.48
L06	0.24	0.68	94.3	97.67	8.97	89.6	96.53	21.89	52.7	54.45

Continued on next page

Table 18 (continued): Phi-4-multimodal, Depth-SFT

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L07	**0.12**	**0.38**	**98.8**	**99.54**	**7.62**	**93.4**	**97.76**	19.57	40.6	45.47
L08	**0.12**	0.58	96.8	98.75	8.80	**91.4**	**97.13**	**17.33**	55.9	57.16
L09	0.20	0.66	98.2	99.31	8.68	88.7	96.46	20.08	27.8	37.57
L10	0.31	**0.41**	**99.3**	**99.74**	9.54	82.5	95.02	20.89	28.3	37.83
L11	0.35	1.51	97.5	99.08	9.87	80.0	94.41	20.74	27.6	37.45
L12	**0.16**	1.62	92.7	96.99	9.29	87.4	95.73	19.42	35.7	42.29
L13	0.28	1.19	97.2	98.92	9.49	85.9	95.62	22.64	35.2	41.93
L14	0.24	1.01	98.4	99.40	10.79	83.8	95.26	24.76	28.1	37.75
L15	0.43	1.01	97.9	99.21	9.60	85.9	95.73	24.29	27.8	37.58
L16	0.63	1.22	97.2	98.95	9.79	89.5	96.52	24.43	39.6	44.77
L17	0.71	1.55	97.4	99.03	10.83	85.0	95.41	24.37	29.4	38.50
L18	1.06	2.00	96.3	98.59	11.39	85.1	95.45	24.20	32.1	40.05
L19	1.64	1.88	96.5	98.69	13.28	79.3	94.18	27.29	29.3	38.44
L20	1.65	2.69	94.9	98.03	11.79	83.5	95.00	25.09	30.6	39.16
L21	1.88	2.85	94.0	97.62	11.53	81.8	94.60	23.78	39.6	44.76
L22	3.41	3.64	93.1	97.29	14.10	80.9	94.25	31.14	29.2	38.31
L23	3.88	3.61	92.2	96.83	13.04	81.7	94.20	30.05	31.2	39.52

F.5 MOSS-AUDIO-4B

Tables 19–21 report the complete depth profiles for the three procedures. Under Depth-SFT, every connection in L08–L15 has lower EER than every connection in L24–L30 on 21LA, while the latter region uniformly performs better on ITW. For example, L09 outperforms L27 on 21LA (9.51% versus 16.77%), but their ordering reverses on ITW (22.40% versus 17.17%).

Table 19: MOSS-Audio-4B encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.86	4.72	81.6	90.71	17.36	78.9	92.26	43.76	29.4	37.28
L01	1.06	3.72	84.2	92.37	15.86	81.5	93.64	37.38	33.8	40.64
L02	0.71	3.89	83.1	91.65	14.31	82.9	93.71	37.32	40.9	44.97
L03	0.43	2.75	86.3	93.61	14.55	82.4	94.17	31.98	45.4	48.32
L04	**0.27**	2.64	86.2	93.58	13.90	83.5	94.35	28.42	46.5	49.05
L05	0.39	2.49	88.0	94.58	14.29	83.6	94.60	25.97	55.1	56.13
L06	0.40	1.92	90.6	95.94	14.13	84.3	94.90	**23.86**	53.4	54.98
L07	0.35	1.33	93.0	97.09	13.03	85.4	95.32	**20.22**	55.4	56.71
L08	0.32	1.50	93.2	97.20	12.56	86.4	95.55	24.87	46.2	49.39
L09	**0.31**	1.55	92.9	97.07	12.38	86.4	95.55	25.63	52.1	53.90
L10	0.39	1.48	93.1	97.17	13.40	86.0	95.49	28.62	48.8	51.29
L11	**0.31**	1.35	93.2	97.20	12.90	**86.5**	**95.59**	27.02	52.1	53.89
L12	0.32	**1.29**	**94.0**	**97.56**	12.72	86.4	**95.59**	26.76	47.1	50.04
L13	0.32	**1.32**	**93.8**	**97.48**	12.32	**86.6**	**95.65**	27.14	44.3	48.03
L14	0.35	1.42	92.2	96.72	12.65	86.1	95.44	25.94	49.4	51.79
L15	0.43	2.15	88.8	95.03	13.42	84.4	94.75	27.39	50.6	52.66
L16	0.44	2.67	89.8	95.59	**11.98**	85.3	94.96	27.39	56.6	57.40
L17	0.44	2.60	90.4	95.89	12.17	85.6	95.08	25.96	**58.2**	**58.85**
L18	0.47	2.60	91.3	96.31	**11.71**	85.9	95.17	26.50	**56.7**	**57.56**
L19	0.63	2.58	91.3	96.30	12.30	85.3	95.03	26.77	53.7	55.06
L20	0.63	3.07	89.6	95.46	12.64	84.6	94.69	26.59	50.7	52.65
L21	0.59	3.09	89.2	95.25	12.68	84.4	94.57	28.16	49.8	51.91
L22	0.59	3.30	88.4	94.82	12.35	84.6	94.62	28.96	52.0	53.50
L23	0.51	3.41	88.4	94.84	12.50	84.4	94.59	28.28	49.7	51.79
L24	0.59	3.19	88.9	95.14	12.44	84.5	94.62	27.63	48.1	50.70

Continued on next page

Table 19 (continued): MOSS-Audio-4B, Linear readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L25	0.51	3.18	88.9	95.14	13.19	83.7	94.40	26.95	50.0	52.13
L26	0.71	3.48	88.1	94.71	13.84	82.9	94.21	28.17	49.3	51.58
L27	0.55	3.06	89.1	95.20	13.51	83.2	94.35	28.17	49.0	51.39
L28	0.67	3.00	89.1	95.20	13.92	82.5	94.14	29.09	47.0	49.79
L29	0.63	3.23	88.6	94.96	14.32	81.5	93.87	28.60	48.2	50.77
L30	0.63	3.41	88.4	94.83	15.32	80.4	93.51	27.79	48.7	51.15
L31	0.59	3.62	88.0	94.67	16.22	79.6	93.31	28.35	48.3	50.83

Table 20: MOSS-Audio-4B encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.30	2.69	84.9	92.75	13.91	25.5	25.66	41.44	40.3	63.28
L01	0.12	1.89	86.4	93.65	11.85	21.3	21.31	37.47	38.6	62.82
L02	0.20	1.00	94.6	97.83	10.42	32.1	33.24	33.30	48.1	65.68
L03	0.12	0.65	96.0	98.42	**8.39**	50.6	57.57	**24.14**	46.0	64.81
L04	0.08	1.16	87.7	94.38	10.99	81.6	94.71	30.11	37.5	43.45
L05	0.12	0.58	**97.6**	**99.09**	**8.25**	**91.5**	**97.13**	**16.82**	63.3	63.67
L06	**0.04**	0.67	**97.0**	**98.83**	11.24	76.5	93.73	31.92	33.4	40.88
L07	0.12	0.69	94.0	97.55	10.87	83.5	95.14	31.12	41.5	46.08
L08	**0.04**	**0.50**	95.0	97.99	12.02	78.6	94.12	45.01	37.5	43.39
L09	**0.07**	0.88	92.6	96.89	10.13	88.3	96.15	34.73	42.8	46.93
L10	**0.04**	1.10	87.8	94.45	10.97	87.3	95.81	36.75	47.7	50.48
L11	**0.04**	**0.54**	95.5	98.22	9.76	**89.3**	**96.52**	25.14	54.6	55.90
L12	0.16	1.21	92.1	96.66	10.73	87.6	95.77	28.15	51.5	53.16
L13	0.08	0.92	93.9	97.50	13.21	80.8	94.47	43.96	40.6	45.38
L14	**0.04**	0.63	96.1	98.48	10.01	86.7	95.85	27.95	46.7	49.60
L15	0.12	1.09	91.3	96.31	11.40	80.5	94.35	29.10	42.4	46.62
L16	0.16	1.09	95.9	98.40	11.48	84.4	95.13	27.05	49.9	52.21
L17	0.12	0.79	93.9	97.50	10.81	83.7	94.74	27.95	57.5	58.28
L18	0.11	1.02	93.6	97.38	14.56	81.5	94.03	35.02	48.4	51.04
L19	0.27	1.59	91.1	96.22	14.57	81.8	93.47	29.75	59.9	60.44
L20	0.21	1.08	95.1	98.05	14.98	81.9	94.45	27.65	46.8	49.89
L21	0.16	1.74	88.7	94.93	14.98	81.5	93.58	35.88	54.4	55.61
L22	0.16	1.59	90.8	96.06	16.21	81.0	93.89	30.41	59.1	59.73
L23	0.16	1.33	92.7	96.97	14.91	81.0	93.14	37.47	48.0	50.75
L24	0.16	1.50	94.9	97.99	13.92	82.0	93.66	30.43	58.3	58.99
L25	0.24	1.13	93.4	97.30	14.85	80.9	93.18	28.33	61.7	62.10
L26	0.12	0.95	95.9	98.38	16.56	79.0	92.15	24.25	66.1	66.21
L27	0.16	1.07	94.0	97.56	14.72	78.8	91.39	25.64	**69.2**	**69.17**
L28	0.16	1.07	95.9	98.40	14.31	80.3	92.92	27.16	59.0	59.74
L29	0.24	1.40	95.0	97.99	16.75	78.3	91.96	32.27	45.5	48.91
L30	0.27	1.45	94.9	97.95	18.08	78.5	92.45	31.46	47.8	50.60
L31	0.24	1.33	92.6	96.93	18.31	72.7	87.14	26.86	**69.8**	**69.87**

Table 21: MOSS-Audio-4B encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	2.12	5.03	87.2	94.28	13.36	83.1	94.09	36.75	37.7	42.26
L01	0.98	3.34	89.8	95.61	11.78	87.3	95.52	31.86	52.2	53.25
L02	0.51	1.65	96.3	98.55	10.76	88.5	96.27	27.83	42.9	46.93

Continued on next page

Table 21 (continued): MOSS-Audio-4B, Depth-SFT

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L03	0.47	0.83	98.1	99.29	9.78	**90.4**	96.89	30.59	39.0	44.40
L04	0.43	0.78	**98.4**	**99.39**	10.13	89.7	96.73	29.23	40.2	45.18
L05	0.35	0.82	**98.5**	**99.44**	10.00	**91.1**	**97.11**	26.60	48.9	51.47
L06	0.31	0.73	98.3	99.35	10.90	90.3	**96.90**	24.25	48.0	50.83
L07	0.24	0.84	97.9	99.20	10.31	90.2	96.83	25.60	55.0	56.34
L08	0.20	0.95	91.5	96.39	**9.66**	**90.4**	96.63	22.06	67.5	67.55
L09	**0.12**	1.22	88.7	94.93	**9.51**	89.2	96.08	22.40	61.9	62.36
L10	**0.16**	1.73	88.6	94.90	10.76	87.5	95.57	23.65	60.8	61.26
L11	0.27	1.88	87.1	94.09	9.83	87.9	95.56	23.43	60.2	60.72
L12	**0.16**	1.47	89.8	95.55	10.16	88.4	95.94	23.48	59.5	60.15
L13	**0.16**	1.68	90.8	96.05	10.64	87.5	95.71	22.87	59.3	60.01
L14	**0.16**	1.90	86.1	93.49	12.62	85.2	94.66	25.78	**67.6**	**67.59**
L15	**0.16**	1.66	89.9	95.58	12.44	85.5	95.14	23.77	63.0	63.29
L16	0.20	**0.52**	94.7	97.86	12.51	86.1	95.63	19.72	53.1	54.86
L17	**0.16**	0.65	91.1	96.19	15.54	84.8	95.03	20.94	62.0	62.43
L18	0.24	0.67	90.3	95.77	14.77	84.2	94.80	19.16	60.1	60.76
L19	0.24	0.69	94.0	97.56	15.37	84.5	95.19	18.65	55.0	56.40
L20	0.20	0.76	94.3	97.70	15.46	84.6	95.24	18.03	52.5	54.32
L21	0.24	0.60	95.2	98.09	16.28	85.4	95.50	19.12	53.5	55.16
L22	0.24	**0.57**	97.0	98.86	18.90	82.5	94.91	**17.26**	46.3	49.57
L23	0.20	0.58	96.9	98.78	18.27	84.0	95.22	18.82	53.1	54.88
L24	0.20	0.58	96.9	98.79	19.44	82.6	94.93	19.03	40.4	45.36
L25	0.20	0.67	96.9	98.81	18.38	82.2	94.81	19.08	51.3	53.41
L26	0.20	0.76	94.3	97.68	15.62	83.5	94.95	18.63	57.1	58.11
L27	**0.16**	0.71	96.7	98.73	16.77	81.6	94.67	**17.17**	59.3	60.08
L28	**0.16**	0.61	97.1	98.87	18.11	80.1	94.39	18.29	55.0	56.41
L29	0.20	0.63	97.1	98.86	20.10	80.7	94.48	18.64	55.2	56.57
L30	**0.16**	0.64	96.8	98.76	20.97	81.4	94.63	18.11	58.9	59.77
L31	0.20	0.65	97.2	98.91	21.80	80.3	94.41	22.05	**70.0**	**70.00**

F.6 AUDIO-FLAMINGO-NEXT

Tables 22–24 report the complete depth profiles for the three procedures. The three procedures exhibit distinct depth preferences. On ITW, their lowest observed EERs occur at L13 for Linear (21.81%), L19 for AASIST (17.41%), and L29 for Depth-SFT (27.11%), compared with 35.76%, 49.13%, and 37.41% at L31, respectively. These results further show that favorable encoder depths depend on the downstream detection procedure rather than being shared across readout and adaptation settings.

Table 22: Audio-Flamingo-Next encoder-depth results with Linear readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.50	7.14	62.7	72.91	18.24	67.7	81.09	54.30	29.9	35.96
L01	0.16	5.06	68.3	79.21	16.03	72.2	85.21	45.31	31.8	38.64
L02	0.16	5.62	67.4	78.16	16.74	70.9	84.14	41.72	33.2	39.73
L03	0.11	4.53	69.4	80.28	14.22	68.9	81.47	50.82	29.9	36.89
L04	0.16	4.00	66.8	77.58	15.71	58.4	68.96	51.93	32.0	35.97
L05	0.16	4.49	70.1	80.95	16.02	69.0	81.91	44.25	30.5	38.39
L06	0.08	3.32	72.9	83.63	13.17	77.7	89.29	46.70	28.6	37.48
L07	0.08	2.23	75.0	85.56	12.58	79.9	90.82	40.50	32.3	39.77
L08	0.08	1.93	73.8	84.51	11.75	74.4	86.33	42.91	31.3	38.67
L09	0.04	1.86	75.6	86.08	11.47	80.9	91.37	40.29	32.4	39.36
L10	0.11	1.92	72.8	83.61	11.74	79.9	90.71	37.79	36.7	42.03

Continued on next page

Table 22 (continued): Audio-Flamingo-Next, Linear readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L11	0.04	1.33	76.1	86.42	12.20	81.9	92.36	30.36	40.1	44.86
L12	0.04	0.93	78.0	88.00	11.59	82.5	92.50	28.96	52.8	53.84
L13	0.04	**0.60**	86.9	93.92	9.99	88.5	95.65	**21.81**	**64.4**	**64.51**
L14	0.04	0.64	84.9	92.77	11.12	87.0	95.06	23.95	**66.2**	**66.19**
L15	**0.01**	0.88	82.9	91.47	10.46	87.1	94.85	**22.98**	**64.4**	64.48
L16	0.04	**0.56**	89.4	95.32	10.03	89.2	95.89	25.13	45.3	48.43
L17	0.04	0.86	87.9	94.54	**9.87**	**89.8**	**96.23**	27.85	45.9	48.79
L18	0.04	0.79	90.8	96.04	**9.90**	**90.8**	**96.75**	27.47	42.8	46.81
L19	0.08	0.82	89.1	95.19	10.99	88.2	95.93	24.36	47.4	50.26
L20	0.08	0.77	**91.4**	**96.33**	10.94	87.4	95.72	25.68	39.8	44.93
L21	0.08	0.90	**92.2**	**96.73**	10.31	88.7	96.06	24.09	45.5	48.89
L22	0.07	1.28	83.6	91.95	14.27	80.6	92.66	30.87	33.3	40.70
L23	**0.03**	0.99	83.7	92.00	14.87	80.2	92.58	30.24	40.1	45.08
L24	0.16	1.00	87.5	94.29	16.35	79.2	92.56	32.64	36.1	42.44
L25	0.08	0.99	86.8	93.88	16.91	79.5	92.68	28.37	33.4	40.81
L26	0.08	0.98	88.6	94.90	15.21	81.0	93.34	32.63	30.2	38.90
L27	0.08	0.87	88.1	94.63	14.55	81.5	93.32	34.30	29.9	38.70
L28	0.13	1.59	87.6	94.37	16.03	80.4	93.04	28.93	34.9	41.76
L29	0.08	1.51	86.5	93.73	15.46	80.9	92.84	29.48	30.8	39.26
L30	0.12	1.18	90.8	96.03	15.33	81.5	93.54	33.47	29.9	38.74
L31	0.27	2.24	85.4	93.08	18.46	77.8	91.66	35.76	30.4	38.96

Table 23: Audio-Flamingo-Next encoder-depth results with AASIST readout. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	0.64	10.36	58.2	67.33	35.45	57.9	87.08	54.44	26.5	33.03
L01	2.83	13.19	54.8	62.84	24.87	62.8	77.64	50.78	48.7	48.93
L02	3.45	11.50	75.7	86.99	26.37	54.3	89.97	50.78	27.6	37.37
L03	0.51	7.06	79.1	89.09	24.91	48.3	89.89	46.89	27.1	37.20
L04	0.59	5.87	82.0	91.02	20.79	70.4	87.86	54.59	29.4	37.42
L05	0.31	2.74	71.6	82.47	23.52	65.2	83.81	33.40	30.5	39.01
L06	0.19	4.07	71.9	82.74	21.00	61.2	91.07	51.93	27.1	37.19
L07	0.12	3.04	77.8	87.85	19.09	55.4	90.65	51.08	27.3	37.26
L08	0.24	2.75	76.8	87.02	16.86	57.4	90.91	44.90	27.1	37.18
L09	0.20	4.32	72.5	83.36	21.99	58.6	90.90	38.44	27.1	37.18
L10	**0.08**	2.00	78.8	88.60	16.44	69.8	91.38	41.50	28.5	37.98
L11	**0.04**	1.82	79.6	89.20	17.40	70.7	92.26	31.01	27.7	37.51
L12	0.16	2.30	86.6	93.78	17.61	73.4	88.51	32.84	50.7	52.61
L13	**0.04**	2.25	79.7	89.24	14.81	75.6	**92.94**	25.43	40.2	45.11
L14	0.15	2.32	80.2	89.62	15.57	78.5	92.01	**25.02**	57.0	57.63
L15	0.24	1.69	74.8	85.37	16.86	55.4	65.02	28.22	59.1	**70.15**
L16	**0.08**	1.81	88.8	94.98	15.86	**81.8**	**93.48**	29.71	63.1	63.14
L17	**0.08**	**1.39**	89.5	95.38	**13.70**	**80.3**	91.85	28.82	**65.2**	65.27
L18	0.11	1.82	64.3	91.72	16.38	58.7	91.03	28.55	27.4	37.33
L19	**0.04**	**0.55**	**98.5**	**99.42**	**9.95**	78.9	89.64	**17.41**	**81.2**	**83.06**
L20	0.12	4.03	69.6	80.48	18.69	61.6	73.54	32.18	64.4	68.89
L21	0.24	2.24	88.9	95.08	15.36	76.7	89.16	28.44	65.0	64.97
L22	0.12	2.24	**92.4**	**96.82**	17.17	78.7	92.90	46.41	43.3	45.47
L23	0.16	2.66	54.1	61.85	19.62	44.9	50.10	48.66	47.0	55.95
L24	0.16	3.07	79.6	89.22	18.51	61.4	72.97	56.34	43.0	44.39
L25	0.19	3.83	79.5	89.18	17.58	54.1	62.88	49.26	48.0	53.02
L26	0.16	3.43	74.8	85.40	17.38	61.8	73.46	47.82	50.9	51.19
L27	0.12	3.91	72.3	83.14	17.86	67.8	81.03	39.50	53.7	54.28

Continued on next page

Table 23 (continued): Audio-Flamingo-Next, AASIST readout

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L28	0.15	4.15	78.9	88.68	17.08	73.6	87.34	33.46	34.0	41.13
L29	0.27	3.60	87.9	94.57	15.09	66.3	78.41	40.81	48.6	49.69
L30	0.12	4.26	78.7	88.58	14.98	69.9	82.46	45.54	52.1	52.09
L31	0.24	3.73	80.0	89.51	17.94	64.8	77.21	49.13	50.2	51.78

Table 24: Audio-Flamingo-Next encoder-depth results with Depth-SFT. Highlighting ranks layers within each dataset and metric.

	Dev	19LA eval			21LA			ITW		
Layer	EER↓	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑	EER↓	F1↑	Acc↑
L00	1.22	8.77	72.0	82.92	21.99	68.1	84.40	62.08	27.5	34.54
L01	0.55	7.00	76.6	86.95	23.11	67.6	82.89	55.66	27.6	36.31
L02	0.63	5.93	80.9	90.24	17.20	75.6	90.08	55.91	27.7	36.05
L03	0.16	5.06	83.5	91.90	16.32	76.9	90.49	50.41	27.1	36.31
L04	0.20	3.78	82.0	90.94	14.24	**81.9**	92.88	45.12	28.2	37.58
L05	0.08	3.07	84.4	92.46	14.89	81.1	93.44	53.31	27.6	37.31
L06	0.08	1.58	**91.9**	**96.59**	19.43	**81.2**	**94.11**	37.19	27.3	37.27
L07	0.11	1.31	89.8	95.51	16.64	79.0	**93.46**	45.37	27.2	37.23
L08	0.08	1.85	86.0	93.42	17.46	76.3	92.31	48.40	27.4	37.33
L09	0.04	1.32	86.6	93.79	15.11	80.0	93.05	58.76	27.3	37.09
L10	0.09	4.44	77.4	87.54	19.10	71.1	89.71	39.26	27.4	37.37
L11	0.08	1.39	89.9	95.59	26.71	70.0	91.52	41.56	27.1	37.19
L12	0.12	2.29	82.8	91.44	27.06	72.2	90.84	40.80	27.4	37.36
L13	0.04	**1.12**	88.0	94.55	15.68	79.0	93.19	28.98	28.6	38.04
L14	**0.01**	4.13	79.4	89.02	15.11	76.6	91.46	39.48	27.3	37.28
L15	0.04	**1.25**	88.0	94.57	**12.04**	80.6	93.37	42.12	30.2	38.91
L16	0.04	2.58	78.0	87.95	**13.47**	77.6	91.21	30.35	34.6	41.59
L17	0.08	3.10	71.5	82.33	14.11	77.2	89.76	36.72	**46.7**	**49.51**
L18	0.07	2.09	75.7	86.12	14.65	76.7	91.05	34.07	**37.8**	**43.52**
L19	0.04	1.74	86.6	93.77	14.84	74.2	91.81	36.67	28.0	37.64
L20	**0.00**	1.28	**91.0**	**96.15**	16.55	74.9	92.47	40.07	28.4	37.86
L21	0.12	5.93	68.4	79.31	16.90	72.8	87.93	38.51	35.7	41.90
L22	0.12	4.38	71.5	82.43	23.52	65.7	87.06	48.97	28.8	37.79
L23	0.16	4.58	68.5	79.42	21.88	65.4	85.71	45.56	33.8	40.05
L24	0.19	2.68	73.9	84.63	24.20	66.9	87.35	43.37	28.4	37.79
L25	0.27	5.06	73.8	84.51	24.05	64.6	86.81	43.70	27.5	37.20
L26	0.20	2.82	81.4	90.49	28.20	68.2	89.32	50.01	27.3	37.21
L27	0.12	2.04	84.7	92.64	30.15	69.6	90.13	45.46	27.2	37.23
L28	0.12	1.50	89.2	95.21	27.94	70.8	90.92	38.72	27.4	37.31
L29	0.24	2.34	83.4	91.84	31.57	67.3	89.02	**27.11**	28.4	37.87
L30	0.31	1.41	89.5	95.38	25.05	71.2	91.37	**27.60**	27.2	37.22
L31	0.67	2.05	89.3	95.30	25.49	72.1	91.12	37.41	30.6	38.98