ICLAD: In-Context Learning with Comparison-Guidance for Audio Deepfake Detection

Benjamin Chou[†]
Purdue University, USA
chou150@purdue.edu

Yi Zhu
Reality Defender Inc., USA
yi.zhu@inrs.ca

Surya Koppisetti
Reality Defender Inc., USA
surya@realitydefender.ai

Abstract

Audio deepfakes pose a significant security threat, yet current state-of-the-art (SOTA) detection systems do not generalize well to realistic in-the-wild deepfakes. We introduce a novel **I**n-**C**ontext **L**earning paradigm with comparison-guidance for **A**udio **D**eepfake detection (**ICLAD**). The framework enables the use of audio language models (ALMs) for training-free generalization to unseen deepfakes and provides textual rationales on the detection outcome. At the core of ICLAD is a pairwise comparative reasoning strategy that guides the ALM to discover and filter hallucinations and deepfake-irrelevant acoustic attributes. The ALM works alongside a specialized deepfake detector, whereby a routing mechanism feeds out-of-distribution samples to the ALM. On in-the-wild datasets, ICLAD improves macro F1 over the specialized detector, with up to $2\times$ relative improvement. Further analysis demonstrates the flexibility of ICLAD and its potential for deployment on recent open-source ALMs.

1 Introduction

Generative models can now synthesize highly convincing audio deepfakes from a few seconds of human speech (Le et al., 2023), posing significant societal risks of misinformation and identity forgery (Pender, 2023; Cox, 2023). Audio Deepfake Detection (ADD) has hence become a critical area of research. State-of-the-art (SOTA) ADD models typically rely on fine-tuning large self-supervised models on deepfakes generated from scripted speech collected in a studio (Xiao and Das, 2025; Chen et al., 2024; Truong et al., 2024; Tak et al., 2022b; Müller et al., 2024a), such as the ASVspoof datasets (Wang et al., 2020; Liu et al., 2023; Wang et al., 2024). In contrast to speech recorded "in the wild", scripted studio speech data are collected under controlled and constrained conditions, making it free from real-world extrinsic variations like background noise, room acoustics, and compression artifacts. Furthermore, being non-spontaneous, scripted speech lacks natural disfluencies, such as filled pauses, repetitions, and false starts common in conversation (Nagrani et al., 2017; Shriberg, 2005; McLaren et al., 2016). As a result, performance saturation has been observed on scripted studio deepfakes, accompanied by severe degradation when models are tested on realistic in-the-wild deepfakes (Ge et al., 2025; Zhu et al., 2025a). This generalization gap significantly limits the practical usage of existing detectors, as deepfakes used in real attacks inherently carry in-the-wild characteristics.

A common way to mitigate this generalization issue is to retrain detectors on expanded training data. However, repetitive supervised fine-tuning is both expensive and not sustainable as new deepfake generation methods continue to emerge rapidly. More importantly, real-world speech data are hard to obtain due to privacy concerns, and their specific patterns differ drastically across use cases (*e.g.*, noisy, monotonic telephony speech from call centers versus highly expressive voice from social media platforms). Hence, a generalizable deepfake detector must be able to adapt quickly and effectively to new conditions, without requiring extensive, costly domain-specific retraining.

To bridge this gap, we propose **I**n-**C**ontext **L**earning with comparison-guidance for **A**udio **D**eepfake detection (**ICLAD**), a training-free deepfake detection method that generalizes to unseen in-the-wild deepfakes with textual explanations on the classification outcome. The core idea of ICLAD lies in a Pairwise Comparative Reasoning (PCR) strategy, where the ALM is initially prompted to provide real and fake evidence simultaneously from the selected examples, without access to the ground-truth label. The ground-truth is then

[†]Work done during internship at Reality Defender Inc.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9242–9256

exposed to the model, enabling an iterative self-discovery process where the ALM identifies and filters deepfake-irrelevant attributes and inherent hallucinations. We further complement the ALM with a specialized deepfake detector to focus on subtle deepfake cues that may be overlooked by the ALM. We found that ICLAD yields higher macro F1 scores than specialized detectors on realistic in-the-wild deepfake datasets, while providing rich textual explanations to support its decision.

Our main contributions are:

1. We introduce ICLAD, built on an audio language model, demonstrating the first successful adaptation of in-context learning (ICL) for training-free detection of audio deepfakes.
2. We design a novel *pairwise comparative reasoning* strategy to guide the ALMs to identify and filter out hallucinations as well as deepfake-irrelevant acoustic attributes.
3. We show that ICLAD improves over the specialized detector on unseen in-the-wild deepfake datasets, while providing textual rationales for ALM decisions.

2 Related Work

2.1 Specialized Audio Deepfake Detectors

The dominant paradigm of ADD has been centered on training classifiers in a fully-supervised manner (Yi et al., 2023). SOTA systems typically employ self-supervised learning (SSL) encoders, such as Wav2Vec2 (Tak et al., 2022b) or WavLM (Combei), as frontend feature extractors, with a classification backend to map high-dimensional representations to a binary decision. While these models achieve near-perfect performance on scripted studio deepfake datasets (ASVspoof 2019, 2021) (Wang et al., 2020; Yamagishi et al., 2021), they fail to generalize to unseen attacks and in-the-wild speech (Müller et al., 2024a). While methods such as data augmentation (Tak et al., 2022a,b), frontend fine-tuning (Martín-Doñas and Álvarez, 2022; Wang et al., 2024), and extracting more robust features (Zhu et al., 2025b, 2024) have been proposed to mitigate this issue, they often come at a drastic increase in training cost and still show large discrepancies between performance achieved with scripted studio deepfakes and in-the-wild deepfakes (Müller et al., 2024a). This fundamental generalization gap underscores the need for a more adaptive paradigm that can handle novel threats without requiring constant retraining.

2.2 Audio Language Models for Audio Deepfake Detection

The recent advent of ALMs (Kong et al., 2024; Ghosh et al., 2025; KimiTeam et al., 2025; Goel et al., 2025; Xu et al., 2025) has introduced a new approach for ADD. Unlike specialized detectors, ALMs are pre-trained on a larger corpus of vast, diverse multimodal data, giving them a more general understanding of real-world speech patterns. However, current audio deepfake detection systems remain dominated by bespoke detectors, with no evidence that off-the-shelf ALMs can be deployed without task-specific adaptation (Gu et al., 2025). This gap stems largely from a data mismatch: ALMs are trained overwhelmingly on authentic speech (*e.g.*, Common Voice (Ardila et al., 2020)) with limited exposure to deepfakes. To bridge this gap, recent work has focused on adapting ALMs through Supervised Fine-Tuning (SFT), often by reformulating deepfake detection as an Audio Question Answering (AQA) task (*e.g.*, "Is this audio fake or real?") (Gu et al., 2025). While improved performance has been observed on scripted studio deepfakes, this SFT approach still inherits the limitations of costly finetuning and fails to generalize to in-the-wild data (Gu et al., 2025).

2.3 In-Context Learning for Adapting to Unseen Data

ICL is an emergent capability of large-scale models to perform a new task based on a few examples provided in the prompt, without any parameter updates (Liu et al., 2022; Olsson et al., 2022; Min et al., 2022). While ICL allows for rapid, training-free adaptation, the efficacy of ICL is notoriously fragile. Its performance is highly sensitive to the choice of exemplars, and it tends to learn superficial correlations rather than the underlying logic of a task (de Wynter, 2025; Chen et al., 2023). This brittleness leads to two critical challenges: first, ICL performs poorly on tasks requiring reasoning, and second, its performance degrades on Out-of-Distribution (OOD) data (de Wynter, 2025). Adapting ALMs from their pre-training on bona fide speech to deepfake detection is a clear example of such an OOD challenge. In our proposed framework, ICLAD, this generalization issue is addressed by a new PCR strategy that guides the ALM to self-discover deepfake-relevant attributes by leveraging its general audio understanding capabilities, thus achieving robust generalization.

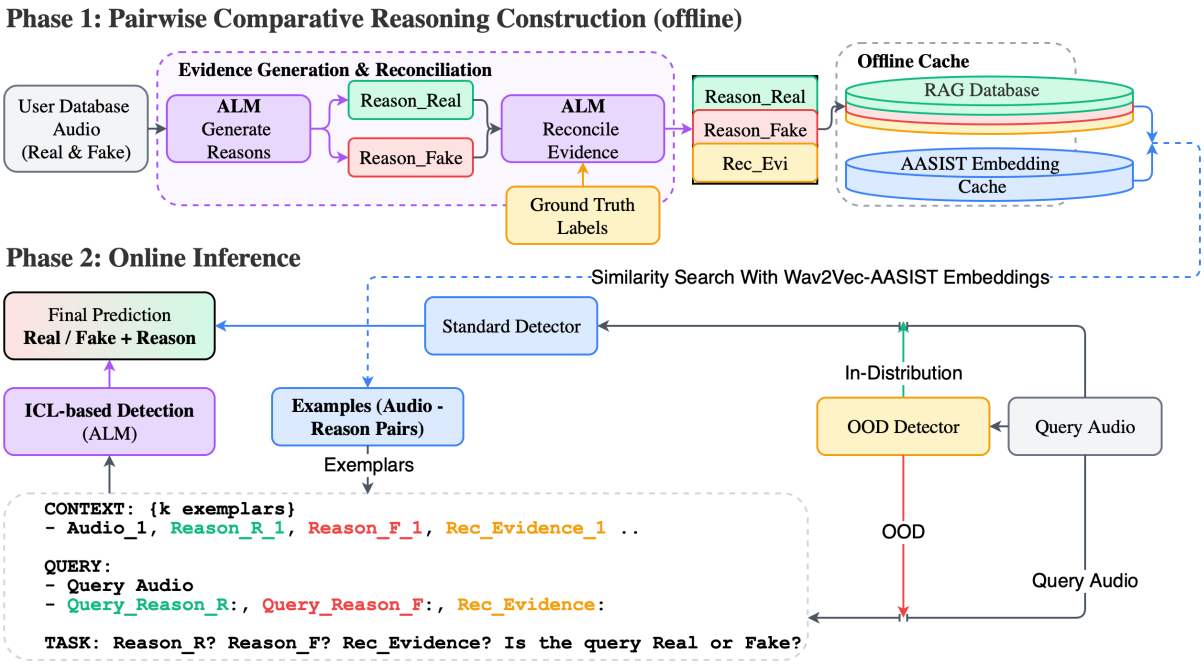

Figure 1: The ICLAD framework has two phases. In Phase-1 (Section 3.1), the ALM is first exposed to a database of labeled audio samples for which it generates evidence supporting both *real* and *fake* classes. Then, in a process we call Pairwise Comparative Reasoning (PCR), the ALM compares this conflicting evidence to produce a reconciled explanation that highlights discriminative attributes consistent with the ground-truth label. During online inference in Phase-2 (Section 3.2), an OOD detector is used to route out-of-distribution samples to the ALM. For a given query audio, we retrieve the most acoustically-similar sounding examples from Phase 1, along with their paired reasons, reconciled evidence, and ground-truth label, and employ in-context learning on the ALM to make a robust final prediction and reasoning on the query audio.

3 ICLAD

Figure 1 provides an overview of ICLAD, a comparative reasoning framework designed to achieve training-free generalization for in-the-wild deepfakes while yielding rich textual explanations. ICLAD entails two phases. In phase-1, the PCR strategy is used to guide an ALM to generate hallucination-aware evidence from a diverse sample pool. Phase-2 retrieves the most acoustically similar examples with their paired evidence and ground-truth from the phase-1 database, concatenates them into the ICL prompt, and directs the ALM to make a robust final decision on the query audio. The following sections detail the workflow and components of both phases.

3.1 Phase-1: Offline Reasoning

3.1.1 Motivation for Comparative Reasoning

We first observed that in a zero-shot setting, ALMs exhibit a strong bias that defaults consistently to a single class prediction (either "real" or "fake"). This demonstrates that ALMs cannot be used as off-the-shelf deepfake detection tools, motivating

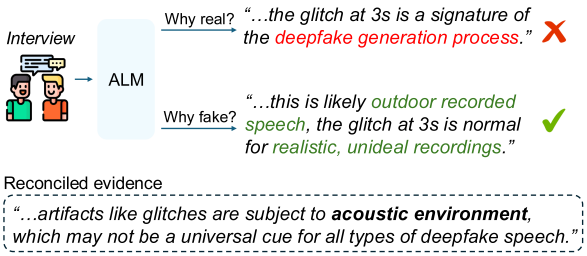

Figure 2: ALMs can cite the same attribute (e.g., a glitch during an interview) as both the sign of a *real* and *fake* speech. Our PCR strategy addresses the issue by forcing a comparison between *real* and *fake* evidence to obtain the *reconciled* evidence, which helps the ALM discover discriminative attributes consistent across datasets.

the use of ICL. Following conventions (Liu et al., 2022; Zhang et al., 2023), we initially tested a simple [AUDIO]-[LABEL] format, interleaving audio tokens and corresponding labels in the prompt before the final query. However, this approach yielded performance no better than random chance. This failure of simple ICL suggests that ALMs could not independently infer the complex acoustic attributes required for deepfake detection, thus motivating

our exploration of more sophisticated reasoning.

During our exploration of different ICL strategies, we observed that ALMs can generate contradictory explanations based on ambiguous acoustic cues (Figure 2). For example, when prompted to provide evidence for both authenticity and forgery of the same audio clip, the might cite the presence of a glitch as evidence of the deepfake generation process, while simultaneously using the same cue as evidence of a real speech. We therefore design the PCR strategy to force a paired real and fake evidence comparison. This strategy compels the ALM to reconcile these contradictory rationales with the ground-truth, enabling it to explicitly pinpoint and filter ambiguous cues, while discovering the discriminative deepfake-related attributes.

3.1.2 Pairwise Comparative Reasoning

Initial Evidence Generation. For each audio sample $\mathbf{A}_i$ with its corresponding label $\mathbf{L}_i \in \{\text{real}, \text{fake}\}$, the ALM is initially prompted without access to the ground-truth label. The prompt is designed to compel the model to simultaneously generate two sets of textual explanations, namely the real evidence $\mathbf{R}_{real,i}$ and the fake evidence $\mathbf{R}_{fake,i}$. This pairwise evidence generation process forces the ALM to consider both sides of $\mathbf{A}_i$.

Evidence Reconciliation. To mitigate the evidence contradiction observed in the initial evidence generation step, we introduce a reconciliation step to discover and filter out ambiguous acoustic cues. For the audio sample $\mathbf{A}_i$, the ALM is prompted with $\mathbf{R}_{real,i}$, $\mathbf{R}_{fake,i}$, along with the ground-truth $\mathbf{L}_i$ to generate a reconciled evidence $\mathbf{R}_{reconciled,i}$. Intuitively, the ALM is tasked to review and reconcile its previously generated (potentially contradictory) descriptions based on the ground-truth. The goal of reconciliation is two-fold. First, we aim to identify and discount acoustic attributes that are not indicative of the true label, while being present. Second, to filter out hallucinated attributes that do not exist in the audio.

Offline Cache. The generated evidence $\mathbf{R}_{real,i}$, $\mathbf{R}_{fake,i}$, and $\mathbf{R}_{reconciled,i}$ for all training samples are stored in an offline cache, which serves as a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) database of high-quality, hallucination-aware explanations. To facilitate fast retrieval during inference time, for each audio sample in the database, we extracted embeddings using a pre-trained specialized deepfake detector Wav2Vec2-AASIST (Tak et al., 2022b), resulting in an associated AASIST embedding cache.

3.2 Phase-2: Online Inference

3.2.1 Example Retrieval

Upon receiving a query audio $\mathbf{A}_q$, we first extract the Wav2Vec2-AASIST embeddings, which are used to query the embedding cache constructed in phase-1. This retrieval-augmented process identifies the K most acoustically similar exemplar entries from the cache, where each entry includes the audio, label, and evidence information $(\mathbf{A}_K, \mathbf{L}_K, \mathbf{R}_{real,K}, \mathbf{R}_{fake,K}, \mathbf{R}_{reconciled,K})$. To optimize the selection of examples, we experimented with different embedding choices and found Wav2Vec2-AASIST with the best detection results. The comparison is detailed in Section 5.3.

3.2.2 Dynamic routing

While ALM demonstrates strong capabilities in general audio understanding, specialized deepfake detectors may excel at capturing fine-grained deepfake acoustic artifacts. To leverage the complementarity between the two types of models, the query audio $\mathbf{A}_q$ is passed through an Out-of-Distribution detector. This detector assesses which detector $\mathbf{A}_q$ should be routed to. If $\mathbf{A}_q$ is classified as in-distribution (ID) (i.e., similar to studio speech/deepfakes), it is routed to a specialized deepfake detector to obtain a final decision. In this case, ALM is not involved. If $\mathbf{A}_q$ is classified as OOD (i.e., in-the-wild sample), it is sent to the ALM for further processing. For the OOD detector, we employ a standard k-Nearest Neighbor (k-NN) approach (Bukhsh and Saeed, 2023), implemented with the FAISS library.

Inference. For OOD samples, we retrieve K examples, each in the format of $(\mathbf{A}_K, \mathbf{L}_K, \mathbf{R}_{real,K}, \mathbf{R}_{fake,K}, \mathbf{R}_{reconciled,K})$, then embed into the ICL prompt. Similar to the reasoning process in phase-1, the ALM is asked to provide real, fake, and reconciled evidence and provide a binary decision (i.e., real or fake).

4 Experimental Setup

4.1 Datasets

We evaluate on five datasets: ASVspoof 2021 (21DF) and MLAAD-v3 (studio-scripted), and ITW, SpoofCeleb, and DFEval 2024 (in-the-wild). These cover 126,348 clips across 42 languages. See Table 12 and App. A.4 for detailed split sizes, languages, and licenses.

4.2 Evaluation Protocol

Data. We evaluate on **126,348** audio clips across the five corpora. (ASVspoof2021: **29,738**, MLAAD: **35,000**, ITW: **31,280**, SpoofCeleb: **18,226**, Deepfake-Eval-2024: **12,104**). For datasets without train/test splits, we subsample the ICL set disjointly from the test set. All audio files are truncated to 4 s to match the input prepossessing pipeline of the baseline.

Metrics. While Equal Error Rate (EER) has been the dominant performance metric for ADD (Yamagishi et al., 2021; Wang et al., 2025), its reliance on continuous scores (*e.g.*, raw logits) makes it unsuitable for evaluating our framework, where the ALM outputs a binary decision. Furthermore, since the calculation of EER does not require a binarized threshold, a low EER value may not directly translate into high accuracy (see Figure 3). To better reflect performance in a practical deployment scenario, where a hard classification is required and class imbalance is often present, we report macro F1-score and accuracy as our primary metrics. It should be noted that a fixed binarization threshold of 0.5 is used for calculating macro F1 and accuracy, rather than relying on the dataset-specific EER threshold. This approach accurately mimics real-world deployment conditions where optimal thresholds cannot be pre-calculated.

Hardware. Experiments were conducted on an NVIDIA A100 40 GB GPU.

4.3 ICLAD Setup

ALM choice. While ICLAD can be applied to any ALMs, we use Gemini-2.5 Flash as it demonstrates significantly stronger audio understanding capabilities on multiple benchmarks (Kong et al., 2025; Team et al., 2025). We further performed ablations with other open-source ALMs, such as Audio Flamingo 3 (AF3) in Section 5.5.

ICL hyperparameters. We select samples using embedding similarity as the retrieval mechanism from a RAG database consisting of 500 samples subsampled from 19DF(19DF) (Wang et al., 2020) and 500 from the target set's training split. We select 10 examples as the context for inferring each query audio, with 5 real and 5 fake.

OOD detector. Our OOD kNN router uses hyperparameters from the original paper (Bukhsh and Saeed, 2023): (k=5, $Threshold$=95%).

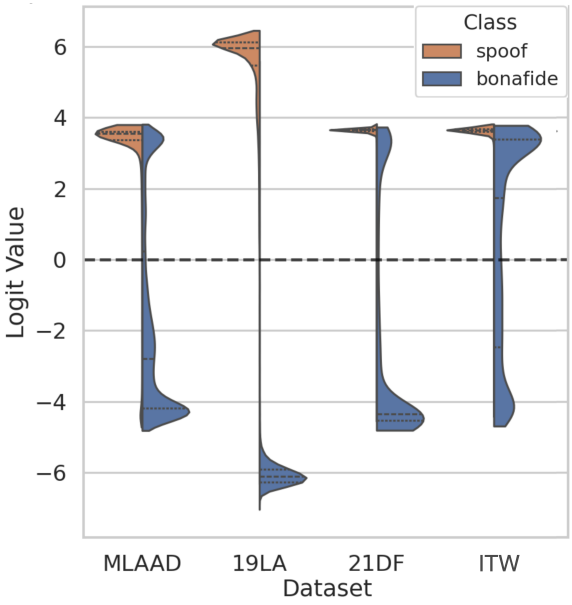

Figure 3: Logit distributions of Wav2Vec2-AASIST on ID (ASVspoof 2021) vs. OOD (ITW, SpoofCeleb) datasets. Overlap between classes is significantly more common on OOD data.

4.4 Baseline

For our comparative analysis, we selected Wav2Vec2-AASIST (Tak et al., 2022b) as the baseline deepfake detector. This choice was motivated by its demonstrated superior performance and stronger generalization capability when evaluated on in-the-wild speech compared to the five other specialized detectors shown in Table 1.

Table 1: Macro-F1 scores of SOTA detectors. Wav2Vec2-AASIST shows the best overall generalization.

Model	21DF	ITW	MLAAD
XLSR Mamba	**0.936**	0.523	0.558
WavLM ASP	0.786	0.474	0.398
XLSR Conformer	0.924	0.644	0.716
RawBMamba	0.738	0.444	0.695
Wav2Vec2-AASIST	0.925	**0.681**	**0.733**

Logit distributions. To further understand the generalization gap of specialized detectors, we analyzed the output logit distributions of the Wav2Vec2-AASIST baseline across different datasets (see Figure 3). On ID data like 19DF, the distributions for real and fake classes are well-separated. However, on in-the-wild datasets, the distributions significantly overlap, leading to poor classification performance and explaining the drop

in Macro F1 scores. This highlights the issue with using EER as the primary metric, because the optimal threshold varies between datasets, and a good threshold for one dataset can potentially lead to very poor performance on another.

5 Results and Analysis

5.1 Performance Comparison

Generalization to in-the-wild deepfakes. Table 2 shows the comparison between ICLAD and the SOTA specialized deepfake detector. The baseline detector retains an expected advantage on its matched training domains (MLAAD and ASVspoof 2021) due to its specialized fine-tuning. In sharp contrast, the generalization of ICLAD is evident across the other three in-the-wild datasets (i.e., ITW, SpoofCeleb, and DFEval 2024), where it consistently outperforms the baseline. The most significant discrepancy occurs on SpoofCeleb, where ICLAD's macro F1 score of 0.665 represents nearly a 2× improvement over the baseline's 0.334.

Explainability. Unlike specialized detectors that output only a score, our model produces textual reasoning alongside its decision, enabling qualitative inspection of the detection process. As illustrated by examples in Table 3, these rationales often focus on synthesis artifacts or physiological markers. At the same time, rationale quality is not guaranteed for every sample; we therefore report hallucination-focused listening-test results for the prompting strategies (Tables 5 and 6 and appendix A.2).

5.2 Comparison between ICL strategies

Setup. We run the ablations on five-datasets, totaling at **6,965** clips. We compare the *pairwise comparative* strategy with two other baseline strategies, namely a *simple* prompting strategy and a *knowledge-guided* prompting strategy. The former introduced a reasoning step to create an (AUDIO | REASON | LABEL) structure. The descriptions of the in-context examples were generated offline by prompting the model with the audio and its ground-truth label, compelling it to create a justification. The latter prompts the ALM to analyze a predefined set of acoustic attributes that are deemed discriminative by humans, including intonation and emotion, speech quality and audio artifacts, biological signs, and natural pacing and hesitations.

Table 2: Performance of ICLAD against a specialized detector. Best Macro F1 (↑) per dataset is bolded (see Appendix A.3 for statistical analysis results).

	Baseline		Ours	
Dataset	**Acc.**	**F1**	**Acc.**	**F1**
21DF	0.868	**0.866**	0.825	0.822
MLAAD	0.800	**0.798**	0.593	0.593
ITW	0.691	0.674	0.778	**0.777**
SpoofCeleb	0.501	0.334	0.668	**0.665**
DFEval 2024	0.500	0.367	0.550	**0.550**

Table 3: Correct and incorrect classifications made by ICLAD. GT: Ground-truth. Pred: ICLAD's output.

GT: Fake, Pred: Fake	
Reasoning	Speech flow is overly smooth and consistent, with no micro-hesitations ("um", "uh"), stutters, or self-corrections. Pauses are timed with machine-like precision, lacking natural irregularities.
GT: Real, Pred: Real	
Reasoning	Natural inhalations occur before phrases (*e.g.*, "conducts", "uh insane"), and spontaneous fillers ("uh") are audible.
GT: Real, Pred: Fake	
Reasoning	Although breathing is present, the broader range of physiological noises — lip smacks, tongue clicks, swallows — is missing.
GT: Fake, Pred: Real	
Reasoning	A clear, naturally timed inhalation at the very start, along with additional breath cues, strongly suggests human physiology.

Quantitative results. Table 4 presents the quantitative ablation results, showing that on average the PCR strategy outperforms both the simple and knowledge-guided strategies. The knowledge-guided strategy exhibits the largest performance variance, an expected outcome given its strong inductive bias. This bias is beneficial when its predefined attributes align with a dataset's distribution,

Table 4: Prompt ablation results across datasets. Best Macro F1 (↑) per dataset and on average is in bold.

Dataset	Metric	Simple	Explicit	*PCR*
21DF	Accuracy	0.8251	0.8267	0.8442
	Macro F1	0.8210	0.8231	**0.8422**
MLAAD	Accuracy	0.6355	0.5551	0.6111
	Macro F1	**0.6395**	0.5808	0.6110
ITW	Accuracy	0.7844	0.7359	0.8071
	Macro F1	0.8022	0.7204	**0.8045**
SpoofCeleb	Accuracy	0.6213	0.5976	0.6527
	Macro F1	0.6097	0.5662	**0.6511**
DFEval 2024	Accuracy	0.5661	0.5489	0.5411
	Macro F1	0.5554	**0.5834**	0.5410
Average	Accuracy	0.6865	0.6528	0.6917
	Macro F1	0.6856	0.6548	**0.6905**

but it becomes detrimental in other settings.

However, the superiority of PCR is not uniform. The limitations are most apparent on datasets where real and fake audio share ambiguous or overlapping cues. For example, in MLAAD, the scripted real class contains fewer physiological markers than its fake counterpart. In these challenging scenarios, PCR's reconciled evidence marks most extracted cues as unreliable. Consequently, the model defaults to its zero-shot bias toward predicting audios as real and leads to random chance performance. This reveals a fundamental trade-off: while PCR is an effective technique to filter hallucinated evidence, this process can be overly aggressive, inadvertently removing the discriminative cues required for a reliable in-context analysis.

Qualitative analysis. Since our method forces the ALM to self-explore acoustic attributes associated with deepfake labels, a common issue that we observed is hallucination, where non-existent attributes are generated in the textual explanation to justify the ground-truth. While hallucination may not necessarily lead to a decrease in accuracy, it will result in incorrect textual explanations. In our qualitative analysis, 22 human annotators with audio analysis experience were recruited to identify whether an audio recording is paired with hallucinated ALM explanation. Table 6 first provides an overview of different hallucination categories observed using the simple prompting strategy. Across 120 generated explanations, 18.3% has hallucinations, where prosody and naturalness related attributes are shown as the leading category. In contrast, only 10.0% of the explanation generated were identified with hallucinations using the PCR strategy. We further provide per-sample annotation result in Appendix A.2.

Table 5: Distribution of hallucination categories. Majority voting results indicate that the leading category is naturalness (37.4%).

Category of reason	Count	Proportion (%)
Prosody/Naturalness (pitch, intonation, pacing)	46	37.40
Other (semantic, uncategorized)	36	29.27
Physiological signals (breathing, throat)	20	16.26
Acoustics/artifacts (noise, distortion, clipping)	19	15.45
Unseen language	2	1.63
Total	**123**	**100.00**

With knowledge-guided prompting, we notice two distinct problems. First, without explicit guidance to treat the attributes separately, the ALM tends to describe different acoustic aspects with high correlation. For example, a single perceived cue in one category would cause it to invent conforming evidence in the others, resulting in a high hallucination rate of approximately 50%.

Secondly, we find that introducing human knowledge can severely bias ALM when testing on certain datasets. For example, the model learns to associate the absence of biological signs (*e.g.*, breaths or pauses) with fake audio. This heuristic failed on datasets like MLAAD, where the real speech is often professionally recorded and scripted, naturally containing fewer biological markers.

Table 6: Human evaluation of explanations generated using the *Simple* strategy. N=120.

Rating	Count	Percentage
High	64	53.3%
Medium	36	30.0%
Low	20	16.7%
Hallucinations	22	18.3%

Table 7: RAG ablation results across datasets. 'Detector' shows the performance of our non-ICL baseline: Wav2Vec2-AASIST. Best Macro F1 (↑) per dataset is in bold, second best is underlined.

Dataset	Metric	Detector	Wav2Vec2	AASIST	Text	AASIST+Text
21DF	Accuracy	0.9162	0.8415	0.8442	0.7740	0.8463
	Macro F1	**0.9148**	0.8394	0.8422	0.7656	<u>0.8443</u>
MLAAD	Accuracy	0.8025	0.6041	0.6111	0.5199	0.6182
	Macro F1	**0.7999**	0.6030	0.6110	0.5029	<u>0.6163</u>
ITW	Accuracy	0.6479	0.7552	0.8071	0.7403	0.7881
	Macro F1	0.5998	0.7483	**0.8045**	0.7237	<u>0.7850</u>
SpoofCeleb	Accuracy	0.5008	0.6497	0.6527	0.6441	0.6326
	Macro F1	0.3396	0.6462	**0.6511**	0.6472	<u>0.6287</u>
DFEval 2024	Accuracy	0.4990	0.5441	0.5411	0.4988	0.5170
	Macro F1	0.3654	0.5441	**0.5435**	0.5160	<u>0.5170</u>
Average	Accuracy	0.6733	0.6789	0.6912	0.6354	0.6804
	Macro F1	0.6039	0.6762	**0.6905**	0.6311	<u>0.6783</u>

5.3 Retrieval Embedding Ablation

Setup. We run the ablations on five-datasets, totaling at **6,965** clips. We conducted a systematic comparison of four different embedding choices to identify the optimal one for example retrieval. These embeddings include Wav2Vec2-XLSR (Baevski et al., 2020), Wav2Vec2-AASIST (Tak et al., 2022b), Qwen3-0.5B text embeddings computed from the evidence (Yang et al., 2025), and a combination of the audio and text embeddings. With the former three, we adopted cosine similarity for finding the most similar examples; with the audio+text embeddings, Maximal Marginal Relevance (MMR) was used to find examples with maximized audio similarity for relevance and minimized text similarity for diversity consideration.

Results. Table 7 compares the performance achieved using different embedding strategies for RAG, benchmarked against the specialized baseline **'Detector'**. Results confirm that Wav2Vec2-AASIST embeddings lead to the best performance on average, surpassing the baseline detector by 8.66%. This suggests that task-specific embeddings are more effective for finding acoustically meaningful examples than relying on general-purpose audio or text representations. In contrast, using text embeddings leads to consistently poorer performance. This is potentially caused by the high similarity in the textual description of real and fake classes, which may cause the PCR mechanism to overly aggressively filter attributes as unreliable. We attempted to exploit this phenomenon by combining Wav2Vec2-AASIST embedding similarity with text dissimilarity for retrieval, which led to marginally better performance on ID datasets (21DF: +0.21%; MLAAD: +0.53%). However, the approach degraded performance on in-the-wild datasets, indicating that a simple text dissimilarity is insufficient. Further investigation is required to effectively integrate textual semantics into the retrieval process.

5.4 Importance of Dynamic Routing

Table 8: OOD ablation. Best Macro F1 (↑) per dataset is in bold, second best is underlined.

Dataset	Strategy	Accuracy	Macro F1
21DF	PCR	0.6460	0.6456
	Baseline	0.9162	**0.9148**
	OOD	0.8442	<u>0.8422</u>
MLAAD	PCR	0.5265	0.5120
	Baseline	0.8025	**0.7999**
	OOD	0.6111	<u>0.6110</u>
ITW	PCR	0.6427	<u>0.6424</u>
	Baseline	0.6479	0.5998
	OOD	0.8071	**0.8045**
SpoofCeleb	PCR	0.5593	<u>0.5577</u>
	Baseline	0.5008	0.3396
	OOD	0.6527	**0.6511**
DFEval 2024	PCR	0.5374	<u>0.5281</u>
	Baseline	0.4980	0.3648
	OOD	0.5411	**0.5410**

Table 8 compares the performance obtained with and without dynamic routing. Results clearly reveal a performance trade-off between the two core components: the specialized baseline maintains its

superiority on ID datasets (MLAAD and 21DF), while the ICL framework demonstrates superior generalization on the diverse in-the-wild datasets (ITW, SpoofCeleb, and DFEval 2024).

The dynamic routing is designed to capitalize on this trade-off. It effectively routes ID samples to the baseline detector; for 21DF and MLAAD, it routes the majority of samples to the specialized detector, leading to significant macro F1 increases on 21DF (19.6%) and MLAAD (+9.9%). Furthermore, even on in-the-wild datasets, the OOD detector successfully identifies a small but critical subset of ID data where the specialized detector remains most accurate. This dynamic routing mechanism helps ICLAD exceed the performance of both the ICL and the specialized detector alone, yielding better macro F1s on all three in-the-wild datasets.

5.5 Open-Source ALM Evaluation

Table 9: Comparison between Gemini-2.5 Flash and Audio Flamingo 3 (AF3) using the *simple* ICL strategy. *For this test, AF3 is prompted using explanations pre-generated by Gemini.

Dataset	Model	Accuracy	Macro F1
21DF	AF3*	**0.6772**	**0.6507**
	Gemini	0.6192	0.6086
ITW	AF3*	**0.7890**	**0.7885**
	Gemini	0.6951	0.6909

We evaluate AF3 (7B parameters) as an open-source alternative to Gemini-2.5 Flash. However, AF3's instruction-following capabilities were severely limited, consistently outputting only a binary label without intermediate reasoning (See Appendix A.5). To enable comparison, we bypassed phase-1 and directly provided AF3 with the explanations generated by Gemini as examples and used the *simple* prompting strategy for ICL (Table 9).

Surprisingly, with pre-generated explanations from Gemini, AF3 obtains higher accuracy than Gemini-2.5 Flash. This suggests AF3 may possess good audio understanding capability but lacks the necessary thinking capability, likely due to its training objective being focused on direct output. This finding suggests a path forward for a fully open-source and more accurate version of ICLAD.

6 Conclusion

We introduce ICLAD, a deepfake detection paradigm that leverages in-context learning capabilities of ALMs to improve generalization to in-the-wild audio deepfakes. The ALM is guided with a Pairwise Comparative Reasoning (PCR) strategy, in order to generate evidence on discriminative per-class artifacts that are consistent across diverse audio samples. In our experiments, ICLAD improves over the specialized detector on three in-the-wild datasets, while the specialized detector remains stronger on scripted in-distribution datasets. ICLAD also produces textual rationales, and our listening-test analysis indicates lower hallucination rates for PCR than for simple prompting. The proposed framework is flexible to operate with recent open-source ALMs (*e.g.*, Audio Flamingo 3), supporting its practical deployment potential.

Limitations

A current limitation is ICLAD's performance degradation on scripted-speech datasets like MLAAD. Our results suggest that while ALM reasoning captures high-level inconsistencies, it cannot yet fully replace specialized detectors for low-level acoustic artifacts in studio settings. A further limitation of our work is its reliance on proprietary Gemini-2.5 Flash for its offline evidence generation and reconciliation phase. This dependency on a proprietary model was necessary due to the instruction-following capabilities required for the comparative reasoning strategy, which we found lacking in currently available open-source ALMs. However, our results demonstrate that recent open-source ALMs, such as Audio Flamingo 3, can obtain better performance when using Gemini-generated cues. With further improvements to instruction-following capabilities, open-source ALMs can be prompted with the proposed reasoning strategy, which enables a completely self-contained, highly performant version of the ICLAD framework.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Zaharah Bukhsh and Aaqib Saeed. 2023. On out-of-distribution detection for audio with deep nearest neighbors. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv preprint*. ArXiv:2503.02857 [cs].

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore. Association for Computational Linguistics.

Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Lv Zhao, and Cunhang Fan. 2024. RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection. *arXiv preprint*. ArXiv:2406.06086 [cs].

Combei. 2024. *WavLM model ensemble for audio deepfake detection*. arXiv. ArXiv:2408.07414 [eess].

Joseph Cox. 2023. How i broke into a bank account with an AI-generated voice. `https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice`. Accessed: 2024-04-30.

Adrian de Wynter. 2025. Is in-context learning learning? *Preprint*, arXiv:2509.10414.

Yixin Dong, Charlie F Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. Xgrammar: Flexible and efficient structured generation engine for large language models. *Proceedings of Machine Learning and Systems 7*.

Noam Gat. 2023. lm-format-enforcer. `https://github.com/noamgat/lm-format-enforcer`. Accessed: 2025-10-03.

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for deepfake speech detection. *arXiv preprint arXiv:2506.21090*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. ArXiv:2507.08128 [cs].

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. 2025. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACMMM '25)*, New York, NY, USA. Association for Computing Machinery.

Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, and Shinji Watanabe. 2025. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 6:68–77.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-Audio Technical Report. *arXiv preprint*. ArXiv:2504.18425 [eess].

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25125–25148. PMLR.

Zhifeng Kong, Arushi Goel, Joao Felipe Santos, Sreyan Ghosh, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2025. Audio flamingo sound-cot technical report: Improving chain-of-thought reasoning in sound understanding. *Preprint*, arXiv:2508.11818.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in*

Neural Information Processing Systems, volume 36, pages 14005–14034. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2507–2522.

Juan M. Martín-Doñas and Aitor Álvarez. 2022. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245.

Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The Speakers in the Wild (SITW) Speaker Recognition Database. In *Interspeech 2016*, pages 818–822. ISCA.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2024a. Does Audio Deepfake Detection Generalize? In *Interspeech 2022*, pages 2783–2787. ISCA. ArXiv:2203.16263 [cs].

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024b. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620. ArXiv:1706.08612 [cs].

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Terry Pender. 2023. AI threatens courts with fake evidence, UW prof says. `https://www.jdsupra.com/legalnews/ai-threatens-courts-with-fake-evidence-7371356/`. Accessed: 2024-05-05.

Elizabeth Shriberg. 2005. Spontaneous speech: how people really talk and why engineers should care. In *Interspeech 2005*, pages 1781–1784. ISCA.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022a. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon Jung, Junichi Yamagishi, and Nicholas Evans. 2022b. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection. In *Interspeech 2024*, pages 537–541. ArXiv:2406.17376 [cs].

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, and 10 others. 2025. ASVspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech. *arXiv preprint*. ArXiv:2502.08857 [eess].

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. Asvspoof 5: crowdsourced speech

data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Yang Xiao and Rohan Kumar Das. 2025. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 32:1276–1280.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint*. ArXiv:2503.20215 [cs].

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint*. ArXiv:2109.00537 [eess].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio Deepfake Detection: A Survey. *arXiv preprint*. ArXiv:2308.14970 [cs].

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? In *Advances in Neural Information Processing Systems*, volume 36, pages 17773–17794. Curran Associates, Inc.

Yi Zhu, Heitor R. Guimarães, Arthur Pimentel, and Tiago Falk. 2025a. Auddt: Audio unified deepfake detection benchmark toolkit. *Preprint*, arXiv:2509.21597.

Yi Zhu, Surya Koppisetti, Trang Tran, and Gaurav Bharaj. 2025b. Slim: style-linguistics mismatch model for generalized audio deepfake detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yi Zhu, Saurabh Powar, and Tiago H. Falk. 2024. Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 139–143.

A Appendix

A.1 Potential Risks

The reliance of our framework on ALMs may introduce risks related to algorithmic bias. For example, biases that might be present in the ALMs' training data (*e.g.*, gender, accent, language) may unintentionally affect the detection process and certain user groups. However, the preliminary results obtained with the open-source AF3 model suggest that a fully transparent system capable of bias discovery and auditing is feasible for future implementation.

A.2 Listening Test Results

To investigate the reliability of ALM generated textual explanations and how they align with human perception, we conducted a systematic listening test to quantify model hallucination rate using the proposed Pairwise Comparative Reasoning (PCR). We evaluated the ALM's PCR outputs using 50 randomly selected audio samples (10 per benchmark dataset). We recruited 22 human annotators with verified experience in audio and speech processing. Each annotator was tasked with validating the PCR's reasoning against the actual audio characteristics. To ensure high-quality feedback, each annotator focused on a specific attribute: speech content, prosody/naturalness, room acoustics, physiological cues, or background noise/artifacts. Annotators chose between three labels: (i) no hallucination, (ii) hallucination (requiring a specific reason), or (iii) unsure. A hallucination was strictly defined as reasoning describing an audible event demonstrably absent from the sample. We employed an overlapping design, assigning 20 samples per person to ensure each sample received multiple expert labels. Following rigorous quality checks, 8 annotators were excluded due to missing data or failure to follow the hallucination definition, leaving a robust set of expert-validated labels.

A.2.1 Quantitative Analysis

We applied majority voting across annotations to determine the final hallucination status for each sample. The detailed distribution of these human labels per sample is provided in **Table 10**.

Overall, the results indicate that the PCR strategy significantly mitigates model errors. Only 10% of ALM responses using PCR contained hallucinations, a marked improvement over the 18% hallucination rate observed with a simple prompting strategy (as noted in Table 4 of the main manuscript).

Table 10: Human annotation distribution. Each sample received annotations from multiple experts to ensure robust and high-confidence labels.

Sample	% Hall.	% Not Hall.	% Unsure
9	69.2	30.8	0.0
11	57.1	42.9	0.0
3	50.0	37.5	12.5
28	50.0	50.0	0.0
5	33.3	66.7	0.0
1	28.6	71.4	0.0
26	28.6	71.4	0.0
8	28.6	71.4	0.0
15	27.3	72.7	0.0
13	25.0	75.0	0.0
18	25.0	50.0	25.0
2	25.0	75.0	0.0
16	25.0	62.5	12.5
30	25.0	75.0	0.0
23	22.2	77.8	0.0
14	22.2	77.8	0.0
21	20.0	80.0	0.0
27	20.0	80.0	0.0
22	20.0	60.0	20.0
20	18.2	81.8	0.0
17	10.0	80.0	10.0
12	10.0	90.0	0.0
7	9.1	90.9	0.0
24	7.7	84.6	7.7
19	0.0	100.0	0.0
10	0.0	100.0	0.0
25	0.0	100.0	0.0
6	0.0	83.3	16.7
4	0.0	66.7	33.3
29	0.0	100.0	0.0

This reduction demonstrates that the comparative framework effectively grounds the ALM's reasoning in actual acoustic evidence.

To understand the nature of the remaining errors, we categorized the reasons provided by annotators for "hallucinated" labels using keyword matching. As shown in **Table 5**, the primary category for disagreement is *Naturalness* (37.40%). Analysis of these cases reveals a consistent contradiction: ALMs frequently label scripted speech with a steady, robotic pace as "unnatural," regardless of whether the source is real or synthetic.

This specific failure mode explains why the ALM excels on in-the-wild spontaneous speech but struggles with scripted datasets like ASVspoof and MLAAD. These findings validate that while PCR reduces hallucinations, the model's internal bias toward "naturalness" in spontaneous speech remains a target for future alignment.

A.3 Statistical Significance Testing

To verify the observed performance differences between the baseline and our Gemini ICL frame-

Table 11: Paired t-test statistics for the accuracy comparison between the baseline and ICLAD. All results are statistically significant.

Dataset	t**-statistic**	p**-value**
21DF	8.97	$< .001$
MLAAD	61.11	$< .001$
ITW	-7.82	$< .001$
SpoofCeleb	-32.95	$< .001$
DFEval 24	-7.98	$< .001$

work, we performed paired t-tests on the results. As detailed in Table 11, the results are statistically significant ($p < .001$) across all five datasets.

A.4 Datasets

Table 12: Dataset details. No model training was performed. ICL examples were drawn from a database of 500 samples from the target's train split and 500 from 19DF. **Licenses**: ODC-By (Open Data Commons Attribution); Apache-2.0 (Apache License 2.0); CC BY 4.0 (Creative Commons Attribution 4.0); CC BY-SA 4.0 (Creative Commons Attribution-ShareAlike 4.0).

	Test Size		**RAG (Train)**	
Dataset	**Main**	**Abl.**		**License**
19DF	—	—	500	ODC-By
21DF	29,738	1,394	500	ODC-By
MLAAD	35,000	2,210	500	Apache-2.0
ITW	31,280	1,160	500	Apache-2.0
SpoofCeleb	18,226	1,200	500	CC BY 4.0
DFEval 2024	12,104	1,000	500	CC BY-SA 4.0

We evaluate our framework on five deepfake datasets representing two distinct conditions: scripted studio speech and challenging in-the-wild audio. For scripted studio datasets we use **ASVspoof 2021 (21DF)** (Yamagishi et al., 2021) and **MLAAD** (Müller et al., 2024b). 21DF contains **English** read-speech from the VCTK corpus. We use a subset of MLAAD that covers the eight languages present in both spoofed and real audio (German, Polish, English, French, Italian, Spanish, Russian, and Ukrainian). The in-the-wild datasets feature spontaneous speech from public figures. **In-the-Wild (ITW)** (Müller et al., 2024a) contains audio of **58 politicians and celebrities** collected from social networks and video platforms. **SpoofCeleb** (Jung et al., 2025) is built upon the VoxCeleb1 dataset, featuring voices from **1,251 celebrities**. The **DFEval 2024** (Chandra et al., 2025) benchmark contains content from 88 websites and **42 languages**, with its audio subset being 78.7% **English**. **ASVSpoof 2019 (19DF)** (Wang et al., 2020) is used exclusively to supplement the RAG database. Table 12 provides a detailed breakdown of data splits and licenses. We follow the intended use of all the datasets.

A.5 Instruction-Following Failures in Audio Flamingo 3

As noted in Section 5.5, Audio Flamingo 3 (AF3) exhibited significant limitations in following complex instructions for structured data generation. Table 13 provides several examples of these failure modes. In each case, AF3 was prompted to provide a reasoned analysis within a specific JSON schema. However, the model frequently produced outputs that violated the prompt's constraints, such as omitting required rationales or echoing schema placeholders verbatim.

We attempted to mitigate these issues using generation enforcement libraries like `lm-format-enforcer`(Gat, 2023) and `xgrammar`(Dong et al., 2024). While these tools forced AF3 to produce syntactically valid JSON, they could not prevent the model from oftentimes generating semantically illogical content.

Table 13: Examples of instruction-following and logical failures in AF3.

Omitted Rationale

The model returns a schema but leaves the required analytical fields empty.
Example: `{"Reconciled_Evidence": ""}`

Echoed Placeholders

Instead of generating new content, the model copies placeholder text from the prompt verbatim.
Example: `{"Final_Answer": "real | fake"}`

Format Violation

The model ignores a JSON-only instruction and outputs a free-form prose sentence.
Example: "The audio clip is real"

Illogical Content

The model produces semantically nonsensical and syntactically valid output. Examples:

- Returning the audio transcription
  Example: "Because men groping in the Arctic darkness had found a yellow metal"
- Filling competing hypothesis fields with the same irrelevant text.
- A non-meaningful response
  Example: "The audio clip is a recording of a human voice"

A.6 LLM Use in Manuscript Preparation

We used OpenAI's GPT-5 (via ChatGPT) to help with wording clarity and grammar edits. All scientific claims, experimental design, data analysis, and conclusions remain the responsibility of the authors.