From Scores to Evidence: Auditable Decisions Can Improve Speech Deepfake Detection

Anonymous authors
Paper under double-blind review

Abstract

One score per utterance does not show which component measurements supported a speech deepfake decision. We ask whether adding two absolute score differences to cross-fitted logistic fusion lowers equal error rate (EER): one difference compares the passive and watermark scores, and the other compares their average with retrieval. The decision record retains four primary scores and an auxiliary neighbor-distance field for post hoc diagnosis. The record stores a passive detector score on original audio, a watermark score on a marked copy, a retrieval vote from ten nearest support examples, and a speaker-profile margin based on nearest bona fide and spoof distances. It also stores distance to the nearest support example. For spoof examples, retrieval excludes the target synthesis family. On 4,080 ASVspoof 5 Track 1 development examples with all four primary scores, the absolute-difference model reaches 8.43% EER. Cross-fitted scalar fusion reaches 11.91%, and the best of 22 passive WavLM runs reaches 6.71%. Across eight held-out families, the mean fold EER reduction is concentrated in A11 and A12. The controls do not identify what causes the reduction. The fixed rule averages passive and watermark scores, then averages that result with retrieval. Its minimum detection cost is lower than the learned model's. In an in-sample, post hoc analysis of the same examples, sorting by distance to the final-score threshold places 88.66% of observed errors among the 33.75% of examples closest to the threshold. This analysis does not evaluate a deployment policy or analyst utility.

1 Introduction

Speech deepfakes can mimic a speaker's voice closely enough to deceive people and automated systems. That risk has motivated extensive work on speech deepfake detection. Yet most benchmarks and papers end with one score per utterance. A final score can rank utterances, but it does not show which component measurements supported or contradicted each decision. One example may have a large passive–retrieval raw-score gap; another may lack a probe altogether. If only the final scalar is retained, that distinction disappears.[1]

We retain those fields in an inspectable decision record: the component values remain available for post hoc diagnosis. We do not test whether a person can use them in an audit, establish source provenance or identity, or evaluate an operational review workflow. The matched analysis uses four measurements. The passive score is measured on the original audio. The probe score measures the response to a known watermark key in a derived audio copy. For spoof examples, the retrieval reference pool excludes the target synthesis family. The profile margin is computed from support examples. Raw neighbor closeness is retained as an auxiliary diagnostic. After these measurements are assembled, calibration produces one scalar score while leaving the component values available for later analysis.

The four scores come from different inputs and reference sets. The passive detector scores the original waveform. A keyed probe measures whether a known mark is detected in a derived copy; without the key, embedding rule, or copy, it is unavailable and does not authenticate the source (Wu et al., 2025; Özer et al., 2026). Retrieval and profile cues depend on support-pool coverage and speaker

[1]Anonymous 4open.science Git repository: `https://anonymous.4open.science/r/iclr27-speech-record-7d42/`. An MIT-licensed reproducibility source ZIP accompanies the submission.

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.

context (Liu et al., 2025), and the profile margin does not establish identity, prompt, channel, or recording-source independence.

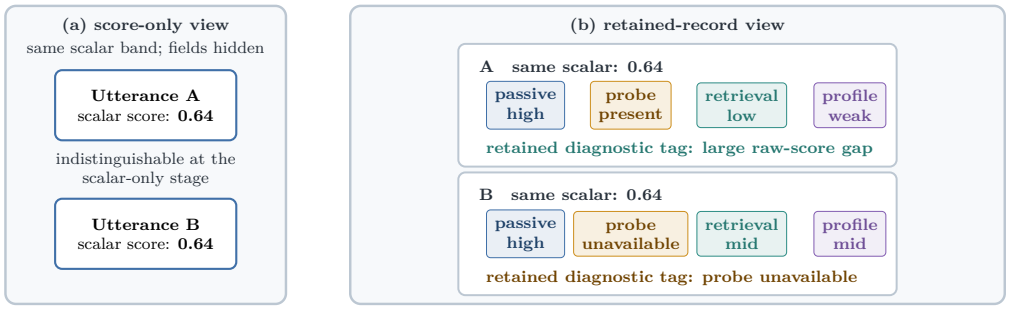

Figure 1: Two examples with the same schematic scalar score (0.64) retain different component measurements: a large passive–retrieval gap in A and an unavailable probe in B. The decision record keeps passive, probe, retrieval, and profile fields through calibration; raw neighbor closeness is auxiliary. Values are schematic, and the figure does not test deployment-time review.

Figure 1 shows two utterances in the same score band: one has a large passive–retrieval raw-score gap, while the other has an unavailable probe field. The performance question is whether adding two absolute differences to cross-fitted logistic score fusion lowers EER on the matched subset. Separately, the record keeps its component fields visible for post hoc diagnosis; the comparison does not show that field visibility causes the EER difference.

We evaluate this design on 4,080 ASVspoof 5 Track 1 development utterances for which all four cues are available. Tables 1 and 2 compare score performance on this subset. Appendix Figure 3 and Appendix Table 25 show which errors appear in a fixed-size post hoc review queue; they do not evaluate a deployment policy. WavLM is the strongest passive-only baseline on this subset. Our question is whether adding two absolute-difference features to cross-fitted logistic fusion lowers EER on this subset while the component measurements remain available for post hoc diagnosis.

Contributions. This paper makes three contributions. First, it specifies an utterance-level decision-record schema that keeps aligned passive, probe, retrieval, and profile fields visible alongside a final score. Second, on this subset, adding two absolute-difference features to cross-fitted logistic score fusion lowers EER relative to the otherwise identical model without those features. The fold-level reductions are concentrated in A11 and A12, and the controls do not identify what causes the difference. Third, it evaluates the record using held-out-family calibration, post hoc error labeling and ranking at fixed review budgets, and worked examples. The keyed probe is an auxiliary measurement only; the study does not establish watermark robustness, attribution, analyst utility, or deployment-time inspection decisions.

2 Related Work

ASVspoof evaluations compare passive detectors across synthesis families and conditions (Wang et al., 2026b;a; Ge et al., 2025a). RawNet2, AASIST, and self-supervised speech models provide scalar references (Tak et al., 2021; Jung et al., 2022; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2021); calibration work asks whether those scores are reliable as probabilities (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Our matched-subset analysis keeps the component measurements that a final scalar omits.

Proactive watermark studies separate mark detection from passive spoof detection and show why a probe depends on its key, embedding rule, and threat model (Wu et al., 2025; Ge et al., 2025b; San Roman et al., 2024). Self-voice conversion and watermark shortcuts further limit an interpretation of the probe as stand-alone class evidence (Özer et al., 2026; Müller & Debus, 2026). Retrieval and profile matching provide context from related synthesis examples and speakers (Liu et al., 2025;

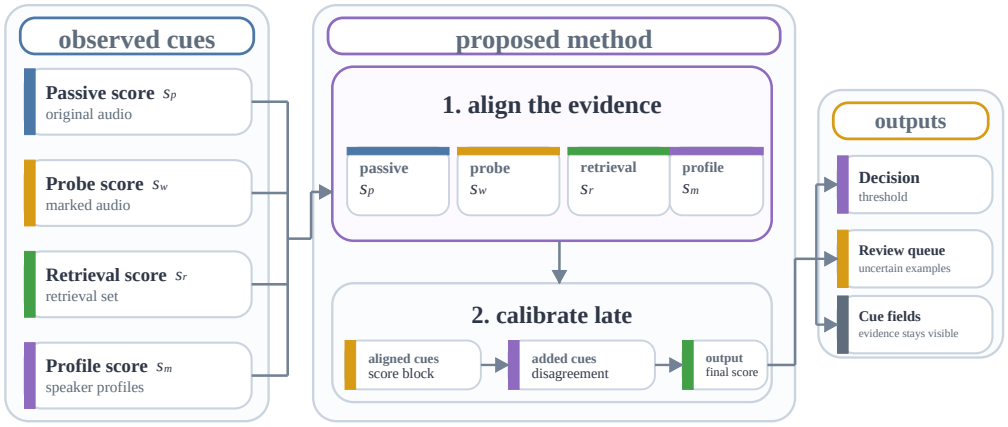

Figure 2: Method overview. The record keeps four measurements for each utterance: a passive detector score, a keyed probe score, retrieval support, and a profile margin. It also stores raw neighbor closeness. A logistic model combines the original scores with two absolute differences between selected scores to produce the operating score; the component measurements remain available for thresholding and post hoc diagnosis.

Zeng et al., 2024). Separate work on interpretable fusion and selective classification motivates inspecting decisions after scoring (Wang et al., 2024; Negroni et al., 2026; Geifman & El-Yaniv, 2017). Appendix Table 3 compares these approaches by retained fields and evaluation target. The present study evaluates one decision record and ranks errors on a matched subset with synthesis-family holdout; it does not introduce a new acoustic detector or validate an operational review policy.

3 Method

Each utterance is represented by a fixed-format decision record before final calibration. The record preserves what each stream says about the same utterance instead of replacing the streams with an early fused score. For an input segment x, larger values of class-oriented fields support bona fide speech or mark consistency. The record contains five scalar fields. The passive score $s_p(x)$ is the detector probability for bona fide speech on the original audio. The keyed spread-spectrum probe score $s_w(x)$ estimates whether the known watermark is present in a marked copy. The retrieval score $s_r(x)$ is the inverse-distance-weighted bona fide vote among $k = 10$ support neighbors under a k-nearest-neighbor (kNN) rule. The profile margin $s_m(x)$ compares distances to the nearest bona fide and spoof examples in the support set; larger values are oriented toward bona fide speech. The raw nearest-neighbor closeness $c_r(x)$ records how close the nearest support example is. We keep c_r unchanged and let the cross-fitted calibrator learn its association with the bona fide target.

These fields have different semantics. The passive detector scores the original utterance. The probe is meaningful only when a known key, the recorded embedding rule, and a watermarked copy of the utterance are available; otherwise its field is missing. Retrieval and the support-set profile margin provide support-pool and speaker-context evidence, while c_r records raw neighbor closeness without a pre-imposed class orientation. The profile field does not identify a speaker or certify prompt, channel, or recording-source independence. Calibration occurs only after the measurements are written side by side in the record. Appendix Table 16 summarizes what each field means and when it applies.

Figure 2 highlights the modeling choice. The original utterance feeds the passive detector, a watermarked copy feeds the keyed probe, and the held-out support set feeds retrieval and profile scoring. The learned step is not early fusion of all streams into one hidden score. Instead, late calibration uses the original score features together with two absolute differences between selected raw scores; thresholding and post hoc diagnosis can still use the underlying fields.

From these channels we form three fixed averages and two visible raw-score gap features. The fixed averages are passive–watermark f_{pw}, retrieval-augmented fusion f_{pwr}, and profile-augmented fu-

sion f_{pwrm}; the gap vector is $d = [|s_p - s_w|, |f_{pw} - s_r|]$. The cross-fitted calibrator is

$$s_{\text{rec}}(x) = \sigma\left(\beta_0 + \beta_z^\top z(x) + \beta_d^\top d(x)\right), \tag{1}$$

where $z = [s_p, s_w, f_{pw}, s_r, s_m, c_r, f_{pwr}, f_{pwrm}]$. The averages are fixed before evaluation: $f_{pw} = 0.5s_p + 0.5s_w$, $f_{pwr} = 0.5f_{pw} + 0.5s_r$, and $f_{pwrm} = 0.25(s_p + s_w + s_r + s_m)$.

The model adds only two raw-score differences to the original score features. The fixed averages are linear combinations of s_p, s_w, s_r, and s_m, so the eight values in z contain five linearly independent features once c_r is included. The main ablation tests whether adding the two gap features improves performance over logistic fusion of the original score features. Table 2 reports the explicit $\beta_d = 0$ ablation, passive-only transforms, no-watermark retrieval/profile controls, single-difference controls, a squared-difference parameterization, and product terms.

We use the following notation throughout. The "fixed retrieval-augmented rule" is f_{pwr}, the "profile-augmented rule" is f_{pwrm}, and "score fusion" is the same cross-fitted logistic protocol with $\beta_d = 0$. The learned decision-record family includes absolute-difference and squared-difference parameterizations of s_{rec}. Unless stated otherwise, worked examples and post hoc analyses use the absolute-difference variant because its retained coordinates are easier to interpret. Its small difference from the squared-difference variant is unresolved by the paired interval. The logistic calibrator standardizes features, initializes its intercept from the training-fold class prior, and uses an L_2 penalty of 10^{-3}. The calibrator learns the sign of c_r from its training fold. Appendix Table 6 separately shows what happens when the passive-score direction is chosen using each family's labels. All learned-model predictions come from calibrators that did not train on the scored examples. Spoof examples from a held-out synthesis family are evaluated with a calibrator trained on the other seven families; a stable hash of each utterance ID assigns bona fide examples to folds. The final record retains component scores, the operating score, calibration bin, nearest-neighbor metadata, watermark status, and consistency flags; field retention alone does not establish source provenance, identity, or operational inspection utility.

We test whether auxiliary measurements reduce errors on held-out families and whether the stored fields help rank errors for post hoc inspection. WavLM-large supplies the scalar-detector reference, and the decision record keeps complementary evidence visible.

Data-handling boundary. The retained field list is an analysis schema, not a release specification. This manuscript does not establish permission to redistribute audio, speaker-linked identifiers, embeddings, neighbor metadata, or keys. Any reproduction artifact must follow the source-corpus terms and minimize or withhold these fields when redistribution is not authorized.

4 Experimental Protocol

Data and splits. We match passive/probe scores to retrieval/profile scores by utterance ID and hold out synthesis families A09–A16. The resulting ASVspoof 5 Track 1 development subset contains 4,080 examples from 705 speakers: 3,215 bona fide and 865 spoof utterances. All central comparisons use these same examples; larger score sets provide context only. For each family, the matched retrieval EER falls within the bootstrap interval obtained by retaining the same number of bona fide and spoof examples from the full score set. For A13, matched EER (19.19%) minus the random-retention bootstrap estimate (23.49%) gives an absolute EER difference of -4.30%, the largest magnitude across families (Appendix Tables 17 and 18). These checks do not make the matched subset representative of all retrieval predictions.

Fixed score streams. The main passive stream is a compact spectrogram convolutional neural network (CNN) with four-second, 16 kHz input. Frozen self-supervised learning (SSL) models, including WavLM-large, provide stronger passive references on the same matched examples (Appendix Tables 10 and 11). The probe uses one development key and a −32 dB spread-spectrum mark in the 3.0–7.6 kHz band on a derived track. The same watermarking condition is applied to examples from both classes. The probe therefore measures a response to that derived input; it is not independent evidence of class. The main comparison uses the HuBERT-large 20k default retrieval setting with $k = 10$, 38,797 support embeddings, and 40,000 predictions before matching on all four cues. For each spoof family, its support examples are removed before scoring (Liu et al., 2025). The

retrieval-overlap check finds no exact query, speaker, or audio-path reuse and no held-out-family top-10 neighbor among 20,000 spoof queries (Appendix Table 19). The only source tag is corpus-level, leaving prompt, channel, and recording-source independence unresolved. Appendix Table 4 fixes the full protocol. Appendix Table 33 lists 13 configurations before matching on all four cues. Those scores do not establish which setting would perform best on the matched examples.

Metrics and uncertainty. Equal error rate (EER) uses one pooled threshold unless a table specifies the mean over held-out-family or calibration folds. We also report expected calibration error (ECE) from 15 equal-width probability bins and bona fide-target minimum detection cost (minDCF; $P_{\text{tar}} = 0.05$, $C_{\text{miss}} = 1$, $C_{\text{fa}} = 10$). Because the cost depends on the target convention, we emphasize EER, ECE, and paired intervals; Appendix Table 7 reverses the target. Matched EER intervals use 5,000 paired class-stratified resamples, with fold- and speaker-resampled checks for learned comparisons (Appendix Tables 27 and 29). Source-level resampling is unavailable with the coarse source tag.

5 Results

Main matched comparison. All values in this paragraph refer to the 4,080-utterance matched subset. Table 1 reports the HuBERT-large 20k default on this subset. The retrieval sweep in Appendix Table 33 uses examples before matching on all four cues. Its scores are not directly comparable with the matched results. The family-wise EER column averages eight evaluations, each holding out one synthesis family while sharing a bona fide pool. The WavLM entry reports the best run from a 22-run matched passive sweep; its median is 9.60%. The fixed rule's family-macro advantage over the passive CNN largely disappears when A12 and A13 are excluded (Appendix Table 21); the passive CNN is anti-oriented on A12 and near chance on A13 (Appendix Table 6). The fixed retrieval-augmented rule lowers pooled EER from 15.84% to 11.91% and family-macro EER from 13.85% to 10.27%. The fixed rule improves ranking but not calibration: ECE rises from 0.1451 to 0.2609. The learned cross-fitted record reaches 8.43% EER with a 0.0709 ECE point estimate on the same setup. The record reaches 8.43% EER, versus 11.91% for both the fixed retrieval-augmented rule and cross-fitted scalar fusion. Its absolute EER difference from either baseline is -3.48% (8.43% - 11.91%). Paired 95% confidence intervals are [-4.61%, -2.08%] for the fixed rule and [-4.60%, -2.34%] for scalar fusion (Appendix Table 27). This result is stronger than the WavLM sweep median, but it does not exceed the best passive WavLM run at 6.71% EER.

The other auxiliary channels are weaker. Nearest-distance retrieval reaches 62.68% EER on its own, so raw closeness is not enough. The probe is also weak as a stand-alone score: probe-only EER is 48.67%, and passive + probe remains close to the passive CNN (21.03% versus 21.15%). Appendix Table 13 shows the same pattern after removing s_w: pooled EER rises only slightly from 11.91% to 12.46%. We retain the probe as a conditional, inspectable field. These experiments do not establish an independent predictive benefit from the probe.

Adding the profile margin lowers family-macro EER from 10.27% to 9.86% and $\text{minDCF}_{\text{BF}}$ from 0.8283 to 0.7673. It also raises pooled EER from 11.91% to 12.96% and ECE from 0.2609 to 0.3270. The matched results show that retrieval accounts for most of the fixed-rule improvement; adding the profile margin shifts family-macro EER and minDCF but raises pooled EER and ECE. The learned record ranks errors differently after scoring, while the controls leave the source of its score difference unresolved.

Interpretation boundary. Here, "disagreement" and "conflict" refer only to numerical gaps between selected raw scores. These gaps do not make the scores measure the same quantity, and the experiments do not show why performance changes. The ECE values are out-of-fold point estimates. We use the paired intervals in Appendix Table 28 only for its named control contrasts, not to create a global calibration ranking.

Table 2 compares additional features after passive and retrieval scores are included. Fold EER averages the eight calibration folds. It should be compared with other fold-level controls in the same table, not with the held-out-family EER column in Table 1. The direct comparison is the explicit $\beta_d = 0$ score-fusion ablation of Eq. 1. This control improves on the fixed mean, reaching 11.91% EER and a 0.0840 ECE point estimate. The proposed record lowers ECE by 0.0130 with a paired 95% interval of [-0.0159, -0.0090] against this named control (Appendix Table 28). The weaker controls identify

Table 1: ASVspoof 5 comparison on the 4,080 development utterances where passive, probe, retrieval, and profile measurements are all available. Lower EER and minDCF are better. The highlighted fixed retrieval-augmented rule is the main fixed-fusion comparison; Table 2 tests whether two absolute-difference features improve on logistic fusion of the original scores.

System	Pooled EER (%) ↓	Family EER (%) ↓	$\text{minDCF}_{\text{BF}}$ ↓	ECE ↓
Passive CNN	21.15	24.44	1.0000	0.2826
Passive WavLM-large (best of 22)	6.71	5.33	0.9776	0.0453
Watermark probe	48.67	49.08	0.9972	0.4889
Passive + watermark probe	21.03	24.39	0.9984	0.3755
Learned passive + watermark	21.03	24.61	0.9984	0.0663
Retrieval kNN	15.84	13.85	0.9406	0.1451
Distance-only retrieval	62.68	65.38	1.0000	0.1708
Profile margin	34.58	38.01	0.9944	0.2833
Fixed retrieval-augmented rule	**11.91**	**10.27**	**0.8283**	**0.2609**
Fixed retrieval + profile	12.96	9.86	0.7673	0.3270

what does not explain the gain. Passive shape alone remains poor: $|s_p - 0.5|$ gives 24.04% EER, and s_p with s_p^2 gives 21.50%. Retrieval with profile but without the passive score reaches 16.64% EER. Adding passive to retrieval without a watermark field reaches 11.67% EER, indicating that most class information already lies in the passive and retrieval channels. The controls do not isolate the source of the remaining difference or establish a separate contribution from the probe. The no-watermark scalar expansion with s_p^2 and $|s_p - 0.5|$ reaches 8.67% EER. The absolute EER difference between the proposed record and this control is -0.24% (8.43% - 8.67%; 95% CI [-0.81%, +0.46%]). The two single-gap controls reach 8.90% for $|s_p - s_w|$ and 8.78% for $|f_{pw} - s_r|$. As a negative control, fold shuffling breaks the example-level link in s_w while preserving each fold's watermark-score distribution. It reaches 8.78% EER, only 0.35% from the full model. The no-watermark and shuffled-probe controls leave the source of the difference unresolved: nonlinear score transforms and information shared with other fields remain possible explanations.

The learned record adds two absolute differences between raw scores. The squared-difference model gives the lowest point estimate at 8.33% EER and 0.0709 ECE. We report the absolute-difference variant in the main analysis because its coordinates are easier to read in worked examples and post hoc diagnostics. The absolute EER difference between the proposed record and squared-difference model is +0.10% (8.43% - 8.33%). Its paired 95% CI [-0.26%, +0.57%] includes zero (Appendix Table 27). The paired absolute ECE difference from the squared-difference model is +0.0001, with a 95% CI of [-0.0016, +0.0021] (Appendix Table 28). The no-watermark nonlinear control remains near the full model, as quantified above. Speaker-resampled intervals in Appendix Table 29 favor the absolute-difference model over the product-term nonlinear control. Their absolute EER difference (record minus control) is -1.72% (8.43% - 10.16%; 95% CI [-2.72%, -0.65%]), computed from unrounded EER values. These controls do not isolate raw-score gaps as the source of the performance change.

We use the absolute-difference model for worked examples because its raw-score differences are easy to interpret. It is not a conclusively better scoring rule: its EER is close to the squared-difference and no-watermark nonlinear variants, and the controls do not identify how the score streams interact.

Where performance changes. The held-out-family results show that the improvement is concentrated, not uniform. Relative to cross-fitted score fusion, the absolute-difference model lowers mean fold EER by 2.02%, from 9.23% to 7.21%. Per-family changes range from -0.12% on A14 to +6.57% on A12. The reductions on A11 (5.11%) and A12 (6.57%) contribute $(5.11 + 6.57)/8 = 1.46\%$ to the 2.02% absolute EER reduction averaged over the eight held-out families.

A12 illustrates where the model with absolute-difference features differs most from scalar fusion. Appendix Table 6 shows that spoof examples receive higher passive bona fide probabilities than bona fide examples on average (0.945 versus 0.627). The passive score consequently yields 81.31% EER, whereas retrieval is correctly oriented at 2.03%. In Table 30, the absolute-difference model lowers the same family's fold EER from 13.01% to 6.44%. Appendix Table 31 repeats the learned fits over ten alternative assignments of bona fide examples while keeping spoof folds fixed. The absolute-difference model's pooled EER ranges from 8.33% to 8.55%, while score fusion ranges from 11.69% to 12.05%. These runs preserve the pooled EER difference but do not establish that

Table 2: Cross-fitted calibration ablation on the 4,080 matched examples. Pooled EER uses one global threshold. The controls compare the two absolute-difference features with transformations of the passive and retrieval scores; they do not identify the source of a performance difference.

Setting	Feature set	EER (%) ↓	Fold EER (%) ↓	$\text{minDCF}_{\text{BF}}$ ↓	ECE ↓	Brier ↓
Fixed retrieval rule	fixed mean	11.91	10.04	0.8283	0.2609	0.1415
Cross-fitted scalar fusion	$\beta_d = 0$ scalar scores	11.91	9.23	0.9107	0.0840	0.0797
Passive margin only	$\lvert s_p - 0.5 \rvert$	24.04	23.92	0.9991	0.2064	0.1824
Passive shape control	s_p plus s_p^2	21.50	24.14	1.0000	0.2305	0.1451
Retrieval + profile only	no probe	16.64	13.58	0.9978	0.1385	0.1129
Passive + retrieval only	no probe	11.67	9.26	0.9238	0.0904	0.0797
No-probe nonlinear control	passive + retrieval shape	8.67	7.34	0.9984	0.0756	0.0645
Passive-probe gap	scalar + $\lvert s_p - s_w \rvert$	8.90	7.99	0.9611	0.0761	0.0680
Shuffled-probe control	fold-shuffled W + gap	8.78	7.99	0.9652	0.0759	0.0673
Fusion-retrieval gap	scalar + $\lvert f_{pw} - s_r \rvert$	8.78	7.30	0.8591	0.0752	0.0691
Squared-difference record	squared score differences	8.33	6.78	0.9742	0.0709	0.0631
Product-term control	score products	10.16	8.00	0.8986	0.0780	0.0751
Proposed decision record	**absolute gaps**	**8.43**	**7.21**	**0.9365**	**0.0709**	**0.0640**

each family-level change is stable. These results associate the largest performance change with a passive failure mode; they do not identify its mechanism.

Passive-reference context. The stronger passive baseline narrows the interpretation. Under the same cross-fitted protocol on the same WavLM-aligned examples, the WavLM decision-record model reaches 9.14% EER, compared with 11.79% for passive calibration alone. Against the broader passive sweep, however, the best of 22 WavLM-large runs still reaches 6.71% EER, while the sweep median is 9.60%. The record has lower EER than the same-protocol passive calibration control (11.79%) and the passive-sweep median (9.60%), but higher EER than the best passive run (6.71%). At fold level, the ordering reverses: the record reaches 6.56% EER versus 5.00% for passive cross-fitted calibration. We use WavLM as the scalar reference for exactly that reason.

The 4,080-example matched WavLM analysis in Appendix Tables 10 and 11 supplies the passive reference used here. Appendix Tables 15 and 34 support the narrow interpretation of the probe: direct marked/unmarked pairs separate cleanly, whereas source-pair and self-voice-conversion controls remain near chance. The probe is therefore a conditional measurement tied to the watermarked copy, key, and threat model, not a stand-alone detector.

Post hoc error ranking. For this descriptive analysis, each system's EER threshold is estimated on the evaluated matched set, as stated in Table 23. The analysis compares how many observed errors each score ranks within a fixed-size review queue; it does not test a deployment-time inspection policy. In the HuBERT retrieval-and-profile setting, the cross-fitted absolute-difference record ranks its observed errors most effectively. In the first 10% of the ranked queue, it captures 47.97% of its own errors, compared with 38.89% for the fixed retrieval-augmented rule and 29.26% for retrieval kNN (Table 23).

Pooled EER does not determine which errors appear near the top of this queue. After excluding the highest-ranked 10%, the residual error rate is 4.87%, compared with 8.09% for the fixed retrieval-augmented rule and 12.45% for retrieval kNN. The fixed rule is slightly more precise at the same budget (46.32% versus 40.44%), but the learned record pulls more total mistakes toward the inspection boundary.

The WavLM analysis also separates detection from post hoc error ranking. The best passive WavLM run reaches 6.72% full-set error. Adding auxiliary fields does not improve that decision rule, but the corresponding decision-record model still captures 50.40% of its own errors in the first 10% of the ranked queue. The record therefore orders observed errors differently even when it is not the best detector.

Appendix Figure 3 separates three post hoc questions: which cues help before calibration, how well the learned score is calibrated, and what the visible cues add at a fixed queue size. In panel (a), threshold margin is the strongest pre-calibration triage signal. At a 20% review load, it captures 61.5% of fixed-rule errors. Nearest-neighbor distance captures 24.3%, and the raw-score gap captures 18.5%, compared with 20.0% for a random queue. Margin therefore gives the clearest pre-calibration ordering; distance and the raw-score gap remain secondary warning signals.

Panel (b) uses the same 15 equal-width bins as the ECE calculation in Tables 1 and 2. All 15 cross-fitted bins sit above the diagonal, indicating mild under-confidence on bona fide examples. The largest gap occurs in the bin centered near 0.436, where the mean score is 0.436 and the empirical bona fide rate is 0.833. Appendix Table 32 repeats the calculation with equal-mass bins; for the cross-fitted model, ECE changes only slightly, from 0.0709 to 0.0715. A foldwise isotonic refit lowers it further to 0.0178.

At a fixed review load, we compare the four-cue union with score-margin queues by the fraction of the record's observed errors each queue captures and by its precision. Appendix Table 24 defines these four cues: record-threshold proximity, passive–record decision mismatch, retrieval–record decision mismatch, and a fusion–retrieval gap. Appendix Figure 3(c) compares them at a fixed budget, Appendix Figure 3(d) breaks down covered errors, and Appendix Table 25 gives the exact values. At a shared review load of 1,377 examples (33.75%), the four-cue union covers 82.85% of the calibrated rule's observed errors at 20.70% precision; the record-margin queue covers 88.66% at 22.15% precision. Passive-margin and retrieval-margin queues cover 32.85% and 40.99%, respectively. The cue union retains which fields flagged each example. At this fixed review load, the score-margin queue finds more errors; the cue union instead shows which component checks triggered for each selected example. This analysis does not establish that the cue union improves a deployment-time review policy.

Among the 285 covered errors, 177 trigger at least two cues (62.1%), as shown in Appendix Figure 3(d). Single-cue cases are mostly ordinary near-threshold decisions: 13 are flagged only by threshold proximity. Appendix Figure 4 shows two matched examples: a correctly classified spoof whose passive score favors bona fide, and a spoof error for which the retrieval score favors bona fide. In the worked records, NN denotes the nearest neighbor.

In the first worked example, C1 is a correctly identified spoof from family A12. Its passive score remains on the bona fide side (1.00), while retrieval is 0.00 and the calibrated score is 0.28, below threshold. C2 shows the opposite failure in family A13: passive evidence points toward spoof (0.07), retrieval points toward bona fide (1.00), and the raw-score gap remains large ($|f_{pw} - s_r| = 0.82$) although the final decision is wrong.

The paired bootstrap intervals preserve the same ordering. The fixed retrieval-augmented rule lowers pooled EER from 21.15% for the passive CNN and 15.84% for retrieval kNN to 11.91%. The fixed rule's absolute EER difference from passive CNN is -9.24% (11.91% - 21.15%). Against retrieval kNN, the absolute difference is -3.92% (11.91% - 15.84%). Both use unrounded EER values (Appendix Table 27). The learned absolute-difference model has lower EER than score fusion, the single-difference controls, and the product-term nonlinear control, while remaining statistically tied with the squared-difference variant. These paired intervals are not adjusted for multiple comparisons, and the controls do not isolate which fields or nonlinear terms account for the difference. Appendix Table 7 shows that $\text{minDCF}_{\text{BF}}$ still favors the fixed retrieval-augmented rule (0.8283 versus 0.9365). We therefore use the fixed rule for threshold-oriented discussion and use learned models only for post hoc error ranking, calibration, and diagnosis.

Family-level variation. Family-level comparisons show that the pooled gain is concentrated, not uniform. Against the passive CNN, the fixed retrieval-augmented rule improves on A12, A13, A14, and A16, is unresolved on A09, A10, and A15, and regresses on A11. A12 is the largest single-family contributor in this error accounting. Appendix Table 6 shows that reversing the passive score direction within each family using its labels lowers family-macro EER from 24.44% to 15.86%.

The error accounting tells the same story. Bona fide examples contribute 297 of the 377 net fewer errors, A12 and A13 contribute 90 and 38, and A11 plus A15 offset the gain by 30 and 12 errors. If A12 and A13 are removed, the family-macro EERs become nearly identical (10.20% for the passive CNN and 10.28% for the fixed retrieval-augmented rule). We therefore interpret the pooled

improvement as recovery from a small set of passive failure modes, not as a uniform advantage over the passive stream. Appendix Figure 5 and Appendix Tables 22 and 21 give the full breakdown and paired intervals.

Component ablations. The separate 10,000-example passive/probe experiment finds no pooled benefit. Passive scoring reaches 22.42% EER, versus 22.54% for fixed fusion and 22.56% for learned fusion; both paired intervals cross zero (Appendix Table 14). The A12 condition reaches 82.66% EER for learned fusion against the reused bona fide pool, although the three systems have similar pooled EERs. The pooled comparison therefore obscures this condition-specific error rate. This separate setting differs from the four-cue matched record and does not establish a predictive contribution from the probe.

6 Conclusion

On the 4,080-example matched ASVspoof 5 Track 1 development subset, the learned decision record reaches 8.43% EER, below cross-fitted scalar fusion (11.91%) but above the best passive WavLM run (6.71%). The fixed retrieval rule has lower minimum detection cost (minDCF) than the learned model. The EER reduction over scalar fusion is concentrated in A11 and A12, which together contribute 1.46% to the 2.02% mean fold EER reduction across eight held-out families. Nearby controls do not identify its source: the no-probe nonlinear model reaches 8.67% EER and the squared-difference model reaches 8.33%, with paired intervals that include zero against the learned record. The keyed probe depends on a known key and a marked copy; these results do not show a separate predictive benefit from the probe.

The fixed-rule comparisons also depend on the metric. On this subset, the fixed rule lowers pooled EER from 21.15% for the passive CNN and 15.84% for retrieval kNN to 11.91%, and lowers minDCF from 1.0000 and 0.9406 to 0.8283 (Table 1). Its ECE is higher than retrieval kNN's (0.2609 versus 0.1451), while the best passive WavLM run has lower EER at 6.71%. Adding the profile margin lowers family-macro EER from 10.27% to 9.86% and minDCF from 0.8283 to 0.7673, but raises pooled EER from 11.91% to 12.96% and ECE from 0.2609 to 0.3270. No system is best on every reported criterion. These metrics answer different questions: EER balances false accepts and false rejects, minDCF searches thresholds under the specified cost convention, and ECE summarizes calibration across score bins. Their disagreement reflects a trade-off among ranking, operating cost, and calibration; it does not yield a universal ordering of systems. Application-specific thresholds require separate validation under the intended costs.

The error-ranking analysis addresses a different question from detection accuracy. At the shared review load of 1,377 examples (33.75% of the matched set), the score-margin queue covers 88.66% of the learned record's observed errors at 22.15% precision. The four-cue union covers 82.85% at 20.70% precision and retains which cues flagged each example (Appendix Table 25). Both summaries use labels from the same development examples to identify errors. They describe post hoc rankings, not a prospective inspection policy, and do not show whether analysts would find either queue useful.

Requiring all four measurements also selects the evaluated examples. In the retrieval sensitivity analysis, A13 has 19.19% matched-subset EER versus a same-size random-retention bootstrap estimate of 23.49%, an absolute EER difference of -4.30% (matched minus estimate; Appendix Table 18). This is the largest family-level shift in that analysis; the matched subset therefore does not represent all retrieval predictions. The 8.43% versus 11.91% comparison tests adding two absolute-difference features: both models receive the underlying score fields, so it does not establish a classification benefit from displaying component values. Taken together, the findings support a limited conclusion: retaining aligned component scores enables post hoc diagnosis, while adding those features lowers EER relative to the specified scalar-fusion control on this selected development subset. Performance on the official evaluation set and other corpora remains unmeasured, and the diagnostic rankings require prospective evaluation with analysts before they can inform an inspection workflow.

Reproducibility Statement

Sections 3–4 and Appendix Tables 4, 5, and 19 specify score construction and evaluation. The MIT-licensed supplementary source ZIP provides code for score-level analyses and a synthetic validation path, but it does not regenerate detector scores because licensed speech audio, checkpoints, full score files, and watermark keys are not redistributed.

AI Use Statement

We used generative AI to help polish the writing and to find related work. The paper describes these uses.

References

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing, 2021.

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. Post-training for deepfake speech detection. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8, 2025a. doi: 10.1109/ASRU65441.2025.11434709.

Wanying Ge, Xin Wang, and Junichi Yamagishi. FakeMark: Deepfake speech attribution with watermarked artifacts, 2025b.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367–6371, 2022. doi: 10.1109/ICASSP43922.2022.9747766.

Xuechen Liu, Xin Wang, and Junichi Yamagishi. Zero-day audio deepfake detection via retrieval augmentation and profile matching, 2025.

Nicolas M. Müller and Pascal Debus. The watermark shortcut: How provenance marking sabotages audio deepfake detection, 2026.

Viola Negroni, Xin Wang, Wanying Ge, Paolo Bestagini, Junichi Yamagishi, and Stefano Tubaro. Toward interpretable speech deepfake detection using artifact-specific experts and calibrated detection scores, 2026.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430.

Yigitcan Özer, Wanying Ge, Zhe Zhang, Xin Wang, and Junichi Yamagishi. Self voice conversion as an attack against neural audio watermarking, 2026.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43180–43196, 2024.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6369–6373, 2021. doi: 10.1109/ICASSP39728.2021.9414234.

Xin Wang, Tomi Kinnunen, Kong Aik Lee, Paul-Gauthier Noé, and Junichi Yamagishi. Revisiting and improving scoring fusion for spoofing-aware speaker verification using compositional data analysis, 2024.

Xin Wang, Héctor Delgado, Nicholas Evans, Xuechen Liu, Tomi Kinnunen, Hemlata Tak, Kong Aik Lee, Ivan Kukanov, Md Sahidullah, Massimiliano Todisco, and Junichi Yamagishi. ASVspoof 5: Evaluation of spoofing, deepfake, and adversarial attack detection using crowdsourced speech, 2026a.

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, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95: 101825, 2026b. doi: 10.1016/j.csl.2025.101825.

Chia-Hua Wu, Wanying Ge, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. A comparative study on proactive and passive detection of deepfake speech. In *Proceedings of INTERSPEECH 2025*, pp. 2595–2599, 2025. doi: 10.21437/Interspeech.2025-1419.

Chang Zeng, Xiaoxiao Miao, Xin Wang, Erica Cooper, and Junichi Yamagishi. Spoofing-aware speaker verification robust against domain and channel mismatches, 2024.

A Additional Evidence

A.1 Supporting Comparisons and Post Hoc Examples

Appendix Table 3 compares the related detector, probe, retrieval, and diagnostic studies that motivate the retained fields. Appendix Table 4 summarizes the fixed matched-subset protocol.

Table 3: Closest-work comparison. The fields column describes values retained for a final score or diagnosis; it does not imply a human review process.

Work	Task / data	Method or retained fields	Evaluation target	Relationship to this study
Wu et al. (Wu et al., 2025)	Proactive and passive speech-deepfake detection	Passive detector and keyed-mark system scores	System-level proactive/passive comparison	Here the keyed probe is one auxiliary field; its weak stand-alone signal is reported explicitly.
Liu et al. (Liu et al., 2025)	Zero-day detection on DeepFake-Eval-2024	Training-free retrieval augmentation and profile matching	Detection and cross-database generalization	Here retrieval and profile values remain visible in a fixed ASVspoof 5 matched decision record.
Negroni et al. (Negroni et al., 2026)	Interpretable speech-deepfake detection	Trained artifact-specific experts and calibrated log-likelihood ratios	Expert semantics plus aggregate classification	Here operational raw scores and gaps remain visible; the record does not infer named artifact identities.
This work	4,080 matched ASVspoof 5 development examples	Passive, keyed probe, retrieval, profile, and raw-score difference fields	Out-of-fold score evaluation and post hoc diagnosis	Does not establish source provenance, a human review policy, or deployment utility.

Appendix Figure 3 reports the post hoc error-ranking, reliability, and fixed-load analyses discussed in Section 5. Appendix Figure 4 shows a correctly classified spoof whose passive score favors bona fide and a spoof error for which the retrieval score favors bona fide.

The appendix collects the protocol details, passive-reference comparisons, boundary checks, and stress analyses that support the main text.

A.2 Matched Subset and Operating Convention

The matched development subset contains 4,080 utterances: 3,215 bona fide and 865 spoof examples from 705 speakers. Table 5 gives its family and source-tag counts. Table 6 shows that passive-score polarity varies across families, which motivates learning its direction only inside each training fold.

Table 5: Coverage of the score subsets used in the paper. BF/S gives bona fide (BF) and spoof counts; source-tag counts use broad corpus labels. Duration is computed from referenced audio files when available.

Subset	Examples	Utter.	BF/S	Spk.	Source tags	Families	Duration
Matched main	4,080	4,080	3,215/865	705	1	8	8.01 h
Retrieval before cue matching	40,000	40,000	20,000/20,000	n/a	n/a	8	not available
WavLM matched	4,080	4,080	3,215/865	705	1	8	8.01 h

Table 4: Matched held-out-family protocol used in the main comparisons.

Component	Specification
Matched examples	ASVspoof 5 Track 1 development split; 4,080 matched examples with 4,080 unique utterances and 705 speakers (3,215 bona fide / 865 spoof); 8 held-out A09–A16 folds with 91–124 spoof examples each. Audio-path metadata covers 8.01 h over 4,080 files.
Passive score	Archived compact spectrogram-CNN bona fide probability for every matched comparison; the stored score is treated as a fixed input stream. No per-family score-polarity correction is applied, so family failures stay visible in Table 20. Stronger frozen-SSL passive comparisons appear in the appendix.
Watermark score	Spread-spectrum presence probability with one development key, strength -32.0 dB, and 16 kHz audio. The matched fusion examples all carry the same derived watermark condition across bona fide and spoof labels, so this stream tests keyed-probe behavior and a raw-score feature while keeping watermark presence independent of the class label.
Retrieval/profile	HuBERT-large archive; $k = 10$ neighbor-label retrieval; 38,797 support embeddings and 40,000 predictions before matching on availability of all four cues from our reimplementation, distinct from the published paper's reported scores. The identifier-overlap analysis finds 0/20,000 spoof examples with any top-10 neighbor from the held-out family, with 0 query utterance-ID matches, 0 speaker ID matches, and 0 exact audio-path matches across the top-10 list; source tags are corpus-level fields and are not used to certify prompt or recording-source separation.
Calibration	8 leave-family-out folds; 10 features including passive–watermark and fusion–retrieval raw-score gaps; seed 20260821; intervals use 5,000 bootstrap resamples.

Table 6: Passive-score orientation analysis for the held-out synthesis-family protocol. The gap is the mean passive bona fide probability on bona fide examples minus that on spoof examples; negative gaps indicate anti-oriented scores. Inverted EER uses the deterministic score complement.

Family	Spoof	$\bar{p}_{\mathrm{BF}}^{\mathrm{bon}}$	$\bar{p}_{\mathrm{BF}}^{\mathrm{spoof}}$	Gap	Passive EER (%) ↓	Inv. EER (%) ↓	Call
A09	111	0.627	0.088	+0.539	9.92	90.08	original
A10	124	0.627	0.072	+0.555	8.73	91.27	original
A11	121	0.627	0.060	+0.568	6.43	93.57	original
A12	91	0.627	0.945	-0.318	81.31	18.69	anti-oriented
A13	100	0.627	0.522	+0.105	53.00	47.00	near chance
A14	102	0.627	0.062	+0.565	5.65	94.35	original
A15	116	0.627	0.285	+0.342	23.27	76.73	original
A16	100	0.627	0.077	+0.550	7.22	92.78	original

Table 7 reports minDCF under the bona fide target convention used throughout; values near 1.0 indicate little reduction under that default cost.

Table 7: Cost-orientation sensitivity for the 4,080 matched examples. $\mathrm{minDCF}_{\mathrm{BF}}$ treats bona fide speech as the target class; $\mathrm{minDCF}_{\mathrm{spoof}}$ inverts labels and scores. Values near 1.0 indicate little reduction from the normalized default-cost reference.

Group	Score	EER (%) ↓	$\mathrm{minDCF}_{\mathrm{BF}}$ ↓	$\mathrm{minDCF}_{\mathrm{spoof}}$ ↓
Main ablations	Passive CNN	21.15	1.0000	0.8579
Main ablations	Watermark probe	48.67	0.9972	0.9988
Main ablations	Fixed record	21.03	0.9984	0.9549
Main ablations	Retrieval kNN	15.84	0.9406	0.9988
Main ablations	Fixed record + retrieval	11.91	0.8283	0.6659
Main ablations	Profile fusion	12.96	0.7673	0.6974
Main ablations	Distance-only control	62.68	1.0000	1.0000
Calibration	Fixed record + retrieval	11.91	0.8283	0.6659
Calibration	Cross-fitted score fusion	11.91	0.9107	0.7295
Calibration	Cross-fitted nonlinear control	10.16	0.8986	0.7618
Calibration	Cross-fitted absolute-difference record	8.43	0.9365	0.6913

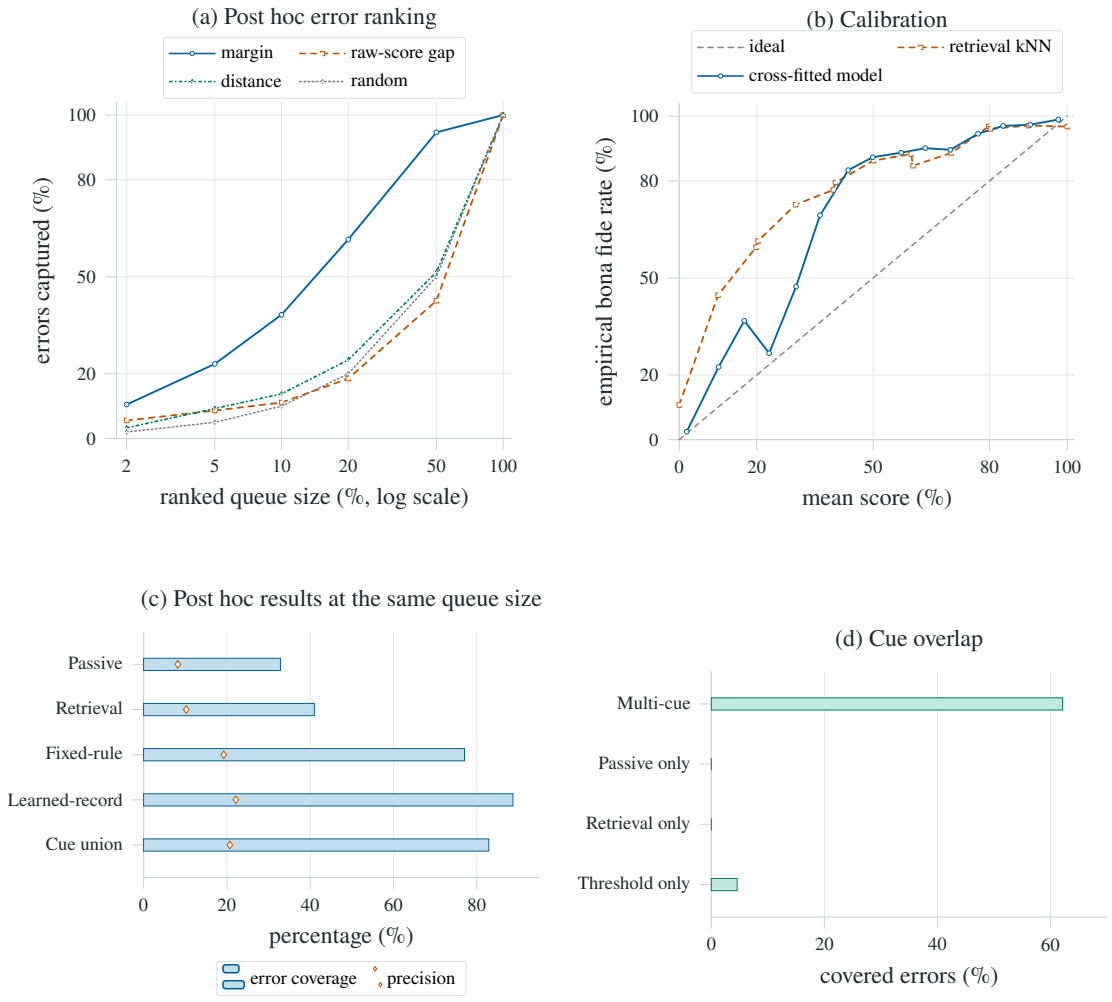

Figure 3: Post hoc calibration and error analysis on the matched examples. (a) Each curve shows the share of fixed-rule errors ranked within a queue of a given size; residual EER and area under the risk–coverage curve (AURC) are reported in Appendix Table 23. (b) Fifteen equal-width reliability bins compare retrieval kNN with the cross-fitted model. (c) At the same 33.75% queue size, the four-cue union covers 82.85% of errors at 20.70% precision; the learned-record margin queue covers 88.66% at 22.15% precision. The cue union shows which stored fields flagged each example. (d) Most covered errors trigger more than one cue; passive–record decision mismatch is the most common single cue. Exact values appear in Appendix Table 25. This figure does not evaluate a deployment-time inspection policy.

A.3 Passive-Detector Context

These tables place the matched decision-record comparison beside stronger passive SSL references and the self-voice-conversion stress setting. Appendix Table 8 shows that longer crops do not improve frozen SSL baselines on these development runs. Appendix Tables 10 and 11 place the matched WavLM reference on the same 4,080 examples as the decision-record analysis. Appendix Table 12 summarizes the self-voice-conversion stress condition.

For WavLM-large, the 4 s full-development runs have 7.00% best and 8.57% median EER; the 6 s runs have 9.03% best and 9.70% median EER. The tested longer crop does not improve this passive reference.

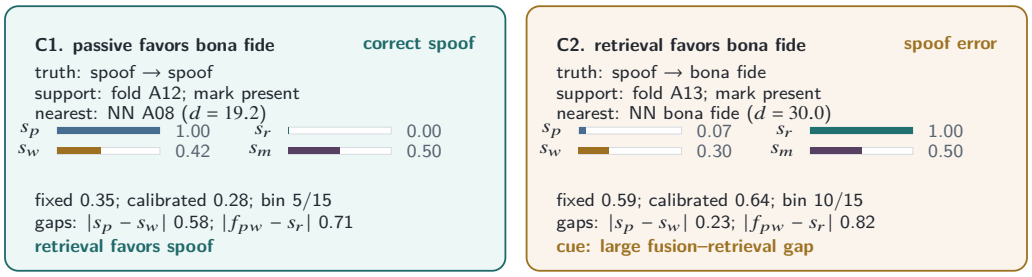

Figure 4: Worked matched-intersection examples from the main decision-record model. C1 is a correctly identified spoof from family A12: the passive score favors bona fide, retrieval favors spoof, and the calibrated score is below threshold. C2 is a failure from family A13: the passive score favors spoof, retrieval favors bona fide, and a large raw-score gap remains visible in the final record.

Table 8: Full-development SSL duration controls for passive detector context. Entries use 20k-per-class ASVspoof 5 Track 1 development runs with 40,000 scored examples. The last column is the signed absolute difference in median EER (%), crop minus the same backbone's completed 4 s reference, computed from unrounded medians.

Backbone	Crop	Runs	Best EER (%) ↓	Median EER (%) ↓	EER range (%) ↓	Δ median EER (%)
WavLM-large	4 s	12	7.00	8.57	7.00–11.16	+0.00
WavLM-large	6 s	8	9.03	9.70	9.03–10.87	+1.13
WavLM-large	8 s	5	11.15	12.80	11.15–14.92	+4.22
WavLM-large	10 s	8	10.91	13.17	10.91–15.04	+4.60
WavLM-large	12 s	4	13.00	13.54	13.00–13.93	+4.97
WavLM-large	14 s	4	11.29	11.64	11.29–12.35	+3.07
WavLM-large	16 s	4	8.67	9.03	8.67–9.32	+0.45
Wav2Vec2-large LV60	4 s	11	9.21	9.45	9.21–9.73	+0.00
Wav2Vec2-large LV60	6 s	7	9.04	9.34	9.04–9.83	-0.11
Wav2Vec2-large LV60	8 s	4	9.84	10.13	9.84–10.44	+0.68
Wav2Vec2-large LV60	10 s	7	9.80	10.20	9.80–10.88	+0.74
Wav2Vec2-large LV60	12 s	4	11.27	11.43	11.27–12.62	+1.98
Wav2Vec2-large LV60	14 s	4	10.96	11.34	10.96–12.03	+1.89
Wav2Vec2-large LV60	16 s	4	10.48	11.02	10.48–11.20	+1.57
HuBERT-large	4 s	11	8.94	9.21	8.94–9.73	+0.00
HuBERT-large	6 s	7	9.45	10.04	9.45–10.64	+0.83
HuBERT-large	8 s	4	16.07	17.73	16.07–18.19	+8.52
HuBERT-large	10 s	7	19.07	19.98	19.07–21.00	+10.78
HuBERT-large	12 s	4	17.95	18.84	17.95–19.34	+9.63
HuBERT-large	14 s	4	19.57	20.17	19.57–21.34	+10.96
HuBERT-large	16 s	4	20.71	22.52	20.71–23.01	+13.31

Training-size effects do not explain the decision-record comparison. In Table 9, WavLM-large's best EER rises from 7.00% at 20,000 training examples per class to 11.17% at 40,000, while Wav2Vec2-large LV60 and HuBERT-large improve modestly at the larger caps. We use the matched WavLM results only as passive reference context.

Table 9: Full-development SSL training-size controls for passive detector context. Entries aggregate completed 4 s ASVspoof 5 Track 1 development runs by frozen SSL backbone and per-class training cap. The last column is the signed absolute difference in median EER (%), entry minus the same backbone's 20k-per-class reference, computed from unrounded medians.

Backbone	Size/class	Examples	Runs	Best EER (%) ↓	Median EER (%) ↓	Δ median EER (%)
WavLM-large	20,000	40,000	12/12	7.00	8.57	+0.00
WavLM-large	40,000	71,334	10/10	11.17	11.92	+3.35
WavLM-large	80,000	111,334	10/10	9.58	10.52	+1.94
WavLM-large	120,000	140,950	10/10	9.53	10.35	+1.78
Wav2Vec2-large LV60	20,000	40,000	11/11	9.21	9.45	+0.00
Wav2Vec2-large LV60	80,000	111,334	10/10	8.92	9.02	-0.43
Wav2Vec2-large LV60	120,000	140,950	10/10	8.89	8.94	-0.51
HuBERT-large	20,000	40,000	11/11	8.94	9.21	+0.00
HuBERT-large	80,000	111,334	10/10	8.37	8.65	-0.55
HuBERT-large	120,000	140,950	10/10	8.41	8.55	-0.66

On the same 4,080 examples, the 22 WavLM-large runs range from 6.71% to 13.63% EER, with a 9.60% median. Under the same cross-fitted protocol, WavLM calibration alone reaches 11.79% pooled EER and the decision record reaches 9.14%, although their fold-level ordering reverses (5.00% versus 6.56%). These tables distinguish the selected best passive run from the same-protocol comparison.

Table 10: Matched frozen SSL passive context on the same 4,080 examples as Table 1. Runs vary backbone, seed, and training size. The best entry is chosen within each backbone; no entry includes watermark, retrieval, or profile evidence.

Backbone	Runs	Examples per run	Best EER (%) ↓	Median EER (%) ↓	EER range (%) ↓	minDCF / ECE ↓
WavLM-large	22/22	4,080	6.71	9.60	6.71–13.63	0.9776/0.0453
HuBERT-large	11/11	4,080	8.21	9.24	8.21–9.48	0.9792/0.0416
Wav2Vec2-large	11/11	4,080	9.14	9.83	9.14–10.05	0.7801/0.0318

Table 11: WavLM passive context on the same 4,080 matched examples as Table 1. The passive entry is the lowest-EER WavLM-large run among 22 matched passive SSL runs. Each Δ EER is the absolute difference in EER (%) between that row and the best passive run at 6.71% (row EER minus reference EER), computed from unrounded scores. The sweep median is 9.60% EER and is shown as context. W, R, and P denote watermark, retrieval, and profile fields.

System	Fit	Pooled EER (%) ↓	Family EER (%) ↓	Fold EER (%) ↓	ΔEER (%) ↓	$\text{minDCF}_{\text{BF}}$ ↓	ECE ↓
WavLM passive (best run)	passive	6.71	5.33	5.00	+0.00	0.9776	0.0453
Retrieval kNN	no fit	15.84	13.85	13.69	+9.12	0.9406	0.1451
WavLM + watermark	fixed	10.86	8.82	–	+4.15	0.9919	0.2813
WavLM + watermark + retrieval	fixed	11.33	9.53	8.80	+4.61	0.9872	0.2019
WavLM + watermark + retrieval + profile	fixed	10.17	8.09	7.83	+3.46	0.9708	0.2861
WavLM decision record	**learned**	**9.14**	**–**	**6.56**	**+2.43**	**0.9984**	**0.0657**
WavLM passive (cross-fitted)	passive fit	11.79	–	5.00	+5.08	0.9823	0.1076
WavLM score fusion	learned scalar	9.72	–	6.85	+3.01	0.9991	0.0712

On the 2,000-example self-voice-conversion condition, the 21 HuBERT-large runs average 29.89% EER and the 26 Wav2Vec2-large runs average 29.65%. This passive-score context concerns converted speech and does not test whether the keyed probe authenticates the source.

Table 12: Frozen-SSL self-voice-conversion stress summary. Each entry aggregates completed runs on the same 2,000-example converted-speech condition used for the passive CNN entry in Table 34. Ranges show run-to-run spread within a model family; EER values are percentages.

Model family	Runs	EER range (%) ↓	Mean EER (%) ↓	minDCF range ↓	Mean ECE ↓
HuBERT-large	21	25.60–34.50	29.89	0.9480–1.0000	0.4564
Wav2Vec2-large	26	23.70–36.50	29.65	0.9890–1.0000	0.4331
WavLM-base-plus	12	27.30–30.00	28.21	0.9990–1.0000	0.4431
WavLM-large	34	32.90–43.40	37.25	0.9970–1.0000	0.4700

A.4 Probe and Retrieval Boundary Checks

Removing the probe field from the matched record changes pooled EER from 11.91% to 12.46%; the probe alone reaches 48.67% EER. Table 13 shows that the probe is conditional context, not a separately useful detector.

Table 13: Effect of removing or changing the probe-score field on the 4,080 matched examples. We remove, neutralize, or invert the stored probe values while holding other scores fixed. The table separates probe availability from failure-mode behavior.

Scenario	W field	Pooled EER (%) ↓	Family EER (%) ↓	$\text{minDCF}_{\text{BF}}$ ↓	ECE ↓	Brier ↓	Worst family ↓
Full record: passive, probe, retrieval	yes	11.91	10.27	0.8283	0.2609	0.1415	A15 (25.85)
Probe unavailable: reweighted passive/retrieval	no	12.46	9.35	0.7692	0.1878	0.1098	A15 (21.69)
Probe unavailable: neutral slot	no	12.46	9.35	0.7692	0.2582	0.1375	A15 (21.69)
Inverted probe	yes	12.37	9.51	0.7552	0.2095	0.1130	A15 (22.40)
Full record + profile (all streams)	yes	12.96	9.86	0.7673	0.3270	0.1834	A15 (22.33)
Probe unavailable + profile	no	12.72	9.84	0.7496	0.2522	0.1366	A15 (22.42)

The full-development check in Table 14 uses 5,000 bona fide and 5,000 spoof examples. Adding the probe to the passive score does not lower EER in either the fixed or learned fusion setting. These utterances and score definitions differ from the 4,080-example matched decision-record analysis, so this check only bounds the probe's stand-alone predictive interpretation.

Table 14: Passive/probe fusion on the separate 10,000-example ASVspoof 5 development set. EER differences are absolute candidate-minus-passive differences in EER (%); confidence intervals come from the paired bootstrap audit. The learned setting fits an earlier passive/probe model, not the four-cue decision record.

Setting	EER (%) ↓	Absolute EER difference (%)
Passive CNN	22.42	Reference
Fixed passive/probe fusion	22.54	+0.12 [-0.08, +0.30]
Learned passive/probe fusion	22.56	+0.14 [-0.10, +0.32]

At the recorded threshold, all 10,000 directly marked examples are accepted and all 10,000 unmarked examples are rejected. Source-pair and self-voice-conversion controls remain near chance; Appendix Table 15 reports the raw statistic and threshold, while Table 34 gives the control counts.

Table 15: Operating counts for watermark-probe conditions. Probe mean, probe standard deviation (SD), and threshold come from the raw keyed-probe statistic; scores at or above the threshold are accepted as marked. Direct pairs use marked and unmarked pairs, and the remaining conditions come from the conversion study.

Condition	*n*	Probe mean	Probe SD	Probe threshold	Marked acc./rej.	Unmarked acc./rej.	Interpretation
Direct marked versus unmarked pairs	20,000	0.01258	0.01261	0.01421	10,000/0	0/10,000	positive control
Source-pair control	2,000	0.00007	0.00099	-0.00481	1,000/0	1,000/0	threshold accepts both classes
Converted-pair condition	2,000	0.00000	0.00086	-0.00085	905/95	896/104	weak mark readout after conversion

The field-meaning table distinguishes a known-key response on a marked copy from authentication of the original utterance. The available source tags do not establish prompt, channel, or recording-source independence.

Table 16: Meaning and applicability of each decision-record field. The table states how each channel is used in this paper and distinguishes a keyed-probe response from authentication of the original utterance.

Scenario	Field(s) used	Reading in this paper	Open requirement
Known key, directly marked signal	s_w on marked/unmarked pairs	probe response and operating threshold under the recorded key	performance on unmarked source audio
Derived marked copy in matched examples	s_p, s_w, s_r, s_m and gaps	keyed-probe response and cross-stream raw-score gaps under the same derived condition for both classes	proof that the original benchmark file carried a mark
No key or absent mark	s_p, s_r, s_m plus availability status	passive, retrieval, and profile evidence stays available without the probe field	watermark authentication
Removal, corruption, or conversion	probe scores under stress	near-chance scores identify a failure condition for the probe channel	mark survival under attack
Retrieved/profile support available	s_r, s_m, c_r and neighbor metadata	support-set context after family exclusion and overlap analysis	prompt, channel, or collection independence

Matching on availability of all four cues retains a different fraction of each synthesis family than the full retrieval score set. Table 17 reports these family counts alongside the retrieval-score and neighbor-overlap checks; it describes a selected development subset, not a representative test population.

Table 17: Retrieval-subset selection and nearest-neighbor overlap analysis. Full scores are computed before forming the matched subset used in Table 1. The final columns report held-out-family leakage in the nearest neighbor and top-k list; coarse source tags describe corpus-level grouping and do not identify prompt or recording-source separation.

Family	Full spoof	Matched spoof	Retained	Full EER (%) ↓	Matched EER (%) ↓	Same-family NN	Same-family top-k
A09	2,479	111	4.5%	15.36	13.51	0/2,479; 0/111	0/2,479
A10	2,548	124	4.9%	14.84	15.22	0/2,548; 0/124	0/2,548
A11	2,568	121	4.7%	26.05	24.65	0/2,568; 0/121	0/2,568
A12	2,457	91	3.7%	1.96	2.03	0/2,457; 0/91	0/2,457
A13	2,526	100	4.0%	23.48	19.19	0/2,526; 0/100	0/2,526
A14	2,450	102	4.2%	1.95	1.91	0/2,450; 0/102	0/2,450
A15	2,426	116	4.8%	27.41	29.31	0/2,426; 0/116	0/2,426
A16	2,546	100	3.9%	6.48	5.00	0/2,546; 0/100	0/2,546

Subset matching changes the retrieval-score distribution relative to the full score set. For A13, matched EER (19.19%) minus the bootstrap estimate (23.49%) gives an absolute EER difference of -4.30%, the largest magnitude across families. Appendix Table 18 reports the family-level sensitivity; it does not establish that the subset represents all retrieval predictions.

Table 18: Matched-subset sensitivity analysis for retrieval scores. The bootstrap column gives the same-size random-retention EER estimate from the full retrieval score set. The EER-difference column is the absolute difference in EER (%), matched-subset EER minus bootstrap EER. For each spoof family, the bona fide retrieval score difference (%) is the mean score for retained spoof examples minus the mean score for omitted spoof examples, scaled by 100; positive values mean retained spoof examples receive higher bona fide scores. The last column is the Kolmogorov–Smirnov (KS) distance.

Family	Retained	Match EER (%) ↓	Boot. EER (%) ↓	Match - boot. absolute EER difference (%)	Bona fide retrieval score diff. (%)	KS ↓
A09	4.5%	13.51	15.31 [12.60, 18.01]	-1.80	-2.01	0.085
A10	4.9%	15.22	14.77 [12.10, 17.58]	+0.45	+0.21	0.058
A11	4.7%	24.65	26.09 [22.11, 30.58]	-1.44	-3.12	0.078
A12	3.7%	2.03	2.23 [1.34, 3.30]	-0.19	-0.24	0.017
A13	4.0%	19.19	23.49 [19.00, 28.00]	-4.30	-7.45	0.133
A14	4.2%	1.91	1.96 [1.74, 2.18]	-0.05	+0.07	0.009
A15	4.8%	29.31	27.49 [23.07, 32.62]	+1.81	+0.00	0.083
A16	3.9%	5.00	6.27 [4.01, 8.99]	-1.26	-1.80	0.046

Across 20,000 spoof queries, no held-out synthesis-family example appears among the ten nearest neighbors. The available metadata also shows no exact query, speaker-ID, or audio-path reuse; broader source and channel overlap remain unresolved (Appendix Table 19).

Table 19: Identifier-overlap analysis for retrieval/profile nearest neighbors. Counts come from the 40,000-example retrieval set before matching on all four cues with held-out synthesis-family exclusion for spoof examples. The coarse source tag is corpus-level metadata; prompt, channel, and recording-source independence remain unresolved at finer granularity.

Identifier relation	Examples	Matches	Interpretation
Held-out family as nearest neighbor	20,000	0/20,000	blocked by family exclusion
Held-out family anywhere in top-k	20,000	0/20,000	blocked by family exclusion
Same utterance ID anywhere in top-k	40,000	0/40,000	not observed
Same speaker ID anywhere in top-k	40,000	0/40,000	not observed
Same audio path anywhere in top-k	40,000	0/40,000	not observed
Same coarse source tag anywhere in top-k	40,000	40,000/40,000	shared corpus-level tag

A.5 Family and Calibration Evidence

The family analysis shows where the matched-set changes concentrate. Across eight held-out synthesis families, family-macro EER changes from 10.27% for the fixed retrieval-augmented rule to 9.86% when the profile margin is added, but the direction and size vary substantially by family. In particular, the fixed rule has 16.51% EER on A11 and 5.47% on A12, while the passive CNN has 6.43% and 81.31%, respectively; Figure 5 visualizes these uneven family outcomes. Table 20 reports the family-level values and paired uncertainty.

Table 20: Held-out synthesis-family uncertainty for Figure 5. Counts give bona fide and spoof examples. System columns report receiver operating characteristic (ROC)-interpolated EER (%). Δ_P is the absolute EER difference (%) for the fixed retrieval-augmented rule relative to Passive; Δ_R is the same difference relative to Retrieval kNN. Negative values favor the fixed rule. Differences use unrounded EER values. Bold intervals favor the fixed retrieval-augmented rule; † marks intervals against it.

Family	Bona.	Spoof	Passive EER (%)	Ret. EER (%)	Fixed EER (%)	Profile EER (%)	Δ_P (%)	Δ_R (%)
A09	3215	111	9.92	13.51	9.03	**6.31**	-0.9 [-4.3, +3.6]	**-4.5 [-6.3, -2.0]**
A10	3215	124	8.73	15.22	8.09	**5.64**	-0.6 [-3.2, +2.7]	**-7.1 [-8.2, -4.7]**
A11	3215	121	**6.43**	24.65	16.51	13.25	+10.1 [+5.9, +15.8]†	**-8.1 [-10.8, -5.0]**
A12	3215	91	81.31	**2.03**	5.47	13.37	**-75.8 [-79.2, -72.5]**	+3.4 [+2.5, +4.4]†
A13	3215	100	53.00	19.19	15.00	**14.98**	**-38.0 [-48.1, -25.9]**	**-4.2 [-8.0, -1.0]**
A14	3215	102	5.65	1.91	**0.22**	0.97	**-5.4 [-6.7, -3.9]**	**-1.7 [-2.9, -1.0]**
A15	3215	116	23.27	29.31	25.85	**22.33**	+2.6 [-3.6, +8.0]	**-3.5 [-6.9, -0.2]**
A16	3215	100	7.22	5.00	**2.00**	**2.00**	**-5.2 [-9.0, -2.0]**	**-3.0 [-4.6, -1.0]**

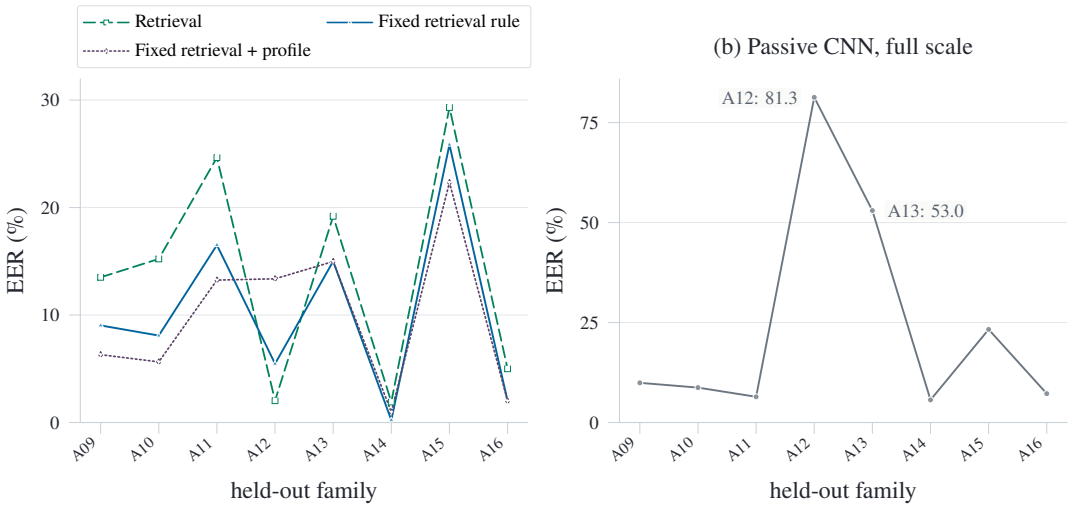

Figure 5: Held-out EER by synthesis family. Panel (a) zooms retrieval and fixed-rule variants; panel (b) keeps the passive-CNN failures on the full scale. Appendix Table 20 gives exact values and paired intervals.

Removing A12 and A13 reverses the family-macro EER comparison: 10.20% for passive CNN versus 10.28% for the fixed rule. The profile variant reaches 8.42% under the same exclusion (Table 21). At pooled EER thresholds, the fixed rule corrects 707 errors made by passive CNN and introduces 330 new errors, for a net reduction of 377. Bona fide examples account for 297 of that reduction; A12 and A13 contribute 90 and 38 net corrected cases, respectively (Table 22). These counts describe matched development examples and do not identify a causal mechanism.

Table 21: Family-influence sensitivity for the matched comparison. Each system entry is family-macro EER (%) after excluding the named held-out spoof family or family pair from Table 20. Δ_P is the absolute difference in family-macro EER (%) for the fixed retrieval-augmented rule relative to Passive; Δ_R is the same difference relative to Retrieval kNN. Differences use unrounded values.

Excluded	Passive EER (%)	Ret. EER (%)	Fixed EER (%)	Profile EER (%)	Δ_P (%)	Δ_R (%)
A09	26.52	13.90	10.45	10.36	-16.07	-3.45
A10	26.69	13.66	10.58	10.46	-16.10	-3.07
A11	27.01	12.31	9.38	9.37	-17.63	-2.93
A12	16.32	15.54	10.96	9.35	-5.36	-4.58
A13	20.36	13.09	9.60	9.12	-10.77	-3.49
A14	27.13	15.56	11.71	11.13	-15.42	-3.85
A15	24.61	11.64	8.05	8.07	-16.56	-3.60
A16	26.90	15.12	11.45	10.98	-15.45	-3.66
A12+A13	10.20	14.93	10.28	8.42	0.08	-4.65

Table 22: Error changes at each system's pooled EER threshold for the passive-CNN versus fixed retrieval-augmented rule comparison. Corrected examples are passive-CNN errors that the retrieval-augmented rule classifies correctly; new errors are retrieval-rule errors on examples classified correctly by the passive CNN. The share divides the net reduction in errors by the total for the full set.

Group	Examples	Passive err.	Fixed-rule err.	Corrected	New err.	Net	Share (%)
All examples	4,080	863	486	707	330	+377	+100.0
Bona fide	3,215	680	383	539	242	+297	+78.8
A09	111	5	8	5	8	-3	-0.8
A10	124	2	6	2	6	-4	-1.1
A11	121	0	30	0	30	-30	-8.0
A12	91	90	0	90	0	+90	+23.9
A13	100	55	17	54	16	+38	+10.1
A14	102	0	0	0	0	+0	+0.0
A15	116	30	42	16	28	-12	-3.2
A16	100	1	0	1	0	+1	+0.3

At their respective EER thresholds, the proposed record makes 344 errors among 4,080 examples (8.43%). Deferring the 10% of examples with the smallest absolute distance from its threshold captures 47.97% of those errors at 40.44% precision, leaving 4.87% error among retained examples. The four diagnostic cues flag 1,377 examples and cover 82.85% of record errors at 20.70% precision (Tables 23 and 24). These quantities use known labels from the development set and do not test a human-review workflow.

Table 23: Post hoc review of matched examples. Each system uses its own EER threshold and defers the 10% of examples closest to that threshold. The area under the risk–coverage curve (AURC) averages retained risk over accepted-prefix coverages and is given with a 95% confidence interval (CI).

System	Errors	Decision error (%) ↓	Error capture at 10% (%) ↑	Review precision (%) ↑	Retained error (%) ↓	AURC (%) ↓ (95% CI)
Passive CNN	863/4,080	21.15	22.83	48.28	18.14	14.85 (13.24–16.47)
Retrieval kNN	646/4,080	15.83	29.26	46.32	12.45	5.74 (5.02–6.50)
Fixed retrieval-augmented rule	486/4,080	11.91	38.89	46.32	8.09	2.49 (2.15–2.84)
Proposed decision record	344/4,080	8.43	47.97	40.44	4.87	1.79 (1.45–2.16)
WavLM passive (best run)	274/4,080	6.72	60.58	40.69	2.94	1.14 (0.84–1.50)
WavLM decision record	373/4,080	9.14	50.40	46.08	5.04	2.96 (2.25–3.74)

Table 24: Post hoc diagnostic-cue analysis for the cross-fitted decision-record model. Cues follow the same EER-threshold convention as Table 23; the last entry is their union. The table does not evaluate an operational inspection policy.

Cue	Rule	Examples	Record errors	Error cov. (%) ↑	Precision (%) ↑
near record threshold	lowest 10% $\lvert s_{\text{rec}} - \tau_{\text{rec}} \rvert$	408	165	47.97	40.44
passive-record decision mismatch	passive and record decisions differ	823	152	44.19	18.47
retrieval-record decision mismatch	retrieval and record decisions differ	564	131	38.08	23.23
large fusion–retrieval gap	top 10% $\lvert f_{pw} - s_r \rvert$	408	66	19.19	16.18
any cue above	union of the four cues	1,377	285	82.85	20.70

At the same 1,377-example load (33.75% of the matched set), the record-margin queue covers 88.66% of record errors at 22.15% precision (Table 25). The four-cue union covers 82.85% of errors at 20.70% precision. It gives up some coverage and precision, while showing which fields flagged each example. No operational policy was evaluated.

Table 25: At a shared load of 1,377 examples, scalar-margin queues are compared with the four-cue union by error coverage and precision, using errors from the calibrated-record threshold as the reference. The four-cue union covers fewer errors at lower precision than the learned-record margin queue, while retaining which cues flagged each example.

Queue	Definition	Examples	Share (%)	Error cov. (%) ↑	Precision (%) ↑
Passive margin	lowest $\lvert s - \tau \rvert$ at the cue-union review load	1,377	33.75	32.85	8.21
Retrieval margin	lowest $\lvert s - \tau \rvert$ at the cue-union review load	1,377	33.75	40.99	10.24
Fixed-rule margin	lowest $\lvert s - \tau \rvert$ at the cue-union review load	1,377	33.75	77.03	19.24
Learned-record margin	lowest $\lvert s - \tau \rvert$ at the cue-union review load	1,377	33.75	88.66	22.15
Four-cue union	union of the four cues	1,377	33.75	82.85	20.70

Appendix Table 26 defines four groups: two correctly classified groups (180 and 146 examples) and two error groups (77 near-threshold errors and 41 cases where retrieval favors bona fide speech). Appendix Figure 6 shows one record from each group and illustrates which measurements agree or conflict. The examples illustrate these groups; they do not estimate prevalence or evaluate a deployment policy.

Table 26: Selection criteria for the four decision-record examples in Figure 6. Counts use the calibrated decision-record operating threshold on the 4,080 matched examples; C1–C2 are correct calibrated decisions and C3–C4 are calibrated-decision errors by construction.

Panel	Selection group	Examples	Share (%)	Group err. (%) ↓	Median margin	$\lvert f_{pw} - s_r \rvert$
C1	**record and retrieval agree**	**180**	**4.41**	**0.00**	**0.475**	**0.043**
C2	**correct spoof; passive disagrees**	**146**	**3.58**	**0.00**	**0.242**	**0.609**
C3	**near-threshold error**	**77**	**1.89**	**100.00**	**0.042**	**0.463**
C4	**retrieval favors bona fide**	**41**	**1.00**	**100.00**	**0.305**	**0.600**

For the main EER contrast, the record's absolute difference from cross-fitted scalar fusion is -3.48% (8.43% - 11.91%; paired 95% CI [-4.60%, -2.34%]; held-out-fold CI [-5.85%, -0.20%]). The absolute EER difference from the no-probe control is -0.24% (8.43% - 8.67%; 95% CI [-0.81%, +0.46%]). From the squared-difference model, it is +0.10% (8.43% - 8.33%; 95% CI [-0.26%, +0.57%]). The latter two intervals include zero. Table 27 reports other paired absolute EER differences as candidate minus baseline.

Table 27: Paired bootstrap absolute differences in equal error rate (EER; %) for Tables 1 and 2. Each difference is candidate EER minus baseline EER; negative values favor the candidate. Differences and intervals use unrounded EER values. Intervals are nominal 95% confidence intervals (CIs) and are not adjusted across comparisons.

Baseline	Candidate	Paired Δ EER (%) CI ↓	Fold-level CI (%)
Passive CNN	**Fixed retrieval rule**	-9.24 [-11.89, -7.15]	–
Passive + watermark	**Fixed retrieval rule**	-9.12 [-11.91, -7.06]	–
Retrieval kNN	**Fixed retrieval rule**	-3.92 [-5.20, -3.22]	–
Passive CNN	**Fixed retrieval + profile**	-8.19 [-10.43, -6.23]	–
Passive + watermark	**Fixed retrieval + profile**	-8.07 [-10.52, -6.13]	–
Retrieval kNN	**Fixed retrieval + profile**	-2.88 [-4.51, -1.51]	–
Fixed retrieval-augmented rule	**Proposed decision record**	-3.48 [-4.61, -2.08]	–
Cross-fitted scalar fusion	**Proposed decision record**	-3.48 [-4.60, -2.34]	[-5.85, -0.20]
Retrieval + profile only	**Proposed decision record**	-8.21 [-9.83, -6.73]	[-12.23, -3.64]
Passive + retrieval only	**Proposed decision record**	-3.24 [-4.27, -2.08]	[-5.41, -0.24]
No-probe nonlinear control	**Proposed decision record**	-0.24 [-0.81, +0.46]	[-0.80, +0.49]
Passive margin only	**Proposed decision record**	-15.61 [-17.12, -14.11]	[-18.04, -13.95]
Passive shape control	**Proposed decision record**	-13.06 [-15.42, -10.89]	[-35.51, -2.49]
Passive-probe gap	**Proposed decision record**	-0.46 [-1.30, +0.14]	[-2.37, +0.11]
Shuffled-probe control	**Proposed decision record**	-0.34 [-1.26, +0.24]	[-2.18, +0.17]
Fusion-retrieval gap	**Proposed decision record**	-0.34 [-0.93, +0.22]	[-1.25, +0.25]
Squared-difference record	**Proposed decision record**	+0.10 [-0.26, +0.57]	[-0.26, +0.97]
Product-term control	**Proposed decision record**	-1.72 [-2.65, -0.67]	[-3.66, +0.12]

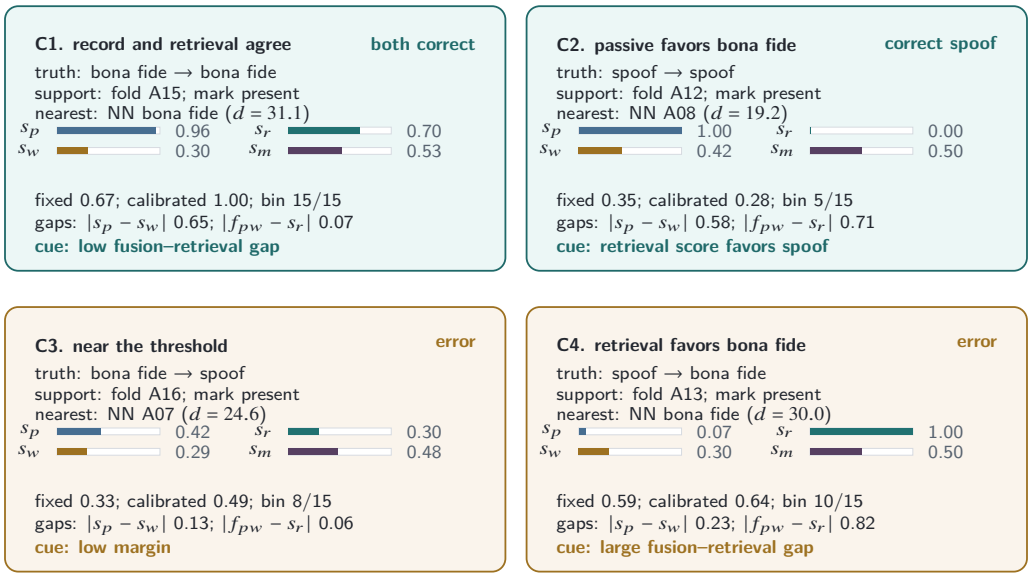

Figure 6: Four example decision records. Each decision uses the calibrated operating threshold for the decision-record model; the fixed score is shown as a reference. Horizontal bars show passive (blue), probe (gold), retrieval (teal), and profile (purple) scores on the same 0–1 scale; each record also lists the decision status, watermark status, calibration bin, nearest-neighbor context, raw-score differences, and diagnostic cue.

The proposed record's ECE is 0.0709, compared with 0.0840 for cross-fitted scalar fusion; the paired difference is -0.0130 [-0.0159, -0.0090]. The no-probe nonlinear contrast is -0.0047 [-0.0090, -0.0020], while the squared-difference contrast is +0.0001 [-0.0016, +0.0021] and includes zero (Table 28).

Table 28: Paired absolute differences in expected calibration error (ECE), proposed record minus each named cross-fitted control, on the 4,080 matched examples. Negative values favor the proposed record. Differences use unrounded ECE values. Intervals are nominal 95% intervals from 5,000 class-stratified bootstrap resamples and are not adjusted across contrasts.

Control	ECE	Absolute ECE difference [95% CI] ↓
Cross-fitted scalar fusion	0.0840	-0.0130 [-0.0159, -0.0090]
Passive + retrieval only	0.0904	-0.0195 [-0.0223, -0.0156]
No-probe nonlinear control	0.0756	-0.0047 [-0.0090, -0.0020]
Product-term control	0.0780	-0.0071 [-0.0093, -0.0035]
Squared-difference record	0.0709	+0.0001 [-0.0016, +0.0021]

Resampling speakers instead of utterances gives an absolute EER difference of -3.48% between the record (8.43%) and scalar fusion (11.91%; 95% CI [-4.63%, -2.34%]). For the no-probe control, the difference is -0.24% (8.43% - 8.67%; CI [-0.82%, +0.48%]). For the squared-difference model, it is +0.10% (8.43% - 8.33%; CI [-0.28%, +0.53%]). Both control intervals include zero (Table 29). The table separates paired-example, held-out-fold, and speaker-resampled intervals because their resampling units differ.

Table 29: Absolute EER differences (%), candidate minus baseline, for speaker-resampled calibration checks. The candidate is the cross-fitted absolute-difference model in Table 2; negative values favor it. Paired intervals resample matched examples, fold intervals resample the eight held-out calibration folds, and speaker intervals resample the 705 speaker IDs. Intervals are nominal 95% and are not adjusted across comparisons.

Baseline	Paired absolute EER difference (%)	Fold CI (%)	Speaker CI (%)
Cross-fitted scalar fusion	-3.48 [-4.60, -2.34]	[-5.85, -0.20]	[-4.63, -2.34]
No-probe nonlinear control	-0.24 [-0.81, +0.46]	[-0.80, +0.49]	[-0.82, +0.48]
Product-term control	-1.72 [-2.65, -0.67]	[-3.66, +0.12]	[-2.72, -0.65]
Squared-difference record	+0.10 [-0.26, +0.57]	[-0.26, +0.97]	[-0.28, +0.53]

Family-held-out results vary. The proposed model reaches 6.42% EER on A11 and 6.44% on A12. The squared-difference setting is lower on A12 (4.62% versus 6.44%) and ties the proposed model on A13 (9.99% each). Across ten alternative bona fide-to-fold assignments, pooled EER ranges from 8.33% to 8.55% for the absolute-difference record and 8.11% to 8.33% for the squared-difference model (Tables 30 and 31). These checks describe fold and assignment sensitivity; they do not remove the family concentration shown above.

Table 30: Per-fold equal error rate (EER; %) for learned calibration on the 4,080 matched examples. Each held-out family is evaluated by a calibrator trained without spoof examples from that family; bold marks the best learned setting in that family.

Family	Passive CNN EER (%)	Fixed rule EER (%)	Score fusion EER (%)	Squared gaps EER (%)	Nonlinear EER (%)	Proposed model EER (%)
A09	9.01	8.30	6.34	6.20	**5.62**	6.34
A10	8.03	6.45	4.74	4.08	**3.95**	4.08
A11	6.66	15.87	11.53	**6.42**	8.97	**6.42**
A12	80.18	6.81	13.01	**4.62**	9.91	6.44
A13	53.06	13.98	13.11	**9.99**	12.86	**9.99**
A14	4.93	0.99	**0.74**	0.99	**0.74**	0.86
A15	23.28	25.92	22.36	**19.96**	**19.96**	21.56
A16	7.94	1.99	**1.99**	**1.99**	**1.99**	**1.99**

Table 31: Calibration sensitivity to bona fide fold assignment. Spoof folds remain fixed by held-out synthesis family; only the stable hash that assigns bona fide examples to those folds changes. Each cell reports mean [minimum, maximum] over ten assignments; EER values are percentages.

Setting	Assign.	Pooled EER (%) ↓	Fold EER (%) ↓	ECE ↓
Score fusion	10	11.89% [11.69, 12.05]	9.45% [9.27, 9.72]	0.0840 [0.0836, 0.0845]
Decision record (abs. gaps)	10	8.42% [8.33, 8.55]	7.06% [6.69, 7.31]	0.0707 [0.0704, 0.0709]
Decision record (sq. gaps)	10	8.22% [8.11, 8.33]	6.73% [6.36, 7.14]	0.0707 [0.0705, 0.0709]

Appendix Figure 3 shows that most miscalibration lies near the upper score range. In Table 32, the record's width ECE is 0.0709 and its equal-mass ECE is 0.0715. The fixed retrieval rule scores 0.2609 and 0.2557, respectively. Isotonic maps fitted on other folds reduce these ECE values to 0.0178 and 0.0321, respectively. These values describe calibration on the matched development set.

Table 32: Calibration-bin sensitivity on the 4,080 matched examples. Width ECE is the main-table value; mass ECE uses 15 equal-mass bins of 272 examples. Fold-iso ECE applies an isotonic map fit on the other folds. Sparse bins count equal-width bins with fewer than 100 examples.

System	Width ECE ↓	Mass ECE ↓	Fold-iso ECE ↓	Sparse bins	Largest bin	Max mass gap
Retrieval kNN	0.1451	0.1451	0.0187	6	33.6%	0.4008
Fixed retrieval-augmented rule	0.2609	0.2557	0.0321	3	18.7%	0.4123
Proposed decision record	**0.0709**	**0.0715**	**0.0178**	**6**	**49.3%**	**0.2832**

The retrieval and profile sweep uses examples before matching on all four cues; its population therefore differs from the 4,080-example main analysis. Under the HuBERT-large 20k default, retrieval EER is 16.46% and profile EER is 59.11%. Under the WavLM-large 20k default, they are 11.12% and 68.12%, respectively. These values describe the tested configurations. They do not show that the downstream setting was selected in advance or predict performance on the matched subset (Table 33).

Table 33: Retrieval and profile measurements across the evaluated configurations, before forming the matched subset. The main comparison uses the HuBERT-large 20k default setting; the WavLM rows provide sensitivity context for examples before matching on all four cues.

Configuration	Support	Examples	Retrieval EER (%)	Profile EER (%)	Worst family EER (%) ↓
WavLM-large 20k default	38,797	40,000	11.12	68.12	A11 / 27.80
WavLM-large 20k d6	38,797	40,000	11.30	38.77	A11 / 21.55
HuBERT-large 20k d6	38,797	40,000	16.27	43.29	A15 / 26.13
HuBERT-large 20k default (main)	38,797	40,000	16.46	59.11	A11 / 27.66
WavLM-large 20k d2	38,797	40,000	17.23	68.83	A11 / 27.08
Wav2Vec2-large 20k d6	38,797	40,000	21.11	60.72	A11 / 34.70
Wav2Vec2-large 20k default	38,797	40,000	21.50	63.32	A11 / 34.52
Wav2Vec2-large 20k d10	38,797	40,000	23.02	59.71	A11 / 33.01
HuBERT-base 40k default	58,797	71,334	23.05	75.37	A11 / 41.56
WavLM-base+ 40k default	58,797	71,334	24.57	71.93	A11 / 48.88
HuBERT-large 20k d8	38,797	40,000	26.04	43.72	A15 / 35.18
Wav2Vec2-base 40k default	58,797	71,334	31.91	56.53	A13 / 40.06
WavLM-large 20k d10	38,797	40,000	43.79	53.30	A13 / 53.87

On the separate probe-control set, the direct marked-versus-unmarked pair has 0.00% EER and 0.0000 minDCF. The source-pair control has 50.00% EER (95% CI [47.7, 52.1]). The self-voice-conversion probe control has 52.00% EER (95% CI [49.6, 54.1]), while passive CNN reaches 48.50% EER on that set (Table 34). These results support a response to the known-key marked-copy condition only. They do not establish authentication of original audio or general probe validity.

Table 34: Stress checks for the auxiliary probe field and passive reference on separate control sets. The detail column reports the relevant probe-control summary. WM denotes watermark, Self-VC denotes self-voice conversion (VC), and CI denotes confidence interval.

Condition	Score	n	Balance	EER (%) ↓	$\text{minDCF}_{\text{BF}}$ ↓	Detail
Probe pair	WM probe	20,000	10,000/10,000	0.00	0.0000	paired originals
Source pair	WM probe	2,000	1,000/1,000	50.00	0.9990	CI [47.7,52.1]; 485 spk.
Self-VC	WM probe	2,000	1,000/1,000	52.00	1.0000	CI [49.6,54.1]
Self-VC	Passive CNN	2,000	1,000/1,000	48.50	1.0000	CI [46.3,50.8]