algorithms

MDPI

Article

MiLTL: A Cross-Modal Contradiction Cascade for On-Device Voice Phishing Detection

Bongjin Jung [1,2] and Joongho Chang [1,*]

[1] Program in AI Convergence Engineering, aSSIST University, Seoul 03767, Republic of Korea; eugene_jung@stud.assist.ac.kr
[2] SDG Management School, 1202 Geneva, Switzerland
* Correspondence: jhchang@assist.ac.kr

Abstract

Voice phishing (vishing) causes financial and psychological harm, yet deployed detectors that scan transcripts for scam vocabulary are more fragile than their benchmark scores suggest: their near-perfect accuracy on standard Korean corpora is memorization, collapsing under scammer paraphrase or language-model rewriting. We construct **KorMMP**, a benchmark keeping real regulator-sourced scam audio and real benign speech but decorrelating transcript from label, and **MiLTL**, an on-device detector built on affective and neutrosophic channels, including a cross-modal contradiction signal (XM) formulated to weigh lexical warmth against vocal coldness. In the reported evaluation, XM's measured contribution is confidence banding and escalation routing rather than ranking accuracy. A scoring rule with zero gradient-learned parameters at inference screens every segment, referring ambiguous calls to a small on-device language model; raw audio stays on the device. We measure both stages on a commodity CPU container and smartphone; these budgets cover the detector only and exclude speech recognition, which we expect to dominate a live end-to-end budget. With no training on the hard benchmark, MiLTL achieves an AUROC of 0.965 (recording-level cluster bootstrap 95% CI 0.943–0.984), while every evaluated comparator stays below 0.69: text-only, audio-only, fusion, 7B multimodal. MiLTL scores lower on the saturated corpus, as expected for a detector designed not to rely primarily on lexical shortcuts. KorMMP is a controlled stress test of lexical decorrelation, not a measure of in-the-wild detection; its harmful and benign audio come from different corpora, so the source and label are structurally confounded, and the residual source effects cannot be fully excluded. XM is author-defined, constructed rather than naturally occurring in the synthetic stratum and not yet validated against human perception; this is the principal open limitation of the work. Within that scope, the work provides a reproducible basis for on-device vishing defense and a benchmark for detector robustness under vocabulary shift.

Keywords: voice phishing; lightweight algorithms; on-device detection; cross-modal contradiction; neutrosophic channels; edge AI; small language model; distribution shift; privacy-preserving detection; Korean speech

Check for updates

Academic Editors: Izhar Ahmed Khan and Junaid Akram

Received: 29 July 2026
Revised: 25 August 2026
Accepted: 26 August 2026
Published: 27 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. The Scale and Human Cost of Voice Phishing

Voice phishing (vishing) is now one of the most damaging forms of consumer fraud worldwide. In the Republic of Korea, reported losses surpassed KRW one trillion in 2025 [1], with a 2.2-fold first-quarter surge [2,3], steeply rising per-victim losses [4], repeated

regulatory escalation [5,6], and severe psychological harm concentrated on vulnerable populations [7,8]. Globally, fraud losses run to hundreds of billions of dollars per year [9,10], driven by industrialized scam centers [11] and telecom-initiated contact [12]. Generative voice cloning lowers the attacker's cost further [13], large language models can be turned against vishing classifiers directly [14], and a survey of the field identifies spoken social engineering detection as an urgent open problem [15].

1.2. Why Transcript Classification Is Fragile

The dominant approach to automated defense couples an automatic speech recognition (ASR) front-end to a text classifier trained on scam vocabulary [16–19]. On in-corpus Korean benchmarks, such detectors appear to be perfect: in our own runs, frozen text encoders reached an AUROC of 1.000 on the transcript-only KorCCViD corpus. We argue that this is an artifact of *shortcut learning* [20]. Standard corpora are lexically trivial, so a text model can memorize surface vocabulary (finance, authority, and urgency tokens) instead of the underlying mechanism of manipulation. Decorrelate that vocabulary from the label, as any scammer does by adopting novel phrasing, and the model collapses. This is the classic gap between in-distribution and out-of-distribution generalization [21,22]. High in-corpus accuracy is therefore not evidence of robustness.

Two further pressures make the transcript-only paradigm untenable for real deployment. First, it is reactive to wording and therefore vulnerable to paraphrasing and LLM-rewritten scripts [14]. Second, running ASR and classification in the cloud raises acute privacy concerns. Korean commercial services that transmit call text or summaries to servers have recently drawn retention controversies and suffered breaches [23–25]. That record makes the case for keeping raw audio and its derivatives *on the device*.

1.3. From Keywords to Nibbles: Our Approach

We propose to detect manipulation from delivery-aware evidence combined with lexical signals, rather than primarily from scam vocabulary. MiLTL (Multimodal In-call Lightweight Threat Locator) does not match keywords. It segments live call audio into short 8 s windows we call *nibbles* and extracts five neutrosophic/affect channels from each: truth (T), indeterminacy (I), falsity (F), emotion energy (E), and cross-modal contradiction (XM). XM weights *warm*, cooperative lexical content against *cold*, controlling prosody—a signal no single-modality framework can compute [26,27]. The deterministic Gate-1 accumulates the channels over the call into a risk score with zero gradient-learned parameters at inference. It escalates only the uncertain calls to a compact small-language-model arbiter (Gate-2). The nibble is the key design unit. It is long enough to expose the *relationship* between what is said and how it is said, and short enough for low-latency, rolling in-call accumulation. Our results locate the robustness in the *fused* per-nibble channels, each of which already combines prosodic and lexical evidence; neither keyword spotting nor affect energy alone suffices. The contradiction signal chiefly governs *where the detector is confident enough to decide alone*, widening the confidently decided region and reducing unnecessary escalation (Section 8.5).

1.4. Contributions

MiLTL is an engineering composition, not a new detection primitive. Each of its parts is well understood on its own: calibrated logistic features, a fixed-weight linear score, a confidence band with a reject option, and a LoRA-adapted small language model. The contribution is the arrangement, which fixes what each stage may see, where the system is allowed to decide alone, and what the system must defer. We characterize that arrangement on a benchmark built to break the shortcut its predecessors rely on.

1. **KorMMP**, a lexically decorrelated Korean multimodal vishing benchmark whose hard slices are defined by observable lexical density, rather than by label, such that high performance on it requires a mechanism beyond vocabulary.
2. **The XM channel**, a modality-agnostic cross-modal contradiction signal, formalized within a neutrosophic (T, I, F) plus emotion energy representation. In the present evaluation, XM contributes primarily to confidence banding and escalation routing, rather than to raw ranking accuracy as measured by AUROC (Section 8.5).
3. **The MiLTL-Cascade**, a lightweight two-stage detection algorithm for edge devices which pairs a deterministic Gate-1 scorer that carries zero gradient-learned parameters at inference with an escalate-only single-pass SLM arbiter. Under the freeze protocol, it outperforms every evaluated comparator on KorMMP by a wide margin, whether text-only, audio-only, naïve fusion, or a 7B multimodal reference (Section 8).
4. An open, reproducible release: the code and derived data are publicly available (Section 10).

Three boundaries frame everything that follows, and we state them here rather than deferring them to the limitations. First, KorMMP is a *controlled stress test*. Its hard cases guarantee lexical decorrelation by construction, so results on it measure robustness to the vocabulary shortcut under a freeze protocol. They do not measure in-the-wild vishing detection on naturally paired calls. Second, the benchmark's harmful and benign audio originate from *different source corpora*: regulator-archived phishing calls versus public daily dialogue. Telephone-band equalization and a provenance audit remove the measurable bandwidth and codec shortcuts, but residual source, speaker, genre, and conversational-role effects cannot be fully excluded. What this paper demonstrates is therefore the detection of delivery-borne signals under this benchmark's conditions, not the detection of "manipulation" as such. Third, the cross-modal construct is an author-defined operational signal that has not yet been validated against human perception (Section 9.2). The remainder of the paper reviews the theoretical background (Section 2) and legacy detectors (Section 3), then states the problem and hypotheses (Section 4). It details the cascaded detector (Section 5) and the KorMMP benchmark (Section 6), specifies the metrics (Section 7), reports the results (Section 8), and discusses the study's limitations (Section 9).

2. Theoretical Background

MiLTL rests on three strands of prior theory, which are condensed here. *Neutrosophic representation* [28,29] models a proposition using three independent membership degrees, truth (T), indeterminacy (I), and falsity (F), which need not sum to one. It generalizes fuzzy and intuitionistic fuzzy sets [30]. Like Dempster–Shafer theory [31], it gives indeterminacy a first-class role. Prior work applies it to decision-making and speech/text sentiment analysis [32,33]. Vishing calls are saturated with indeterminacy: a caller claiming to be a prosecutor may be genuine. Because the T/I/F axes need not trade off against one another as a probability simplex would, Gate-1 can represent "high-truth claim *and* high-falsity evidence", the signature of a manipulative script delivered with false warmth. The score draws on that capacity in *transcript-only* situations, where the voice channels E and XM cannot be computed and F enters the score.

Affect and deception leakage: Affective computing has established that arousal and valence are recoverable from speech [34–36]. Deception research on nonverbal leakage [37] and Interpersonal Deception Theory [38] predicts that a deceiver can control wording more easily than prosody, which produces *divergence*. The E channel operationalizes the affective signals, and XM turns the predicted divergence into a computable quantity. Whether that quantity corresponds to divergence as a human listener would perceive it is untested (Section 9.2). Cross-modal conflict is a classical perceptual finding [26], which

has been exploited computationally for sarcasm and fake news detection [27,39], as well as multimodal deception detection [40]. Those constructions were not designed for a phone-side always-on budget. XM adapts the same idea to that budget: it weights lexical warmth against prosodic valence and dominance directly, rather than fusing the modalities into a joint embedding.

Finally, the *cheap-first cascade* skeleton is inspired by boosted and soft cascades [41,42], early-exit networks [43], and mobile keyword-spotting cascades [44,45]. The decision to escalate only an uncertain band is grounded in reject-option and selective classification [46–48] and learning-to-defer [49], with calibration [50] making the band meaningful. The escalation target is a small, quantizable LoRA-adapted language model [51] of a class that is now deployable on smartphones [52,53], consistent with TinyML and edge inference [54,55]. Local computation also avoids the potential privacy exposure associated with cloud transcription [56].

3. Related Work: Legacy Vishing Detectors

We group prior detectors by the signal they exploit. For each group we note what was done, what was found, and the limitations that MiLTL targets.

3.1. Text- and Transcript-Centric Detectors

Boussougou and Park anchor the Korean line of research. They built the KorCCViD corpus, a Korean-LM classifier, and an attention-based CNN–BiLSTM with FastText embeddings, later strengthened with back-translation and SMOTE tree ensembles [16,17,57,58]. Yu et al. combined named-entity recognition with *n*-grams [18], and others have detailed real-time detection pipelines [59]. *Contributions:* These works established that Korean vishing transcripts are highly separable and that lightweight models suffice in in-corpus settings. *Limitations:* Their discriminative power rests on lexical features, and, as we show, it does not survive lexical decorrelation. This is shortcut learning [20] and the ID–OOD gap [22,60].

3.2. LLM- and Small-LM Detectors

A newer line of research applies language models. Sim and Kim fine-tuned small LMs with expert vishing criteria and a chain-of-thought approach [61]. Park et al. used GPT-4o for data augmentation and classification [62]. Broader studies have probed the state of LLM-based phone-scam detection [63]. *Contributions:* LLMs raise the text-only ceiling and improve scenario coverage. *Limitations:* These models remain transcript-bound and are directly attackable by LLM-rewritten scripts [14]. Continuous cloud LLM inference is also costly and potentially privacy-exposing.

3.3. Multimodal and Audio Approaches

Multimodal vishing detection systems that integrate text and audio have begun to appear [64]. Generic multimodal deception detection approaches combine audio, video, and text streams, typically using fusion over per-modality features [40] or cross-modal cues such as sarcasm incongruity [39]. Anti-spoofing research targets synthetic and replayed speech [65]. *Contributions:* These studies have confirmed that audio carries complementary information. *Limitations:* Naïve fusion tends to average the modalities rather than contrast them. Our fairness controls show audio-only and naïve text+wave fusion sitting near chance on KorMMP, because neither isolates the *contradiction* between modalities. The established multimodal deception systems target interview or courtroom video with visual streams, and no Korean-capable telephone-speech implementation runnable under our freeze protocol has been released. As the closest feasible reference, we evaluated a modern

multimodal speech–language model directly on KorMMP (Qwen2-Audio-7B-Instruct, zero-shot; Section 8).

3.4. On-Device and Commercial Deployments

Korean carriers have shipped on-device or hybrid detectors [66–68] developed through organized data-sharing approaches [69]. Academic on-device research spans Android systems, KoBERT-based in-call detection, distillation, and roadmaps [19,70,71]. *Contributions:* These studies demonstrate real-time feasibility and commercial demand. *Limitations:* Several approaches transmit call text or summaries to servers, creating the very privacy exposure that on-device detection is meant to avoid [24,25,72,73]. Their detection signal also remains largely lexical.

3.5. Positioning: What MiLTL Composes

Each of these lines supplies a component that MiLTL reuses. Contradiction-aware multimodal modeling supplies the cross-modal channel; cascade and reject-option work supplies the escalate-only routing; edge-inference work supplies the adapted small-model arbiter. We did not find the following three components used together in prior vishing work: a lightweight cross-modal contradiction channel, a deterministic front-end with zero gradient-learned parameters, and an escalate-only arbiter that keeps raw audio and its derivatives on the device. MiLTL is that arrangement, evaluated on a benchmark built to defeat the lexical shortcut that inflated earlier reported metrics.

4. Problem Setting and Hypotheses

We address real-time Korean vishing detection on-device, operating on 8 s nibbles within an observation budget of up to 26 nibbles (≈208 s) per call, under distribution shifts resulting in unreliable lexical cues. We test three falsifiable hypotheses, as well as a channel-level sub-hypothesis (H1b), through an evaluation conducted over five fixed seeds with $n = 500$ per corpus.

H1 (Delivery-aware cascade margin)**.** *Under lexical decorrelation, the proposed cascaded detector that combines lexical, prosodic, and cross-modal channel evidence outperforms the strongest text-only detector on KorMMP. The hypothesis is* falsified *if MiLTL-Cascade does not exceed the strongest text-only baseline by at least* 0.05 *in AUROC.* Metric: *AUROC on KorMMP.*

H1b (XM banding contribution)**.** *Under the fixed operating band, XM widens Gate-1's confident coverage and reduces Gate-2 escalation without increasing confidently missed harm beyond the reported operating trade-off. The hypothesis is* falsified *if removing XM leaves the escalation share unchanged or lowers it, or if XM's coverage gain comes at the cost of additional confidently missed harm.* Metrics: *Band statistics at the deployed calibration stage (Table 7).*

H2 (Cascade gain)**.** *Escalating the uncertain band to Gate-2 improves robustness over the deterministic Gate-1 alone. The hypothesis is* falsified *if MiLTL-Cascade does not exceed MiLTL-L1 (Gate-1 only) under the recording-level clustered paired test at* $p < 0.05$ *(Section 7; the row-level DeLong test is also reported).* Metric: *Paired AUROC on KorMMP.*

H3 (Bounded edge cost)**.** *The per-call inference cost is bounded for on-device use. Gate-1 carries zero gradient-learned parameters at inference, holding 42 stored calibration scalars plus fixed integer weights. Gate-2 adds* at most one *SLM forward pass* per call, *not per nibble, and only when the call-level score falls in the uncertain band.* Metrics: *Inference parameter state, the Gate-2 invocation condition, the measured escalation rate, and measured latency and memory for both stages on a*

commodity CPU container and *on phone hardware (Section 5.4). Instrument-grade on-device energy and integrated end-to-end application latency remain directions for future work (Section 9).*

5. The MiLTL Cascade

We use the following terminology consistently from here on. *MiLTL*, or *the cascade*, denotes the detection system as a whole. *Gate-1* is its deterministic *scorer* and *Gate-2* its escalate-only SLM *arbiter*. We call the comparison methods in Section 8 the *baseline arms*.

5.1. Why Nibbles, Not Keywords or Affect Energy Alone

The design hinges on one choice: the unit of observation. Keyword spotting reduces a call to a bag of tokens, which paraphrase and vocabulary shift defeat by construction [14,20]. Affect energy alone reduces a call to a prosodic envelope, which is ambiguous; excitement, distress, and coercion share acoustic signatures. Accordingly, the audio-only control—Wave-Seq, a prosody-only sequence model defined in Section 6.1—sits near chance in Section 8.

MiLTL instead segments a call into 8 s *nibbles* and reads five channels per nibble. These channels capture *what* is said (T, F), *how* it is said (E), *whether the two disagree* (XM), and *how uncertain* the evidence is (I). The 8 s window sets the timescale. It is long enough for a complete prosodic phrase to co-occur with its lexical content, so a contradiction between them can register. It is short enough that accumulating nibbles yields early, low-latency decisions rather than an end-of-call verdict. Per-nibble multichannel extraction therefore keeps evidence that keyword spotting and affect energy alone each discard.

5.2. Gate-1: Five Channels and a Deterministic Scorer

Gate-1 reads five channels from each nibble:

- **Truth (T)**: Lexical–semantic cooperativity with natural, congruent delivery.
- **Indeterminacy (I)**: Epistemic ambiguity between the lexical and prosodic streams.
- **Falsity (F)**: Alignment with coercive or scam lexical templates.
- **Emotion energy (E)**: Vocal arousal determined from telephone-band audio (300–3400 Hz, μ-law equalized).
- **Cross-modal contradiction (XM)**: Warm lexical content weighted against cold, dominant prosody.

Each channel is a fixed-weight logistic over standardized prosodic and lexical features. T, I, and F are calibrated memberships, not a probability simplex. Section 5.3 details the extraction, and Appendix B (Listing A1) provides the released code for the channels and the five-channel output $[T, I, F, E, \mathrm{XM}]$.

The deterministic risk score combines the channels as follows:

$$\mathrm{risk} = E - 2\,T + I + \mathrm{XM}, \qquad P(\mathrm{harm}) = \sigma\big((\mathrm{risk} - b_0)/s\big), \tag{1}$$

where T enters the risk calculation negatively, because warm, cooperative wording is protective unless contradicted, while E, I, and XM all raise risk. The integer weights (w_T = 2, w_{XM} = 1) are fixed rather than learned by gradient descent. We set them from pretests on KorMMP and KorCCViD against the operating point, meaning band coverage and escalation, rather than against ranking AUROC (Section 9.2). Integer weights are a deliberate design choice. They keep the always-on scorer auditable, so a reader can verify every decision by hand from the channel values, and they need no gradient-fit parameters, which can overfit a small calibration pool. The choice is also not delicate: a sensitivity sweep over the weight neighborhood ($\pm$0.5 around each considered weight) shifted the AUROC by less than 0.03 on the pooled set and hard strata (Section 8.5). The map $\sigma(\cdot)$ is a

calibrated logistic whose bias b_0 and scale s are fixed on the validation calibration set, not on KorMMP.

Two design decisions rest on measured evidence. First, **the scam-word channel F is left out of the score whenever audio is available.** On its own, F barely separates scam from benign calls (near-chance AUROC; Section 8.5). In the same pre-tests it produced most of the false alarms on benign calls whose transcripts are rich in the finance and authority words that scams also use, the vocabulary-rich hard-benign cases. Retaining F flagged too many benign calls, so we drop it when voice audio is available. F enters only in *transcript-only* situations, where the voice channels E and XM cannot be computed. There we add it in a form centered on a typical benign call ($+w_F(F - 0.5)$, $w_F = 2$), so that it raises the score only for unusually scam-heavy wording instead of shifting every call by the same amount. Second, **the score is computed once per call, not once per 8 s segment.** We average each channel value over the call's valid segments and apply the risk formula in Equation (1) to that whole-call average.

Gate-1 then bands the call-level score. A call is escalated to Gate-2 only when $P(\text{harm})$ falls in the uncertain band $[\tau_{\text{low}}, \tau_{\text{high}}] = [0.40, 0.90]$; below it the call is cleared, above it the call is flagged. A call that has not accumulated enough evidence, meaning fewer than 80 words with audio or fewer than 12 words in transcript-only mode, is reported as undecidable rather than scored.

Gate-1 carries *zero gradient-learned parameters at inference*, and its always-on path runs without a neural network. Its only stored state is 42 calibration scalars: 18 feature medians, 18 interquartile ranges, and a bias/gain pair for each of T, I, and F, together with the fixed integer weights. The released algorithm also ships a trained extractor and an optional head, which the deterministic path does not invoke. Table 3 calls this deterministic scorer the Gate-1-only arm (MiLTL-L1); see also Figure 1. Section 8.5 quantifies each channel's role. T is the strongest single channel influencing the ranking, and E prevents confidently missed harm. XM widens the region that Gate-1 can decide alone. F sits near chance on the benchmark, which is why it is excluded whenever audio is available.

5.3. Channel Extraction and Calibration

Each channel is a deterministic function of one nibble, computed from two inputs: the nibble's audio and its transcript. A speech recognition model (`faster-whisper`) produces the transcript. In the released demo it transcribes live when the user supplies audio alone.

5.3.1. Prosody

We decode the audio to 16 kHz mono and analyze each 8 s nibble in 25 ms frames taken every 10 ms. Per frame we measure pitch by autocorrelation, searched over 75–300 Hz, loudness as root-mean-square energy, and spectral shape by fast Fourier transform. Averaging and differencing the frames yields 18 descriptors per nibble: voiced ratio; pitch mean, deviation, slope, and range; energy mean, deviation, and slope; zero-crossing rate; a syllable rate proxy; jitter and shimmer; harmonics-to-noise ratio; spectral centroid and tilt; and three pause statistics. The channels read seven of them: energy mean, pitch mean, pitch range, syllable rate, pause ratio, spectral tilt, and harmonics-to-noise ratio. We compute the other 11 but do not use them in the deployed scorer.

5.3.2. Lexicon

The transcript passes through five counters. We match each nibble's words against five small Korean word lists of 8–11 terms each: directive, urgency, threat, scam subversion, and warmth. Each list contributes a normalized match count clipped to $[0, 1]$. These five counts and the seven prosodic descriptors form the complete channel input. No sentence embedding is involved, so the always-on stage never runs a text encoder.

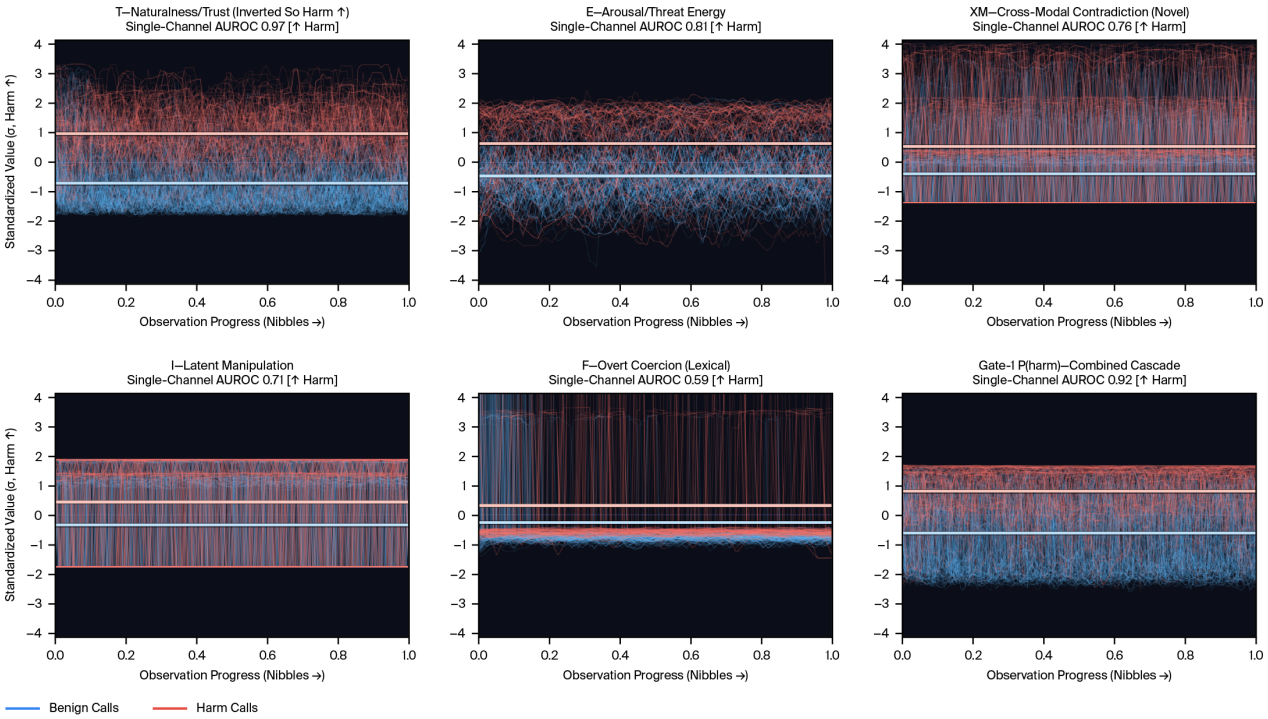


Figure 1. Harm-oriented per-channel discrimination on KorMMP, shown on a shared standardized axis (T inverted for display). Under synthetic decorrelation, the prosody-bound channels hold (E 0.83, XM 0.74), while the lexically bound F inverts (0.98 easy $\rightarrow$ 0.10). The per-channel separation shown here is diagnostic only; the full ablation (Section 8.5) shows the XM contribution to the recomposed score is to banding rather than to ranking.

5.3.3. From Features to Channels

We standardize the prosodic features, centering each by its median and scaling it by its interquartile range, both measured on a benign speech reference. An ordinary benign nibble therefore sits near zero in prosody, and every later quantity reads as a *departure* from ordinary conversation. Three fixed logistics map the standardized features to an arousal–valence–dominance (AVD) description of the voice. Arousal A rises with energy, pitch, and rate. Dominance D increases with energy, pitch range, and rate, but falls with pausing. Valence V rises with a low-frequency-dominant, harmonically clear (warm) voice and falls with a bright, harsh (cold) one. The channels are then formulated as additive evidence over the AVD and lexical signals: F from a cold voice carrying coercive words; the cross-modal contradiction channel $\mathrm{XM} = \mathrm{clip}(\mathrm{warmth} - V, 0, 1) \cdot (0.5 + 0.5D)$, formulated to weight warm lexical content against a cold, controlling voice; I built around XM; and T from a coherent, cooperative delivery. Appendix B details the exact expressions.

5.3.4. Why This Form of XM?

The formulation follows four requirements, not a fit to data. *Directness*: warmth $- V$ is the mismatch itself, namely, lexical warmth minus vocal valence. *One-sidedness*: Only the manipulative direction counts, warm words over a cold voice. The converse is common in benign talk and must not raise risk. *Control gating*: The $(0.5 + 0.5D)$ factor scales the contradiction by vocal dominance, because the target construct is manipulative *control*. *Budget and portability*: Everything follows from the calibrated prosodic features and the small warmth lexicon, with no always-on text encoder. We also tested a trained head over the same channels during development. It underperformed this analytic form, and although it ships in the release, the deployed path does not invoke it. Throughout this paper, cross-modal contradiction is thus an *author-defined operational signal*: motivated by the

deception leakage hypothesis and computable within the edge budget, but not a validated measure of human-perceived incongruence (Section 9.2).

5.3.5. Calibration

Each evidence stream becomes a channel value in $[0, 1]$ through a logistic whose midpoint and slope come from benign percentiles. We weight F and I so that ordinary benign calls read low, and T so that it reads high. These constants are the feature medians and interquartile ranges plus one midpoint/slope pair for each of T, I, and F. They are the only values that must be fitted. We estimate them once on benign training audio and then freeze them; no KorMMP data and no gradient training enter this fit.

5.4. Gate-2: Escalate-Only Single-Pass Arbiter

Gate-2 uses a Qwen2.5-1.5B-Instruct backbone [52] with a LoRA SFT adapter [51] ($r = 8$, $\alpha = 16$, dropout 0.05, target modules $q_{\text{proj}}, v_{\text{proj}}$, AdamW; the exact configuration and prompt are detailed in the released implementation [74]). It runs only on calls whose Gate-1 score falls in the uncertain band. Its prompt is transcript-first: the arbiter judges predominantly from the transcript, with the channel summary, XM leading, appended as *auxiliary*, a deliberately conclusion-free context. A prompt hint can silently dominate a small model, so we measured its effect by ablation. Deleting the channel summary, or replacing it with another call's, left Gate-2's decisions unchanged (Section 8.5, Table 8). Gate-2 emits a binary decision from a *single forward pass* over the yes/no token log-probabilities, with no autoregressive generation.

Lightweight and Measured by Design

The cascade's architecture fixes its cost. Gate-1 is arithmetic over 42 calibration scalars with zero gradient-learned parameters. Gate-2 runs one forward pass of a 4-bit 1.5B model per escalated call: never per nibble, and never continuously. Table 1 reports measurements on a commodity CPU container and on a consumer smartphone. Excluding ASR and cold start, both measured stages complete within the 8 s observation cadence, and Gate-2 adds its ≈1.6 s decision delay only for escalated calls. The 1.5B backbone is the smallest that executes the decision protocol reliably, and adaptation rather than scale determines the size required: a zero-shot 3B backbone did not approach the adapted 1.5B backbone (Table 8). A sustained on-phone stress test of five consecutive benchmark runs, harsher than deployment, shifted throughput by only ≈1%. Container-measured references proxy two things: what consumer iOS does not expose, such as instrument-grade energy; and on-device speech recognition, which needs a developer toolchain to build. We deliberately avoided an Android emulator, which runs on a host CPU.

Table 1. Measured per-call budget; includes neither speech recognition nor cold-start costs. Container: 4-vCPU, fp32. Phone: iPhone 15 Pro Max (A17 Pro, 8 GB, iOS 26.5.2). Gate-1 is the released code run unmodified (median of three runs), Gate-2 is the LoRA-merged 4-bit arbiter (Q4_K_M, 980 MB) timed on the real 801-token prompt under llama.cpp Metal, its on-device decision matching the benchmark. Memory is the app-reported process peak during Gate-2 inference, including the model, runtime, and prompt state, not weights only. The container value is process RSS.

Stage	Model State	CPU Container	iPhone 15 Pro
Gate-1 (always-on, per call)	0 gradient-learned params	0.35 ms · ≈ 45 MB	0.19 ms [†]
Gate-2 (per escalated call)	1.5B, 4-bit ≈ 1 GB	7.3 s (fp32)	≈ 1.6 s · 1 GB peak

[†] Gate-1 phone memory is included in the app process baseline, and is not separately measured.

6. The KorMMP Benchmark

6.1. Construction and Composition

Standard Korean vishing corpora are lexically saturated (Section 1). The real FSS phishing recordings eligible for benchmark construction are few, at 143 calls meeting the duration filter, and mostly vocabulary-dense. We therefore built KorMMP as a controlled stress test that isolates one failure mode a deployed detector would face: lexical decorrelation. It pairs paraphrased or novel scripts, degraded ASR, and legitimate calls whose wording shares the finance and authority vocabulary that scams typically use. It is a diagnostic construction, not a sample of field conditions.

Every case in KorMMP uses a real recording: an FSS phishing call for harm, an AI-Hub dialogue for benign. The harm/benign contrast is carried by the delivery, while the transcript decorrelates vocabulary from the label. Control cases keep the recording's own ASR transcript; hard cases pair the real recording with an independently generated one. Both mismatch directions are included: warm, scam-vocabulary-poor wording carried by cold phishing prosody, and vocabulary-rich finance/authority wording carried by warm benign prosody, such that the label cannot be reliably predicted from the vocabulary. Table 2 details the pooled composition of KorMMP. The synthetic transcripts were drawn from eight harm and eight benign Korean scenario archetypes (harm: institution, loan, family, refund, delivery, acquaintance, investment coaching, and subsidy).

Table 2. KorMMP composition, pooled over five seeds ($n = 500$: 200 harmful, 300 benign). All audio is real (telephone band, μ-law equalized); in 213 instances, the recording's own ASR transcript is retained, while the remaining 287 instances use an independently generated decorrelated transcript. Slices are cut by lexical density, not by label.

Slice	Label	N	%	Audio/Transcript
Control (easy) slices—142 instances				
Easy-harm	Harm	48	9.6	Real FSS/real ASR
Easy-benign	Benign	94	18.8	Real AI-Hub/real ASR
Hard slices—358 instances (72%)				
Hard-harm-natural	Harm	16	3.2	Real FSS/real ASR
Hard-harm-asr	Harm	20	4.0	Real FSS/degraded ASR
Synth-hard-harm	Harm	116	23.2	Real FSS/synthetic
Hard-benign	Benign	35	7.0	Real AI-Hub/real ASR
Synth-hard-benign	Benign	171	34.2	Real AI-Hub/synthetic
Total	—	**500**	100	Real audio; 213 real, 287 synthetic transcripts

Natural vs. Synthetic Hard Cases and What Each Can Show

The hard stratum mixes two kinds of difficulty, and from this point on we report them *separately* (Table 4). The *natural* hard cases (hard-harm-natural, hard-harm-asr, hard-benign) are real calls with their own transcripts. They comprise scam calls that are naturally vocabulary-poor (e.g., grooming and impersonation scripts) or ASR-degraded, and legitimate calls that are naturally rich in finance/authority vocabulary. Their composition places unmodified real scam calls of this lexically atypical kind in the benchmark, namely, 36 of the 200 pooled harm instances. The *synthetic* hard cases pair a real recording with an independently generated transcript. They serve as a *controlled* decorrelation probe, the only way to guarantee by construction that vocabulary cannot predict the label. They also model the attack that motivates the benchmark, in which a scammer re-scripts the call (e.g., with an LLM [14]) while the delivery pressure stays the same. The two strata carry different evidential weight. We demonstrate text-detector *collapse* on the synthetic stratum, where decorrelation is guaranteed. The natural stratum is *not* lexically adversarial, and

frozen text baselines do well on it; it tests the MiLTL behavior on real, unmodified calls. Obtaining naturally occurring calls whose own transcripts defeat a text detector at scale would require new data collection, which we state as future work.

Slices are cut by an observable input feature, lexical density, rather than by the harm label, giving a 2×2 stratification (easy/hard $\times$ benign/harm). Of these, the hard slices carry the diagnostic load. *Hard-harm* reflects scam-vocabulary-poor phishing (e.g., grooming, degraded ASR, or novel scripts) that a lexical detector reads as benign (i.e., false negatives). *Hard-benign* carries vocabulary-rich legitimate finance/authority transcripts over benign audio, which a lexical detector reads as harmful (i.e., false positives). As the split uses an observable feature that is independent of the label and of any detector's output, it follows a standard sub-population shift evaluation approach [21] rather than introducing label leakage.

A boundary statement comes first. This benchmark does not establish that the detected prosodic signal is *uniquely* manipulation specific. Harmful and benign recordings originate from different source corpora: the harm side from real phishing calls archived by the financial regulator, the benign side from real daily dialogue. Residual source, speaker, genre, and interaction-role effects may therefore remain despite telephone-band normalization and provenance auditing. Matched-source collection is the decisive control (Section 9.3).

Within that boundary, we designed the benchmark to test detection ability fairly rather than to favor MiLTL. Four properties keep it honest.

(i) The split is detector-independent. Lexical density sets the easy/hard cut, and no MiLTL output enters it.

(ii) The remaining confound is disclosed, not hidden. Audio source and class label are structurally confounded: *every* harm recording is FSS-sourced and *every* benign recording is AI-Hub dialogue, in the synthetic strata as well. We control what can be controlled and disclose the rest (Section 9.2). All audio for all detectors passes an identical telephone-band plus μ-law codec chain, which removes bandwidth and codec signatures. A provenance audit then checks whether audio alone predicts the label; the audio-only control dropped to AUROC 0.51 once equalization was applied. No component of MiLTL is trained on the KorMMP labels, so MiLTL could not have *learned* a source signature from the benchmark. One baseline arm is an exception, disclosed in full in Section 6.2. The audio-only Wave-Seq control has no KorCCViD audio to learn from, so we fit it on FSS harm audio against KsponSpeech benign dialogue, part of which recurs in the KorMMP evaluation set. That overlap can only act in the baseline's favor, and it still scored AUROC 0.508 (Table 3).

(iii) Two fairness controls see the same two modalities as MiLTL [75,76]. Audio-only **Wave-Seq** is a 1D convolutional network over the prosody sequence. Naïve fusion **MiLTL-Dual** combines a transcript-only detector and Wave-Seq by a max-OR union rather than by their contradiction. Both reach near-chance performance on KorMMP (Table 3), so access to text and audio alone leaves the MiLTL margin unexplained.

(iv) The difficulty applies to MiLTL too. Its Gate-1 uncertainty and residual errors concentrate on the hard slices (Section 8), and on the saturated KorCCViD control it scores lower than the legacy detectors.

The 500 pooled instances are drawn across the 5 seeds from 235 unique real recordings (106 harm, 129 benign) within a 284-recording pool, such that each used recording appears approximately twice on average. The lexical shortcut that KorMMP removes is measurable: the AUROC of a transcript proxy fell from 0.502 on KorCCViD to 0.407 on KorMMP. The scenario banks and the exact per-seed bundles are provided in the open repository.

6.2. Datasets, Splits, and Freeze Protocol

From the frozen KorMMP composition, each of the 5 fixed random seeds draws a separately sampled 100-call evaluation set, balanced across the easy/hard × benign/harm slices. Both corpora therefore provide $n = 500$ observations (100 calls × 5 seeds). We record the seeds, so every draw is exactly reproducible. The seeds resample a shared recording pool, so the 500 observations are *not* independent: they derive from 235 unique source recordings. We therefore compute all uncertainty in this study at the *recording* level (Section 7). Under the frozen cross-corpus protocol, we fit all of the MiLTL trainable components on the training data only: the Gate-2 LoRA adapter on KorCCViD-train plus disjoint synthetic data, and the Gate-1 channel calibration and thresholds on benign training audio. KorCCViD-test and all of KorMMP are held out from every such fit. KorMMP enters none of the MiLTL training, early stopping, or synthetic augmentation, so no trainable component of MiLTL is fit on KorMMP [21,22]. One choice is not covered by this freeze. The Gate-1 score's operating point, meaning the integer weights and the F-exclusion rule, was selected from pre-tests on KorMMP and KorCCViD (Section 5.2, Limitation 1). The freeze protocol supports a controlled validation of architectural generalization. It does not demonstrate environment-agnostic deployment, which remains a direction for future work (Section 9).

6.2.1. Disclosure: The Wave-Seq Baseline's Training Audio Overlaps KorMMP

The freeze above covers MiLTL and the text baselines; the audio-only *Wave-Seq* control is the one arm it cannot cover, as KorCCViD provides no audio on which to fit an acoustic model. We instead fit Wave-Seq on 417 FSS harm recordings against 188 KsponSpeech benign dialogues and then froze it. That training set is *not* disjoint from KorMMP: 46 of the 372 unique KorMMP evaluation calls (12.4%), or 73% of the FSS-sourced harm calls, occur in it. At the audio level the exposure is wider still. A total of 104 of the 235 unique evaluation recordings are FSS recordings Wave-Seq saw in training, because the synthetic-hard-harm stratum pairs new transcripts with those same recordings. The overlap can only act in Wave-Seq's favor, since it gives the baseline label-matched training audio that no MiLTL component received. Wave-Seq nonetheless scored AUROC 0.508 on KorMMP (Table 3), which is chance. The provenance-audit conclusion of Section 6.1, that audio alone does not carry the KorMMP label, therefore survives the overlap and is arguably strengthened by it. Naïve fusion *MiLTL-Dual* contains Wave-Seq and inherits the same overlap; it scored 0.561.

6.2.2. Why Evaluation-Only Is a Necessity, Not a Convenience (Measured)

A natural objection is that frozen baselines are handicapped, and that fairness demands baselines *adapted* to KorMMP. We therefore ran a direct test with a leakage-proof cross-fit. We partitioned instances by connected components of the recording–transcript graph, 235 components in all; a plain recording split would leak, because synthetic transcripts recur across seeds with different recordings. We trained each baseline from scratch on one fold and scored it only on the other, so no model ever saw its evaluation instance's recording or transcript. The adapted baselines saturated the benchmark: KLUE-RoBERTa reached an AUROC of 1.000, CatBoost 0.998, and CNN–BiLSTM 0.985, each ≈1.0 on the synthetic stratum, against 0.29–0.49 for their frozen counterparts. This saturation is diagnostic, not a victory. KorMMP decorrelates scam vocabulary from the label by construction, so vocabulary cannot be what was learned; what remains learnable in the synthetic strata is the transcript *generator's fingerprint*. Three checks confirm that reading. Every one of the 16 scenario-archetype × label groups spanned both folds, which is unavoidable at any instance-level split with 8+8 archetypes. A capacity gradient of shallow probes on the identical split tracks expressiveness exactly as artifact learning predicts (word

unigram logistic 0.60 on the synthetic stratum, character 3-gram logistic 0.78, BERT class 1.0). Prediction from transcript length alone sits at chance (0.53). The conclusion motivates the protocol: *any* sufficiently expressive model saturates a finite synthetic generator when trained inside it, so within-benchmark adaptation is not a sufficient measure of vishing generalization. KorMMP is evaluation-only by necessity. This finding also bounds what the benchmark can measure, and we state the bound explicitly. KorMMP cannot rank models that are adapted to it, and saturation under adaptation is a property of any finite synthetic construction, not a defect unique to this one. What KorMMP does measure is whether a detector that was not trained on it retains its accuracy once the lexical shortcut is removed; every conclusion in this paper is conditional on that protocol. The benchmark cannot measure source invariance either because its harmful and benign audio derive from different source corpora (Limitation 2). Questions outside that scope, above all performance after adaptation on realistic external data, are stated as future work rather than claimed (Section 9.3).

7. Experimental Setup

We report standard binary classification metrics per call, pooled over the five seeds ($n = 500$ calls): AUROC, F1, accuracy (ACC), sensitivity (SEN), specificity (SPE), and positive/negative predictive value (PPV/NPV). AUROC is threshold-free; F1 and the confusion matrix metrics (ACC, SEN, SPE, PPV, and NPV) are reported at the Youden's J operating point (i.e., the threshold that maximizes $\mathrm{SEN} + \mathrm{SPE} - 1$).

7.1. Uncertainty and Significance

The 500 pooled observations derive from 235 unique source recordings on KorMMP, since synthetic cases reuse real audio and seeds resample a shared pool, and from 386 unique calls on KorCCViD. Seed-level averaging or a purely row-level test would therefore treat correlated observations as independent and understate uncertainty. All confidence intervals are recording-level cluster bootstrap 95% intervals. We resampled recordings with replacement ($B = 10{,}000$, fixed seed) and recomputed the full metric pipeline in each replicate, including the Youden operating point. Paired detector comparisons share the resample within a replicate, yielding cluster-robust CIs and two-sided bootstrap p-values for AUROC differences across *all* detector pairs. The DeLong test [77] is retained as a secondary, row-level check. Every number in Section 8 is taken verbatim from the consolidated results and the released analysis outputs. The full environment, the freeze protocol training commands, the five-seed benchmark driver, and a CPU-only interactive demo that reproduces these results are documented in Appendix A.

7.2. Baseline Detectors

Table 3 compares MiLTL-Cascade with eight other arms; the shorthand in its rows is defined here. Every text baseline arm is trained or calibrated only on KorCCViD-train and then *frozen*: its weights are fixed and are not adapted to KorMMP. The audio-only *Wave-Seq* arm is the exception. KorCCViD carries no audio, so we fit it on FSS and KsponSpeech audio, which overlaps the KorMMP evaluation set as disclosed in Section 6.2; it too is frozen before evaluation. The baselines are therefore *transfer references* under the freeze protocol, and the comparison measures out-of-distribution robustness, not superiority over competitors tuned to this benchmark (Section 6.2 shows why tuning baselines to the benchmark itself is uninformative). A "frozen" baseline is thus a conventional text detector evaluated exactly as trained: a KLUE-RoBERTa sentence encoder (`hf`), a CNN–BiLSTM with FastText embeddings, or a gradient-boosted decision tree ensemble. The *lexical (text proxy)* arm uses no learned model; instead, it scores a call according to its scam-keyword density

alone as a pure-vocabulary stand-in for a keyword detector. *Bllossom-B3 (public-finetuned)* is an off-the-shelf, publicly released Korean fine-tune of Llama-3-8B, included as a strong text-only LLM baseline; "public-finetuned" means fine-tuned by its original authors, not by us. The remaining arms are the audio-only *Wave-Seq* and naïve fusion *MiLTL-Dual* controls (Section 6.1), plus *MiLTL-L1*, the Gate-1-only ablation of the cascade.

8. Results

Table 3 reports the metrics obtained by the nine detector arms on KorMMP and the seven text-capable arms on the transcript-only KorCCViD control (the audio-only Wave-Seq and fusion MiLTL-Dual need audio, which KorCCViD does not provide).

Table 3. Detection performance on the lexically decorrelated **KorMMP** benchmark (nine arms) and the saturated **KorCCViD** transcript-only control (seven text-capable arms). AUROC with recording-level cluster bootstrap 95% CI; F1/ACC/SEN/SPE at the Youden point; $n = 500$ pooled over five seeds, resampling of 235 unique KorMMP recordings (386 KorCCViD calls). KorMMP is a source-confounded but provenance-audited stress benchmark; these should not be interpreted as source-invariant field results.

Detector Arm	AUROC	95% CI	F1	ACC	SEN	SPE
KorMMP (hard)						
cnn_bilstm (frozen)	0.658	[0.621, 0.694]	0.677	0.632	0.965	0.410
hf (frozen)	0.641	[0.593, 0.688]	0.701	0.658	1.000	0.430
tree (frozen)	0.638	[0.577, 0.696]	0.520	0.620	0.515	0.690
lexical (text proxy)	0.407	[0.337, 0.476]	0.383	0.684	0.245	0.977
Bllossom-B3 (public-finetuned)	0.684	[0.639, 0.729]	0.624	0.720	0.580	0.813
Wave-Seq (audio-only)	0.508	[0.500, 0.527]	0.576	0.410	1.000	0.017
MiLTL-Dual (naïve fusion text + wave)	0.561	[0.505, 0.620]	0.547	0.602	0.600	0.603
MiLTL-Cascade	**0.965**	**[0.943, 0.984]**	**0.939**	**0.950**	**0.955**	**0.947**
MiLTL-L1 (ablation)	0.933	[0.902, 0.959]	0.840	0.872	0.840	0.893
KorCCViD (saturated control)						
hf (frozen)	1.000	[1.000, 1.000]	1.000	1.000	1.000	1.000
cnn_bilstm (frozen)	1.000	[1.000, 1.000]	0.997	0.998	0.995	1.000
tree (frozen)	1.000	[1.000, 1.000]	1.000	1.000	1.000	1.000
lexical (text proxy)	0.502	[0.429, 0.576]	0.353	0.670	0.225	0.967
Bllossom-B3 (public-finetuned)	0.883	[0.842, 0.921]	0.834	0.876	0.780	0.940
MiLTL-Cascade	**0.890**	**[0.840, 0.934]**	**0.881**	**0.912**	**0.815**	**0.977**
MiLTL-L1 (ablation)	0.866	[0.815, 0.911]	0.805	0.840	0.825	0.850

8.1. Robustness Under Lexical Decorrelation (H1)

On KorMMP, MiLTL-Cascade attained an AUROC of 0.965 (95% CI 0.943–0.984) and F1-score of 0.939 (0.911–0.967). The strongest text-only baseline, the public-finetuned Bllossom-B3 [61], reached only 0.684 (0.639–0.729), while frozen encoders obtained AUROC values of 0.638–0.658 and the lexical proxy performed worst, at 0.407. The margin over the best text-only detector was $0.965 - 0.684 = 0.281$ AUROC (paired cluster bootstrap 95% CI 0.233–0.329, $p < 0.001$; no replicate crossed zero), far exceeding the 0.05 threshold. Therefore, **H1 holds**. The two fairness controls are consistent with the margin being architectural rather than a matter of modality access. Audio-only Wave-Seq reached an AUROC of 0.508 and naïve text+wave fusion (MiLTL-Dual) 0.561, both near chance despite having the same raw modalities as the cascaded detector. A *modern multimodal* reference strengthens this control. Qwen2-Audio-7B-Instruct [78] scored each call zero-shot from the audio and transcript jointly, in a single yes/no forward pass on the same equalized audio with a 30 s cap. It achieved an AUROC of 0.667 (95% CI 0.595–0.735; 0.659 on the

synthetic stratum). That AUROC sits inside the text-only band and below Gate-1 alone, even though the model uses both modalities and carries 4.5× the cascade's total parameters. The general multimodal capability of this zero-shot reference did not match the cascade under the KorMMP freeze protocol. The scope of these results is narrow, and we state it precisely. MiLTL exceeded the *evaluated* frozen, zero-shot, and naïve fusion controls. The test does not establish superiority over all modern learned multimodal architectures, whose fair adaptation requires the external data detailed in Section 9.3. Appendix C gives per-detector case evidence for the failure modes: legitimate finance false alarms, scam-word-free grooming, and ASR-degraded audio.

8.2. Stratified Results: Where the Collapse Happens

Table 4 and Figure 2 break each arm down by stratum, separating the naturally occurring hard cases from the synthetic decorrelation probe (Section 6.1). The pattern is sharp. On the *easy* and *natural* strata the frozen text detectors stayed high (KLUE-RoBERTa, AUROC 1.000 on both), which is expected, as these calls are not lexically adversarial. On the *synthetic* stratum, where vocabulary is decorrelated from the label by construction, the same detectors collapse: KLUE-RoBERTa falls to 0.286, CNN–BiLSTM to 0.494, and the pure-vocabulary proxy to 0.029. The lexical proxy lands below chance because the decorrelation procedure deliberately places scam vocabulary in benign calls and removes it from scams. MiLTL-Cascade is the only arm that held a consistent AUROC across all four strata (0.976–0.983), and its synthetic-stratum score of 0.976 (0.952–0.995) measures H1's claim directly. Two qualifications are merited. The natural-hard stratum is small, at 16 harm calls, so its CIs are wide. We *demonstrated* the collapse of the text detectors on the synthetic stratum only. The natural-stratum results indicate that the MiLTL behavior does not degrade on real unmodified calls within that small stratum. They are not evidence of in-the-wild detection performance, and they do not show that the text detectors fail on those calls.

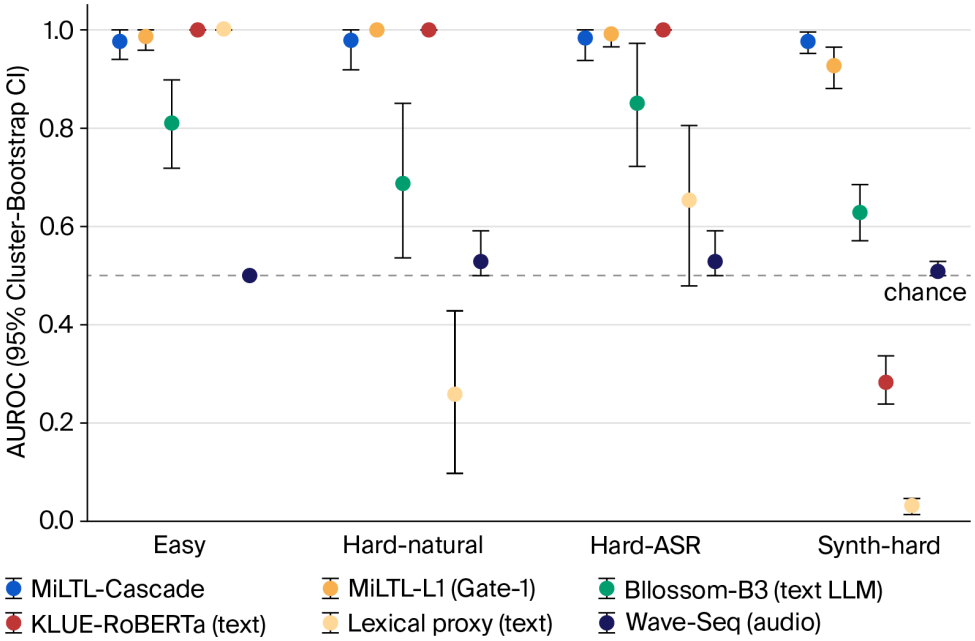


Figure 2. Per-stratum AUROC on KorMMP with recording-level cluster bootstrap 95% CIs; six of nine arms (Table 4 lists all). Points give each arm's AUROC within the four strata (easy, hard-natural, hard-ASR, and synth-hard), with vertical bars spanning the 95% CI and the dashed line marking chance at 0.50. Text detectors collapse only under synthetic decorrelation, with the lexical proxy inverting below chance, while MiLTL-Cascade stays flat across strata.

Table 4. Per-stratum AUROC on KorMMP with recording-level cluster bootstrap 95% CIs. Each harm slice is paired with its class counterpart (the three hard-harm strata share the real hard-benign side). Easy $n = 142$; hard-natural $n = 51$ (16 harm, wide CIs); hard-ASR $n = 55$; synth-hard $n = 287$.

Detector	Easy	Hard-Natural	Hard-ASR	Synth-Hard	Pooled
MiLTL-Cascade	**0.976 [0.940, 1.000]**	**0.979 [0.919, 1.000]**	**0.983 [0.937, 1.000]**	**0.976 [0.952, 0.995]**	**0.965 [0.943, 0.984]**
MiLTL-L1 (Gate-1 only)	0.986 [0.958, 1.000]	1.000 [1.000, 1.000]	0.991 [0.965, 1.000]	0.926 [0.881, 0.964]	0.933 [0.902, 0.959]
Bllossom-B3 (LLM, text)	0.810 [0.718, 0.898]	0.688 [0.536, 0.850]	0.850 [0.722, 0.972]	0.628 [0.571, 0.685]	0.684 [0.639, 0.729]
KLUE-RoBERTa (text)	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.286 [0.239, 0.336]	0.641 [0.593, 0.688]
CNN–BiLSTM (text)	0.977 [0.945, 0.998]	1.000 [1.000, 1.000]	0.964 [0.893, 1.000]	0.494 [0.479, 0.506]	0.658 [0.621, 0.694]
Tree ensemble (text)	0.811 [0.705, 0.903]	0.821 [0.644, 0.933]	0.849 [0.715, 0.950]	0.426 [0.358, 0.493]	0.638 [0.577, 0.696]
Lexical proxy (text)	1.000 [1.000, 1.000]	0.255 [0.097, 0.428]	0.649 [0.479, 0.805]	0.029 [0.014, 0.046]	0.407 [0.337, 0.476]
Wave-Seq (audio)	0.500 [0.500, 0.500]	0.529 [0.500, 0.591]	0.529 [0.500, 0.591]	0.509 [0.500, 0.528]	0.508 [0.500, 0.527]
Naïve fusion (text + wave)	0.642 [0.517, 0.775]	0.577 [0.337, 0.817]	0.653 [0.454, 0.859]	0.415 [0.351, 0.482]	0.561 [0.505, 0.620]

8.3. Cascade Gain and Ablation (H2)

Removing Gate-2 (yielding the MiLTL-L1 ablation, Gate-1 only) decreased the KorMMP AUROC from 0.965 to 0.933. The paired recording-level cluster bootstrap results indicate a gain attributable to Gate-2 (95% CI 0.010–0.058, two-sided bootstrap $p = 0.003$; row-level DeLong $p < 0.001$ agrees). Therefore, **H2 holds** under clustering. The gain is concentrated where it should be: on the synthetic decorrelation stratum, the cascaded detector lifts Gate-1 from 0.926 to 0.976 (Table 4). The Gate-2 ablation test (Section 8.5) decomposes the gain further. The SFT adapter carries it *entirely*: a zero-shot backbone in the escalate band yields 0.834, *below* Gate-1 alone. The channel summary hint makes no measurable contribution. Gate-1 alone already exceeds every text-only and single-modality baseline (0.933 vs. $\leq$ 0.684), which places most of the detector's robustness in the deterministic channel front-end.

Gate-1's AUROC coexists with an escalation rate near two-thirds because AUROC and the escalation band measure different things. AUROC is threshold-free and reports only how well the scores *order* harmful above benign calls. The band is a statement about *confidence*. A scorer can order calls almost perfectly while placing most of them in the middle of the probability scale, inside the class-overlap region that no fixed threshold separates safely. Gate-1 is calibrated to commit only at the extremes and to defer otherwise.

Nothing reviews the calls Gate-1 decides alone, so we audited whether those decisions were correct. Table 5 reports the two committed bands on KorMMP. Gate-1 leaked 1 of the 200 harmful calls (0.5%) and raised 7 false alarms among the 300 benign calls (2.3%). The escalated remainder was mixed at 28% harmful, and Gate-1 referred those calls rather than deciding them. It acts alone only where it is nearly always correct, and hands the overlap region to Gate-2.

Table 5. Gate-1's two committed bands on KorMMP ($n = 500$, five seeds). Gate-1 decides the cleared and flagged bands alone; the uncertain band is referred to Gate-2; the combined confident-band error is 8/179 (4.5%), matching Table 7.

Gate-1 Band	Calls	Correct	Error
Cleared (benign)	63	62 benign	1 harmful leaked (1/63, 1.6%)
Flagged (harmful)	116	109 harmful	7 false alarms (7/116, 6.0%)
Escalated to Gate-2	321	—	28% harmful base rate (referred)

8.4. The Saturated Control and Memorization (KorCCViD)

On KorCCViD, the frozen text encoders scored perfectly. Those perfect scores reflect vocabulary saturation, not generalization: the same encoders collapsed on KorMMP. MiLTL-Cascade scored lower on this control, at 0.890. We report that lower score as a feature: a detector whose signal is non-lexical should *not* win a memorization contest. The gap

between the MiLTL score on the hard benchmark and its lower score on the saturated one inverts the legacy pattern, in which detectors score near-perfectly in-corpus yet collapse out-of-corpus [20].

8.5. Channel Ablation, Sensitivity, and Error Analysis

8.5.1. Channel Ablation

Table 6 reports the Gate-1 risk score with each channel removed or isolated, reconstructed exactly from the frozen per-call channel values. We state the findings as they are. First, **T is the load-bearing channel**: removing it causes the only catastrophic degradation (0.933 → 0.835 pooled; 0.927 → 0.834 on the hard strata). T is not a *purely* lexical feature: its calibrated membership fuses prosodic arousal–valence–dominance evidence, lexical warmth, speech acts, and an arousal–coherence term. By itself, $-T$ ranks on par with the full formula (0.972 pooled, 0.966 hard; overlapping CIs with the cascaded detector). Second, **the explicit I and XM terms do not add to the ranking accuracy**: dropping either slightly *raises* the raw AUROC. The ablation removes the explicit XM term only. XM also enters the score through I, which is built around it, and through T, both in the same risk-raising direction. These figures therefore understate the score's total dependence on XM. Removing either term withdraws much the same evidence, which is why the *Without I* and *Without XM* variants score almost identically. The measured contribution of I and XM lies where the design places it: in the *banding* that turns a score into a decision (Table 7).

Table 6. Gate-1 channel ablation: AUROC of the recomposed risk score with channels removed or isolated (recording-level cluster bootstrap 95% CIs), pooled and on the combined hard strata. Removing T results in the greatest degradation; removing I or XM slightly *raises* the ranking AUROC, while their measured role is in banding performance (Table 7).

Score Variant	Pooled (n = 500)	Hard Strata (n = 358)
Full: $E - 2T + I + \mathrm{XM}$	0.933 [0.902, 0.959]	0.927 [0.889, 0.959]
Without T	0.835 [0.789, 0.877]	0.834 [0.778, 0.885]
Without E	0.917 [0.885, 0.946]	0.914 [0.874, 0.948]
Without I	0.952 [0.925, 0.974]	0.944 [0.909, 0.971]
Without XM	0.953 [0.926, 0.975]	0.945 [0.911, 0.972]
$-T$ alone	0.972 [0.951, 0.988]	0.966 [0.938, 0.986]
E alone	0.820 [0.751, 0.883]	0.827 [0.753, 0.893]
XM alone	0.767 [0.717, 0.816]	0.748 [0.684, 0.804]
I alone	0.715 [0.663, 0.766]	0.673 [0.610, 0.734]
F alone	0.535 [0.469, 0.597]	0.287 [0.222, 0.353]

Table 7. Banding roles of the channels on KorMMP ($n = 500$). Band statistics at the canonical band $[0.40, 0.90]$ when the risk score is recomposed without a given channel, with the deployed calibration held fixed (the stored p_1 sigmoid is inverted exactly; maximum residual 4.41×10^{-3}). Re-deriving the calibration per variant would require Gate-2 re-runs and is deferred as future work.

Score Variant	Escalated	Confident Calls	Confident Errors	Harm Missed
Full: $E - 2T + I + \mathrm{XM}$	64.2%	179	4.5%	1
Without XM	79.2%	104	1.0%	1
Without E	57.8%	211	5.2%	11
Without I	46.4%	268	5.6%	15
$-T$ alone	0.0%	500	40.0%	200

With the deployed calibration held fixed, removing E left 11 harmful calls confidently dismissed as benign, against 1 with E; E is the safety channel. Removing XM shrank the confidently decided region from 179 to 104 calls and raised the escalation share from 64% to 79%, about 23% more Gate-2 invocations. XM is therefore the coverage channel: it lets

Gate-1 decide alone more often, which is a meaningful trade-off on-device. This is exactly H1b's content. With XM present, the escalation share falls from 79% to 64% (confident region 104 → 179 calls), the confidently missed harm is unchanged at 1, and the cost is seven confident false alarms. **H1b holds** at the reported trade-off. A raw channel score gives no such decision structure: under the calibrated band, $-T$ alone confidently discards every harmful call. We therefore claim XM as a *banding and escalation* signal and as Gate-2's cross-modal hint, not as a ranking improvement. This corrects our own earlier framing, which overstated the latter.

8.5.2. Gate-2 Ablation

For Table 8 we re-decide the 321 escalated calls under different Gate-2 variants and recompose the cascade's score. The recomposition is exact, because Gate-2 depends only on the stored transcript and channel diagnostics; the canonical variant's re-scoring agrees with the stored benchmark decisions on 99.1% of calls. Three findings follow. (i) **The adapter carries the entire gain**. Zero-shot backbones yield AUROC scores of 0.812–0.882, all *below* Gate-1 alone, so un-adapted escalation would not help. With the adapter, the re-scored cascade reaches 0.968 (canonical stored decisions: 0.965). (ii) **The channel hint is decision-inert**. Deleting the summary, or replacing it with another call's diagnostics, alters the composed AUROC by at most 0.0003. We adopted the conclusion-free summary design during development to stop the arbiter from parroting a verdict, and the arbiter evidently disregards it almost entirely. We therefore claim the summary as providing auditability, not accuracy, and the measured contribution of XM to the system remains exactly its Gate-1 banding role (Table 7). (iii) **Adaptation beats scale**. Zero-shot 3B (0.882) performed far worse than the adapted 1.5B (0.968), which reinforces the backbone choice.

Table 8. Gate-2 ablation. Composed cascade AUROC on KorMMP ($n = 500$; recording-level cluster bootstrap 95% CIs) when the 321 escalated calls are re-decided by Gate-2 variants (confident calls keep their Gate-1 decision). The canonical variant's re-scoring agrees with the stored decisions on 99.1% of escalated calls. Reference: canonical cascade, 0.965; Gate-1 alone, 0.933.

Gate-2 Variant (Escalated Band Only)	Cascade AUROC [95% CI]
1.5B + adapter, channel hint (canonical)	0.968 [0.947, 0.985]
1.5B + adapter, *no* channel hint	0.968 [0.947, 0.985]
1.5B + adapter, *shuffled* channel hint	0.968 [0.947, 0.985]
1.5B zero-shot (no adapter)	0.834 [0.792, 0.873]
0.5B zero-shot	0.812 [0.766, 0.856]
3B zero-shot	0.882 [0.845, 0.916]

8.5.3. Sensitivity

The scorer is not delicately tuned. We swept w_T over $[0, 3]$ and w_{XM} over $[0, 2]$. The ± 0.5 neighborhood of the published $(2, 1)$ yields pooled AUROC values of 0.907–0.949 (published cell 0.933) and 0.903–0.942 on the hard strata: a flat plateau, not a spike. The escalation band behaves monotonically in the computable, narrowing direction. Shrinking $[\tau_{\mathrm{low}}, \tau_{\mathrm{high}}]$ from $[0.40, 0.90]$ to $[0.40, 0.70]$ halves the escalation rate (64% → 32%), while the cascade's AUROC falls from 0.965 to 0.939 and the confident-band error rises from 4.5% to 25%. The published band therefore buys confident-decision purity at the cost of escalation. Widening it would escalate calls whose Gate-2 posterior was never computed, which we defer as future work.

8.5.4. Error Analysis

At the pooled Youden operating point, the cascade's confusion matrix is TP 191/FP 16/FN 9/TN 284. The errors are asymmetric by stratum. All 16 false positives are synthetic

hard-benign calls, that is, vocabulary-rich legitimate transcripts, and they make up 9.4% of that slice. All nine false negatives are *real* harm recordings: five easy-harm, two ASR-degraded, and two natural-hard. Scenario archetypes give a second view of the same errors. Archetype labels exist only for the 287 synthetic hard calls, which are drawn from the eight harm and the eight benign archetypes alike. Within that population every one of the 16 errors falls in a benign archetype, and every harm archetype is error-free, delivery impersonation included. Across the benign archetypes the errors are thinly spread. The heaviest is legitimate delivery notification, which accounts for 5 errors out of its 29 calls, so no single scenario dominates. The errors concentrate in the uncertain band, precisely where Gate-2 is invoked; the escalation rate is 64.2% pooled, ranging from 33.6% on synthetic hard-harm to 95.9% on synthetic hard-benign. As Section 8.3 noted, Gate-1 is almost always correct on the calls it settles by itself, which is the intended behavior of a recall-first screener on a benchmark designed to be ambiguous.

8.5.5. Why the Two Stages Are Complementary

The two stages read different evidence, so they capture different information. Gate-1 summarizes the whole call in five values that describe *how* it is spoken. It cannot see *what* is being asked. To Gate-1, a warm, fluent scam and a warm, fluent legitimate consultation look almost the same, and these are the calls that fall in the uncertain band. Gate-2 fills that gap by reading the transcript itself. It looks for the concrete tactics of fraud: someone posing as a bank, a prosecutor, or a family member; pressure to move money to a "safe account"; demands for passwords or verification codes; and threats or refund and loan lures. The channel summary (XM first) is passed along only as a hint, and the arbiter is told that a low XM score does not excuse a transcript full of impersonation or transfer pressure. The two stages therefore read *disjoint* evidence (Gate-1 the delivery, Gate-2 the content), which is why adding Gate-2 improves on Gate-1 alone (Section 8.3) rather than re-scoring the same evidence. The division is measured, not assumed: the ablation above leaves the arbiter's decisions unchanged when the hint is deleted.

8.5.6. Behavior Under Shift

The arbiter's decision criteria are scam *mechanics* rather than surface vocabulary. We fitted the adapter only on KorCCViD-train and disjoint synthetic data, never on KorMMP, yet it converts band cases well enough to raise the cascade's AUROC above Gate-1 alone on that benchmark. That result is evidence that the hand-off transfers, not a demonstration of invariance to the deployment environment (Section 9). Appendix C presents representative KorMMP calls on which each legacy family errs while MiLTL is correct, among them a real recording whose broken ASR defeats the transcript-only detectors but not the delivery-aware cascade (Table A1).

9. Discussion and Limitations

This study's results are consistent with a single thesis. On this controlled decorrelation benchmark, robustness comes from the *arrangement* of the components, meaning calibrated channels that fuse prosodic and lexical evidence together with contradiction-aware escalation, rather than from lexical memorization or raw modality access. The cascade featurizes delivery-borne attributes that a transcript classifier cannot represent, and that a listener may not register in the moment. One caveat qualifies that thesis, as argued in Section 6.1 and Limitation 2: the source confound leaves open how much of the detected signal is manipulation-specific. Section 8.5 attributes the robustness channel by channel and stage by stage, and its main finding is that Gate-1, which has zero gradient-learned parameters, already outperformed every text-only baseline. In that attribution the fused T member-

ship carries the ranking, E guards against confidently missed harm, and XM contributes coverage rather than ranking.

Cost follows from the same arrangement: Gate-1 runs as fixed arithmetic over its stored calibration scalars, and Gate-2 costs at most one forward pass per call, so the per-call computational burden is bounded by construction (H3). The bound derives from the escalation *unit*, not from an assumption that escalation is rare, and it therefore holds even on the hard slices where most calls are routed to the arbiter. We quantified it on physical phone hardware as well as on a reference container (Section 5.4, Table 1), within the counters that a consumer iOS device exposes. Both figures are detector-side only and exclude speech recognition and cold start. The CPU-only interactive demo in Appendix A.3 shows the same profile on any machine, reporting per-stage latencies and process CPU/RAM occupancy in real time.

On-device operation is not merely an efficiency choice. Keeping raw audio and its derivatives on the phone addresses the privacy exposure that has troubled server-side commercial services [24,56]. This is a property of the detector as specified, not a measured deployment. We ran the transcription front-end in a container rather than building it for the device (Section 5.4), so a fully local end-to-end pipeline remains to be demonstrated (Section 9.3). The channel scorer is also scalable, and providers could re-calibrate it on larger curated pools. We state any environment-agnostic claim as a *thesis requiring validation in future work*, not as a demonstrated result: the present evidence comes from a controlled, frozen validation on two Korean corpora.

9.1. Language Dependence and Portability

Concerns about generality reduce to one question: which parts of MiLTL are Korean, and which are not? The *language-bound* components are small and explicit: the five lexical word lists (directive, urgency, threat, scam subversion, warmth; 8–11 Korean terms each), the speech-act lexical proxy, and the calibration constants (medians, IQRs, and per-channel bias/gain), which were fit on Korean benign speech. Everything else is *language-independent by construction*: the 18 prosodic descriptors and their extraction, the AVD mapping, the XM formula, the risk equation and banding, the cascade logic and the calibration *procedure* itself, which runs on any benign speech pool. Porting MiLTL to a new language therefore takes three steps: (i) translate or re-source the five short word lists, (ii) re-run `fit_calib` on benign speech in the target language and telephony condition, and (iii) swap the Gate-2 backbone or adapter for one competent in that language. None of these steps involves gradient training of the always-on path. We state this as a design property, not a result; we have run no non-Korean evaluation (Limitation 5).

9.2. Limitations

The principal limitation of this work concerns construct validity. The cross-modal contradiction signal is an author-defined operationalization that has not been validated against human perception; no rating study was conducted, and one was not within reach in this revision cycle. We state it first because it bounds how the cross-modal channel should be named and interpreted throughout the paper. It is developed as item 9 below, alongside the remaining limitations.

1. **Benchmark scope.** KorMMP is a *controlled stress test* and contains no naturally paired call data. Its synthetic hard cases pair real audio with an independently generated, lexically decorrelated transcript. The transcript is therefore not a faithful transcription of the recorded speech, and the lexical–prosodic contradiction is constructed. We separate the strata in all reported results (Section 8.2). We demonstrate text-detector collapse on the synthetic stratum; the naturally occurring hard stratum, which is

admittedly small, captures those detectors' behavior on real unmodified calls. Read the KorMMP results as an indication of robustness under controlled lexical decorrelation, not as a measurement of in-the-wild vishing detection on naturally paired calls, which remains a direction for future work. KorMMP also cannot rank models that are adapted to it, because any sufficiently expressive model memorizes the finite synthetic generator. Every result reported here therefore uses an evaluation-only protocol (Section 6.2). The protocol has one exception. The Gate-1 score's operating point, meaning its integer weights and the rule that drops F whenever audio is available, was selected from pre-tests on KorMMP and KorCCViD, so the reported figures are not a fully out-of-sample estimate of that choice.

2. **Source confound.** We drew harm and benign audio from different sources, so the confound is *structural*: every harm recording is FSS-sourced and every benign recording is AI-Hub dialogue, in the synthetic strata as well. Uniform telephone-band plus μ-law equalization and the provenance audit removed the bandwidth and codec signature, leaving an audio-only AUROC of 0.508. No component was trained on KorMMP labels, so no detector could have *learned* a source signature from the benchmark. Residual speaker, genre, affect, or conversational-role differences between the pools cannot be excluded, so part of the acoustic signal may still be source-related rather than manipulation-related. The decisive control is natural pairs with matched speakers and recording chains in both classes, which requires further data collection and is planned as future work.
3. **Recording-level dependence.** The 500 pooled evaluations derive from 235 unique recordings, so all intervals are recording-level cluster bootstrap CIs with shared resamples (Section 7). The residual limitation is the recording pool itself, which no resampling scheme can widen.
4. **Cascade attribution.** We decomposed the cascade gain on both sides (Section 8.5, Table 8). The one cell we did not run is an adapter *retrained* per backbone size. Since adaptation dominates scale, we do not expect that cell to change the choice of Gate-2 architecture.
5. **Language coverage.** We evaluated on Korean-only corpora and covered neither cross-lingual nor code-switched speech (Section 9.1 maps which components are language-bound).
6. **Edge-cost measurement.** We measured edge cost on a commodity CPU container and on phone hardware, within the counters a consumer iOS device exposes (iPhone 15 Pro Max: Gate-1 0.19 ms/call; Gate-2 $\approx$1.6 s per escalated decision at 1 GB peak, $\approx$1% sustained load drift; Section 5.4). Four things remain unmeasured: instrument-grade energy, Gate-1 OS-level memory, on-device ASR (which we expect to dominate the live budget), and Android-class devices. End-to-end integrated application measurement requires service-level engineering, which lies beyond this paper's scope.
7. **Calibration drift.** We fixed the calibrated parameters and thresholds after fitting; they may need recalibration under operational drift.
8. **Statistical power.** With $n = 500$, subgroup analyses are underpowered. The per-stratum CIs in Table 4 quantify this, and the natural-hard stratum has only 16 harm calls. We report per-scenario error counts descriptively, with no significance claims.
9. **Construct validity of XM.** This is the study's central unresolved question. The AVD and XM constructs are author-defined operationalizations. Explicit design requirements (Section 5.3) and their measured banding roles (Section 8.5) justify them, but they have not yet been validated against independent human ratings, and an attacker who matches prosody to wording could weaken the signal. One consideration bears on what such a study could establish, though it does not diminish this limitation.

The harmful recordings used here are real phishing calls drawn from a regulator archive, and the outcome of each call is not disclosed to us, so we cannot say whether human listeners did or did not perceive the incongruence these recordings carry. A rating study is planned work. It would speak directly to whether "contradiction" is the right name for this quantity, and it may also bear on the channel's detection value.

9.3. Future Work

The following require new data, GPU training, or hardware, and are stated as commitments rather than claims:

1. Complete the on-device budget assessment beyond what consumer iOS exposes (Section 5.4), covering instrument-grade energy, on-device ASR real-time factors, and Android-class devices. Integrate the cascade into a callable application for end-to-end measurement.
2. Conduct an adapted comparison on *external* adaptation data (i.e., temporally unseen FSS recordings and additional benign pools). The within-benchmark cross-fit in Section 6.2 shows that adaptation on KorMMP itself measures generator memorization, so a meaningful adapted comparison requires data outside the benchmark. A learned late-fusion baseline that stacks text, prosodic, and audio embedding scores belongs to the same external-data protocol. We have already reported a zero-shot multimodal speech–language reference (Section 8).
3. Collect naturally paired and matched-source call data: larger collections of natural mismatch cases, and harm and benign audio drawn from the same speakers and recording chains, which would resolve the source confound (Limitation 2).
4. Validate the XM output against independent human perception of content–delivery mismatch, in a follow-up study designed from the outset to carry that perceptual dimension.
5. Retrain the Gate-2 adapter under various backbone sizes to extend the sweep reported in Table 8 (its zero-shot cells already show that adaptation dominates scale).
6. Extend the channel definitions and calibration process to additional languages and telephony conditions (Section 9.1), strengthen XM against adversarial prosody mimicry and synthetic voice attacks [13,65], and re-fit the channels on larger curated pools to test the training-scalable thesis.

10. Conclusions

We introduced KorMMP, a lexically decorrelated Korean multimodal vishing benchmark, and MiLTL, an on-device-oriented cascaded detector whose deterministic first stage combines prosodic, lexical, and cross-modal evidence and selectively invokes an adapted small language model for uncertain calls. MiLTL composes well-understood parts, and its value lies in the arrangement: a front-end a reader can audit by hand, an explicit rule for when the system may decide alone, and a bounded escalation cost. What is *demonstrated*: Under the freeze protocol, MiLTL-Cascade achieved the highest AUROC and F1-score on KorMMP of any arm evaluated. No evaluated comparator system came close, whether text-only, audio-only, naïve fusion, or the 7B multimodal reference. The detector stayed flat across the benchmark's easy, natural, ASR-degraded, and synthetic strata, while the text detectors degraded markedly on the synthetic decorrelation probe. The cascade's gain over Gate-1 alone survived recording-level clustering (CI 0.010–0.058). The ablations attribute the robustness to specific components: the fused T membership carries the ranking, E guards the confident band against confidently missed harm, and the cross-modal contradiction channel (XM) improves confident coverage and steers escalation. On cost, the always-on stage runs within a fraction of a millisecond per call, at 0.35 ms on a commodity CPU container and 0.19 ms on a consumer iPhone, with zero gradient-learned parameters.

Gate-2 adds at most one small-model forward pass per escalated call. These figures exclude speech recognition, which we expect to dominate the live budget.

What has *not yet been demonstrated*, stated as commitments in Section 9.3, is as follows: in-the-wild detection on naturally paired calls; comparison against baselines adapted to *external* data, since adaptation within the benchmark measures generator memorization (Section 6.2); validation of XM against human perception; instrument-grade energy and Android-class hardware; end-to-end integration including on-device speech recognition; and non-Korean deployment.

Three boundaries qualify all of the above. KorMMP is a controlled stress test of lexical decorrelation rather than a proxy for field performance. Because its harmful and benign audio derive from different source corpora, part of the detected acoustic signal may be source-related rather than manipulation-related, and the decisive matched-source control requires new data collection. And the validation of the cross-modal construct against human perception remains this study's principal open limitation, not merely a direction for future work. Within the stated scope, the cascade offers a reproducible, inspectable basis for on-device vishing defense that keeps raw audio local by design. That claim remains subject to the integrated end-to-end measurement, including on-device speech recognition, that we state as future work. Alongside it we provide a stratified, uncertainty-quantified benchmark for studying robustness to vocabulary shift.

Author Contributions: Conceptualization, B.J.; methodology, B.J.; software, B.J.; validation, B.J. and J.C.; formal analysis, B.J.; data curation, B.J.; writing—original draft preparation, B.J.; writing—review and editing, J.C.; visualization, B.J.; supervision, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was written with the support of a research grant from the Seoul School of Integrated Sciences and Technologies (aSSIST).

Institutional Review Board Statement: Not applicable. This study did not involve new experiments on humans or animals; it uses only pre-existing corpora obtained under their respective data licenses (see the Data Availability Statement).

Informed Consent Statement: Not applicable. No new human-subject data were collected; all speech data originate from pre-existing corpora used under their source licenses.

Data Availability Statement: The MiLTL source code and the derived, non-identifying artifacts (channel definitions, the deterministic scorer, evaluation scripts, and the consolidated results underlying every reported number) are openly available at https://github.com/mrbong77-sys/MiLTL_Open (accessed on 25 August 2026) and archived on Zenodo (DOI: 10.5281/zenodo.21546202) [74]. The KorMMP benchmark is derived from Financial Supervisory Service (FSS) vishing recordings and AI-Hub benign corpora; raw call audio is retained only on local infrastructure under the source data licenses and is *not* redistributed. The FSS text corpus is included, whereas AI-Hub-derived text is represented by pointers and SHA-256 fingerprints rather than content, so that it can be rebuilt from the original source under its own license and verified byte-for-byte.

Acknowledgments: Use of AI tools. In accordance with MDPI's AI ethics policy, the authors disclose the AI models used. *As components of the released system and benchmark:* Qwen2.5-1.5B-Instruct as the Gate-2 arbiter backbone; a KLUE-RoBERTa Korean sentence encoder for the frozen text encoder baseline; Whisper (`faster-whisper`) for offline speech-to-text transcription of the audio corpora; FastText Korean embeddings for the CNN–BiLSTM baseline; and a public fine-tuned Llama-3-8B Korean model (Bllossom), reproduced solely as a baseline. *As general authoring assistants:* Anthropic Claude (Claude Opus 4.8) was used to help with software development, benchmark and evaluation tooling, and the drafting and editing of this manuscript. All experimental data, results, and reported values are genuine outputs of the authors' pipeline; the authors have reviewed and verified all AI-assisted output and take full responsibility for the content. No AI tool is listed as an author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MiLTL	Multimodal In-call Lightweight Threat Locator
KorMMP	Korean Multimodal Manipulation-Phishing benchmark
KorCCViD	Korean Call Content Vishing Detection corpus
FSS	Financial Supervisory Service
ASR	Automatic Speech Recognition
XM	Cross-modal contradiction channel
T, I, F	Truth, Indeterminacy, Falsity (neutrosophic channels)
E	Emotion energy channel
AVD	Arousal–Valence–Dominance (affect intermediate)
SLM	Small Language Model
LoRA	Low-Rank Adaptation
SFT	Supervised Fine-Tuning
AUROC	Area Under the Receiver Operating Characteristic curve
IQR	Interquartile Range
OOD	Out-Of-Distribution

Appendix A. Reproduction and Interactive Demo

Everything needed to reproduce the reported numbers is available from an open repository (https://github.com/mrbong77-sys/MiLTL_Open (accessed on 25 August 2026), archived at Zenodo [74]). Two tiers are supported: a GPU-free tier that verifies the pipeline logic, result consolidation, the deterministic synthetic corpus, and the provenance/decorrelation audits; and a full tier that rebuilds the data pools, trains the components, and runs the five-seed benchmark.

Appendix A.1. Environment and GPU-Free Verification

With Python $\geq$ 3.10, the regression suite and deterministic synthetic generator can be run on any machine with no external data, using the following code:

```
pip install -r requirements.txt
python -m unittest discover -s tests
python scripts/synth_edgecases.py --n-per 40 --seed 42 --out synth42.jsonl
python scripts/consolidate_results.py --dir artifacts/rounds/canonical
```

Appendix A.2. Full Reproduction

The Gate-2 LoRA SFT and the five-seed benchmark are driven by single orchestrators under the freeze protocol (train on KorCCViD-train plus disjoint synthetic data; no KorMMP in training):

```
python scripts/gate2_sft.py \
    --train-bundle artifacts/rounds/train_fair_99.jsonl \
    --band all --fair-audio \
    --gate2-model Qwen/Qwen2.5-1.5B-Instruct \
    --out artifacts/models/gate2_adapter_1.5b --epochs 3 --device cuda
python scripts/canonical_bench.py --rounds 5 --total 100 --device cuda \
    --ckpt artifacts/models/channel_extractors.pt \
    --gate2-adapter artifacts/models/gate2_adapter_1.5b
python scripts/audit_audio_provenance.py ...  # source-shortcut stays low
python scripts/lexical_shortcut.py ...        # decorrelation fails on hard
```

Prosody/XM reproduction requires the original audio (FSS crawl and AI-Hub download); this is a licensing constraint of the source corpora, not a technical one, and every benign transcript is verifiable against the shipped SHA-256 pointer manifest.

Appendix A.3. Interactive Demo

A standalone, CPU-only web demo runs as `python demo/miltl_demo.py`; Gate-1 needs numpy only, and Gate-2 attaches the released LoRA adapter automatically. It renders the live per-nibble channel traces, Gate-1 banding, Gate-2 activation, and per-stage latencies. It accepts pasted transcripts, wave uploads with real prosody plus automatic Korean ASR, and bundled benchmark cases quoted verbatim from the five-seed benchmark. `bash build.sh` regenerates the manuscript's results tables, stratified figure, and bibliography from its single-source-of-truth data: the consolidated results table, the cluster bootstrap analysis outputs, and the citation allow-list. In this compact version the regenerated tables are inlined and restyled rather than included at compile time. The release pins its content in two ways for artifact integrity. AI-Hub-derived text ships as SHA-256 pointer manifests, verifiable byte-for-byte against the original source. The Zenodo deposit (DOI 10.5281/zenodo.21546202) immutably archives the exact released bytes of the code, bundles, and calibration artifacts.

Appendix B. Core Architecture: The Cross-Modal Channels and the Cascade

The MiLTL always-on path is analytic and inspectable. Listing A1 is the released implementation of the neutrosophic/affect channels, including the cross-modal contradiction signal $\mathrm{XM} = \mathrm{clip}(\mathrm{warmth} - V, 0, 1) \cdot (0.5 + 0.5D)$—warm lexical content (warmth) weighted against cold, dominant prosody (low valence V, high dominance D)—and the five-channel output $[T, I, F, E, \mathrm{XM}]$. Memberships are calibrated logistics whose bias and gain are fit on *benign-train only*; there is no gradient training; Gate-1's only inference-time state is the 42 calibration scalars and the fixed integer weights (`miltl/native/channel_calib.py`).

Listing A1. Neutrosophic/affect channels with the cross-modal contradiction signal XM (released code, condensed).

```
def evidence(avd, sa, warmth, z_pause):
    """(A,V,D) + speech-act + warmth -> per-nibble evidence zF/zI/zT and XM."""
    A, V, D = avd[:, 0], avd[:, 1], avd[:, 2]
    directive, urgency, threat, subversion = sa[:,0], sa[:,1], sa[:,2], sa[:,3]
    cold = 1.0 - V
    # F (overt coercion): cold/harsh tone + scam ask + threat + directive
    zF = 0.40*cold + 0.30*subversion + 0.20*threat + 0.10*directive
    # XM = cross-modal contradiction: warm words over cold, dominant voice
    XM = np.clip(warmth - V, 0.0, 1.0) * (0.5 + 0.5*D)
    has_ask = np.maximum(directive, subversion)
    zI = 0.85*XM + 0.15*(warmth * has_ask)          # I is XM-centric
    balancedD = 1.0 - np.abs(D - 0.5) * 2.0
    coercion  = np.maximum(threat, directive)
    zT_local  = 0.30*(1.0 - XM) + 0.20*balancedD + 0.15*warmth - 0.5*coercion
    return {"A": A, "zF": zF, "zI": zI, "zT_local": zT_local, "XM": XM}

def membership(self, key, zc):                       # calibrated logistic (benign-fit b, g)
    b, g = self.thr[key]
    return 1.0/(1.0 + np.exp(-(g * (zc - b))))

def channels5(nci, cal):
    """[L,5] = (T, I, F, E, XM) per nibble -> input to the deterministic scorer."""
    out = np.zeros((nci.prosody.shape[0], 5), np.float32)
    out[:, :4] = channels(nci, cal)                  # T, I, F, E
    ev = evidence(avd_from_z(cal.zfeat(nci.prosody)), nci.speech_act, nci.warmth, ...)
    out[:, 4] = ev["XM"]                             # XM as an explicit channel
    return out
```

The channel vector is collapsed by the deterministic Gate-1 scorer of Equation (1)—risk $= E - 2T + I +$ XM, with fixed integer weights—and only the uncertain band $[0.40, 0.90]$ is escalated to Gate-2. Gate-2 is a Qwen2.5-1.5B-Instruct backbone with a LoRA SFT adapter. Its prompt is transcript-first: the arbiter judges from the transcript, with the channel summary (led by XM) appended as auxiliary, conclusion-free context. The measured contribution of the summary to Gate-2 decisions is practically nil (Table 8), although it was retained for the sake of auditability (exact code and Korean prompt strings: `adapters/baselines/gate2_slm.py` in the release [74]).

Appendix C. Where Legacy Detectors Break and MiLTL Survives

This Appendix details concrete cases drawn from the released per-call result sheets (`artifacts/rounds/canonical/sheet_kormmp_*.csv` [74]). Each case carries its `call_id` so a reader can locate it directly, and the sheets hold every case's score, transcript, and channel values. Outcomes are TP/TN (correct) or FP/FN (wrong), computed at the same Youden's J operating point as Table 3.

Appendix C.1. Automatic Speech Recognition Degradation

Real vishing audio is noisy telephone speech, and its ASR transcript often carries broken and wrongly transcribed words—the input a lexical detector handles worst. Take the real FSS recording `fssaud_36467` (`call_id` `fssaud_36467_asr`, hard-harm-asr, scam-word density 2.31). The ASR mangles the institution and account terms a keyword detector relies on: *ipchulgeum gyejwa* (deposit account) becomes *itulgeum gyedan* ("staircase"), and the agency names collapse to non-words. At the operating point, the lexical proxy scores the call below its harm threshold and misses it (FN) as does the naïve text+wave fusion control. The encoder, tree, and 8B LLM still flag it. MiLTL flagged it too (TP), on non-lexical evidence: the prosodic arousal channel E, which survives transcription errors. The cross-modal channel also registered a non-zero value (XM $= 0.315$). Consistent with the ablation in Section 8.5, we report XM here as diagnostic context rather than as the driver of the correct decision.

Appendix C.2. Legacy Failure Modes vs. MiLTL

On the hard slices, the legacy detectors fail in opposite directions, and no single family avoids both. Table A1 lists five cases by `call_id`. **Over-flagging legitimate calls:** consider a synthetic bank consultation and a civil-service (*minwon*) consultation, both legitimate transcripts over real benign audio. The frozen KLUE-RoBERTa encoder, the tree ensemble, the 8B public LLM, and the audio-only and fusion controls all raise a false positive on the finance and authority vocabulary, as does the lexical proxy on the higher-density case. MiLTL reads the cooperative pattern (high T) and stays benign. **Missing keyword-sparse scams:** a family impersonation lure ("Mom, it's me, my phone broke…") and a loan lure carry no scam keywords (density 0.0). The lexical proxy, the tree ensemble, and the 8B LLM therefore score them as benign. MiLTL flags them on delivery-borne evidence, principally the fused T membership and arousal (E); the XM channel affects where these calls fall in the confidence band rather than the flagging decision itself (Section 8.5). The ASR-degraded scam in Appendix C.1 shows a similar pattern. MiLTL is correct on all five cases shown, while each legacy family errs on at least one. Wave-Seq catches the harm cases only by flagging almost every call as harmful (specificity 0.017, Table 3), which is why it over-flagged the benign cases.

Table A1. Per-detector outcomes on representative KorMMP cases (`call_id`) at the Youden operating point of Table 3. Benign cases pair a synthetic legitimate transcript with real benign audio; the ASR case is a real FSS recording. ✓ correct, – wrong.

Case (`call_id`, Slice)	MiLTL	Text-enc.	Tree	Lexical	LLM-8B	Audio	Fusion
Bank consultation, legit (`synth_hb_146`, synth-hard-benign)	✓	–	–	–	–	–	–
Civil-service consultation, legit (`synth_hb_97`, synth-hard-benign)	✓	–	–	✓	–	–	–
Family impersonation, no keywords (`synth_hh_74`, synth-hard-harm)	✓	✓	–	–	–	✓	–
Loan lure, no keywords (`synth_hh_73`, synth-hard-harm)	✓	✓	–	–	–	✓	–
Real scam, ASR-degraded (`fssaud_36467_asr`, hard-harm-asr)	✓	✓	✓	–	✓	✓	–

Detectors: MiLTL = MiLTL-Cascade; Text-enc. = frozen KLUE-RoBERTa encoder; Tree = frozen tree ensemble; Lexical = keyword proxy; LLM-8B = public Llama-3-8B Korean fine-tune (Bllossom); Audio = audio-only Wave-Seq; Fusion = naïve text+wave fusion. No single legacy family is correct across both failure directions; MiLTL is correct in all five rows shown, although it is not error-free overall (Section 9.2).

References

1. Herald, T.K. [Graphic News] 2025 Voice Phishing Losses Surpass W1tr. (Citing Korean National Police Agency Data). 2025. Available online: https://www.koreaherald.com/article/10631333 (accessed on 24 July 2026).
2. Ilbo, S. 2025 1 3116… 2024 2.2. (Citing KNPA National Investigation Headquarters). 27 April 2025. (In Korean) Available online: https://www.segye.com/newsView/20250427507459 (accessed on 24 July 2026).
3. NewDaily. 7 7766… (Citing National Investigation Headquarters). 28 August 2025. (In Korean) Available online: https://www.newdaily.co.kr/site/data/html/2025/08/28/2025082800070.html (accessed on 24 July 2026).
4. Times, T.K. Voice Phishing Scams Cost W17 Mil in Losses per Victim, up 50% from Year Earlier. (Citing FSS). 11 March 2024. Available online: https://www.koreatimes.co.kr/southkorea/law-crime/20240311/voice-phishing-scams-cost-w17-mil-in-losses-per-victim-up-50-percent-from-year-earlier (accessed on 24 July 2026).
5. Office for Government Policy Coordination. Party and Government Review the Implementation Status of the Comprehensive Measures to Eradicate Voice Phishing. (Inter-Agency Press Release). 30 December 2025. (In Korean) Available online: https://www.opm.go.kr/opm/news/press-release.do (accessed on 24 July 2026).
6. Yoon, H.-S.; Jeon, Y.-S.; Lee, J.-M.; Kim, G.-H. A Study on Countermeasures against Voice Phishing by Stage of the Offence. Korean Institute of Criminology and Justice Policy (KICJ). 2023. (In Korean) Available online: https://www.kicj.re.kr/boardDownload.es?bid=0029&list_no=14162&seq=1 (accessed on 24 July 2026).
7. Kim, M.; Kim, E. Analysis of Voice Phishing Damage Experiences and Influencing Factors. *J. Consum. Policy Stud.* **2021**, *52*, 53–72. (In Korean) [CrossRef]
8. Balcombe, L. The Mental Health Impacts of Internet Scams. *Int. J. Environ. Res. Public Health* **2025**, *22*, 938. [CrossRef] [PubMed]
9. Global Anti-Scam Alliance (GASA); Feedzai. Global State of Scams Report 2024. 2024. Available online: https://gasa.org/knowledge-base/blog/global-state-of-scams-report-2024-1-trillion-stolen-in-12-months-gasa-feedzai (accessed on 24 July 2026).
10. Commission, F.T. New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024. March 2025. Available online: https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024 (accessed on 24 July 2026).
11. INTERPOL. INTERPOL Financial Fraud Assessment: A Global Threat Boosted by Technology. 2024. Available online: https://www.interpol.int/News-and-Events/News/2024/INTERPOL-Financial-Fraud-assessment-A-global-threat-boosted-by-technology (accessed on 24 July 2026).
12. Finance, U. Annual Fraud Report 2025. (Over £1 Billion Stolen in 2024; 16% of APP Fraud via Telecom Networks). 2025. Available online: https://www.ukfinance.org.uk/news-and-insight/press-release/fraud-report-2025-press-release (accessed on 24 July 2026).
13. Barrington, N.; Cooper, S.; Farid, H. People are poorly equipped to detect AI-powered voice clones. *Sci. Rep.* **2025**, *15*, 11004. [CrossRef] [PubMed]
14. Li, W.; Manickam, S.; Chong, Y.W.; Karuppayah, S. Talking Like a Phisher: LLM-Based Attacks on Voice Phishing Classifiers. *arXiv* **2025**, arXiv:2507.16291. [CrossRef]

15. Triantafyllopoulos, A.; Spiesberger, A.A.; Tsangko, I.; Jing, X.; Distler, V.; Dietz, F.; Alt, F.; Schuller, B.W. Vishing: Detecting social engineering in spoken communication—A first survey & urgent roadmap. *Comput. Speech Lang.* **2025**, *94*, 101802. [CrossRef]
16. Boussougou, M.K.M.; Park, D.J. Exploiting Korean Language Model to Improve Korean Voice Phishing Detection. *KIPS Trans. Softw. Data Eng.* **2022**, *11*, 437–446. [CrossRef]
17. Boussougou, M.K.M.; Park, D.J. Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection. *Mathematics* **2023**, *11*, 3217. [CrossRef]
18. Yu, S.; Kwon, Y.; Kim, M.; Lee, K. Korean Voice Phishing Detection Applying NER with Key Tags and Sentence-Level N-Gram. *IEEE Access* **2024**, *12*, 52951–52962. [CrossRef]
19. Kim, Y.-J.; Lee, B.-Y.; Kang, A.-R. Conceptual Design of a Real-Time Voice Phishing Detection Method Using KoBERT. *Proc. Korean Soc. Comput. Inf. Conf.* **2024**, *32*, 95. (In Korean) Available online: https://www.dbpia.co.kr/journal/voisDetail?voisId=VOIS00744133 (accessed on 24 July 2026).
20. Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
21. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I.; et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv* **2021**, arXiv:2012.07421. [CrossRef]
22. Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; Schmidt, L. Measuring Robustness to Natural Distribution Shifts in Image Classification. *arXiv* **2020**, arXiv:2007.00644. [CrossRef]
23. TechCrunch. A Timeline of South Korean Telco Giant SKT's Data Breach. 8 May 2025. Available online: https://techcrunch.com/2025/05/08/a-timeline-of-south-korean-telco-giant-skts-data-breach/ (accessed on 24 July 2026).
24. Etoday. "…AI. (Ixi-O Sends Full Call Text Through Servers; On-Device AI Is Not Fully Safe). 2025. (In Korean) Available online: https://www.etoday.co.kr/news/view/2533875 (accessed on 24 July 2026).
25. Newsway. "LGU+ 'AI '···· (LG U+ AI Call-Summary Security Flaw Exposed 36 Users' Summaries to 101 Others; Detection/STT On-Device, Summary Text Sent to Server, Retained ~6 Months). 11 December 2025. (In Korean) Available online: https://www.newsway.co.kr/news/view?ud=2025121116142551526 (accessed on 24 July 2026).
26. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef] [PubMed]
27. Tan, R.; Plummer, B.; Saenko, K. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. EMNLP. 2020. Available online: https://aclanthology.org/2020.emnlp-main.163/ (accessed on 24 July 2026).
28. Smarandache, F. *A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability*; American Research Press: Rehoboth, NM, USA, 1998. Available online: https://fs.unm.edu/eBook-Neutrosophics6.pdf (accessed on 24 July 2026).
29. Wang, H.; Smarandache, F.; Zhang, Y.; Sunderraman, R. Single Valued Neutrosophic Sets. *Multispace Multistructure* **2010**, *IV*, 410–413. Available online: https://fs.unm.edu/SingleValuedNeutrosophicSets.pdf (accessed on 24 July 2026).
30. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
31. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, USA, 1976. Available online: https://press.princeton.edu/books/paperback/9780691100425/a-mathematical-theory-of-evidence (accessed on 24 July 2026).
32. Mishra, K.; Kandasamy, I.; Vasantha Kandasamy, W.B.; Smarandache, F. A Novel Framework Using Neutrosophy for Integrated Speech and Text Sentiment Analysis. *Symmetry* **2020**, *12*, 1715. [CrossRef]
33. Sharma, M.; Kandasamy, I.; Vasantha, W.B. Emotion Quantification and Classification Using the Neutrosophic Approach to Deep Learning. *Appl. Soft Comput.* **2023**, *148*, 110896. [CrossRef]
34. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997. Available online: https://mitpress.mit.edu/9780262661157/affective-computing/ (accessed on 24 July 2026).
35. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
36. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249. [CrossRef] [PubMed]
37. Ekman, P.; Friesen, W.V. Nonverbal Leakage and Clues to Deception. *Psychiatry* **1969**, *32*, 88–106. [CrossRef] [PubMed]
38. Buller, D.B.; Burgoon, J.K. Interpersonal Deception Theory. *Commun. Theory* **1996**, *6*, 203–242. [CrossRef]
39. Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. *Towards Multimodal Sarcasm Detection (MUStARD)*; ACL: Stroudsburg, PA, USA, 2019. Available online: https://aclanthology.org/P19-1455/ (accessed on 24 July 2026).
40. Chebbi, S.; Jebara, S.B. Deception detection using multimodal fusion approaches. *Multimed. Tools Appl.* **2021**, *82*, 13073–13102. [CrossRef]
41. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001*; IEEE: Piscataway, NJ, USA, 2001. Available online: https://ieeexplore.ieee.org/document/990517/ (accessed on 24 July 2026).

42. Bourdev, L.; Brandt, J. Robust Object Detection via Soft Cascade. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, San Diego, CA, USA, 20–25 June 2005*; IEEE: Piscataway, NJ, USA, 2005. Available online: https://ieeexplore.ieee.org/document/1467448/ (accessed on 24 July 2026).
43. Teerapittayanon, S.; McDanel, B.; Kung, H.T. BranchyNet: Fast Inference via Early Exiting. *arXiv* **2016**, arXiv:1709.01686. [CrossRef]
44. Chen, G.; Parada, C.; Heigold, G. Small-Footprint Keyword Spotting Using Deep Neural Networks. ICASSP. 2014. Available online: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42537.pdf (accessed on 24 July 2026).
45. Gruenstein, A.; Alvarez, R.; Thornton, C.; Ghodrat, M. A Cascade Architecture for Keyword Spotting on Mobile Devices. *arXiv* **2017**, arXiv:1712.03603. [CrossRef]
46. Chow, C.K. An Optimum Character Recognition System Using Decision Functions. *IRE Trans. Electron. Comput.* **1957**, *EC-6*, 247–254. Available online: https://ieeexplore.ieee.org/document/5222035/ (accessed on 24 July 2026). [CrossRef]
47. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. *arXiv* **2017**, arXiv:1705.08500. [CrossRef]
48. Cortes, C.; DeSalvo, G.; Mohri, M. Learning with Rejection. In *Proceedings of the ALT 2016*; LNCS; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9925, pp. 67–82. [CrossRef]
49. Madras, D.; Pitassi, T.; Zemel, R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. *arXiv* **2018**, arXiv:1711.06664. [CrossRef]
50. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*; JMLR: Sydney, NSW, Australia, 2017; Volume 70, pp. 1321–1330. Available online: https://arxiv.org/abs/1706.04599.
51. Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2022**, arXiv:2106.09685. [CrossRef]
52. Team, Q. Qwen2.5 Technical Report. *arXiv* **2025**, arXiv:2412.15115. [CrossRef]
53. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219. [CrossRef]
54. Banbury, C.; Reddi, V.J.; Torelli, P.; Holleman, J.; Jeffries, N.; Kiraly, C.; Montino, P.; Kanter, D.; Ahmed, S.; Pau, D.; et al. MLPerf Tiny Benchmark. *arXiv* **2021**, arXiv:2106.07597. [CrossRef]
55. Elhanashi, A.; Dini, P.; Saponara, S.; Zheng, Q. Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices. *Electronics* **2024**, *13*, 3562. [CrossRef]
56. Goodwin Procter LLP. AI Transcription Tools Under Scrutiny: Navigating Privacy Risks and Practical Mitigation Strategies. 2026. Available online: https://www.goodwinlaw.com/en/insights/publications/2026/04/alerts-practices-dpc-ai-transcription-tools-under-scrutiny (accessed on 24 July 2026).
57. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2015**, arXiv:1409.0473. [CrossRef]
58. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *arXiv* **2018**, arXiv:1706.09516. [CrossRef]
59. Lee, M.; Park, E. Real-Time Korean Voice Phishing Detection Based on Machine Learning Approaches. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 8173–8184. [CrossRef]
60. Miller, J.P.; Taori, R.; Raghunathan, A.; Sagawa, S.; Koh, P.W.; Shankar, V.; Liang, P.; Carmon, Y.; Schmidt, L. Accuracy on the Line: On the Strong Correlation Between OOD and ID Generalization. *arXiv* **2021**, arXiv:2107.04649. [CrossRef]
61. Sim, J.Y.; Kim, S.H. Detecting Voice Phishing with Precision: Fine-Tuning Small Language Models. *arXiv* **2025**, arXiv:2506.06180. [CrossRef]
62. Park, H.; Lee, J.; Han, S.; Byun, H. Enhanced Voice Phishing Detection Using an LLM-Based Framework for Data Augmentation and Classification. *IEEE Access* **2025**, *13*, 152530–152545. [CrossRef]
63. Shen, Z.; Wang, K.; Zhang, Y.; Ngai, G.; Fu, E.Y. Combating Phone Scams with LLM-based Detection: Where Do We Stand? *arXiv* **2024**, arXiv:2409.11643. [CrossRef]
64. Kim, J.; Gu, S.; Kim, Y.; Lee, S.; Kang, C. A Multimodal Voice Phishing Detection System Integrating Text and Audio Analysis. *Appl. Sci.* **2025**, *15*, 11170. [CrossRef]
65. Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H.; Kinnunen, T. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM TASLP* **2023**, *31*, 2507–2522. [CrossRef]
66. Newsroom, S.T. SKT, AI. (On-Device A. Phone Detection). 2025. (In Korean) Available online: https://news.sktelecom.com/217274 (accessed on 24 July 2026).
67. DigitalToday. LG, (ixi-O) . . . AI. (EXAONE-Derived On-Device Model). 2024. (In Korean) Available online: https://www.digitaltoday.co.kr/news/articleView.html?idxno=540316 (accessed on 24 July 2026).

68. Hankyung. AI . . . KT. (KT Service 2.0; ~14.6M Calls Analyzed; Bank FDS Integration). 29 July 2025. (In Korean) Available online: https://www.hankyung.com/article/202507298558i (accessed on 24 July 2026).
69. Boannews. AI. (Six-Agency De-Identified Training-Data MoU). 2024. (In Korean) Available online: https://m.boannews.com/html/detail.html?idx=130294.
70. Kim, J.; Kim, J.; Wi, S.; Kim, Y.; Son, S. HearMeOut: Detecting Voice Phishing Activities in Android. In *Proceedings of the MobiSys 22*; ACM: New York, NY, USA, 2022. [CrossRef]
71. XAI (On-Device Text-based Voice Phishing Detection Using Synthetic Data and XAI). *J. KIIT* **2026**, *24*, 1–15. (In Korean) [CrossRef]
72. Etnews. SKT. (SKT Cuts Server Retention of Call-Summary Files from 1 Year to 6 Months, per PIPC). 27 September 2024. (In Korean) Available online: https://www.etnews.com/20240927000264 (accessed on 24 July 2026).
73. News1. 'AI ' . . . ". (Personal-Data Collection Controversy; SKT Says Only Summaries Are Retained, Originals Deleted). October 2024. (In Korean) Available online: https://www.news1.kr/it-science/cc-newmedia/5568872 (accessed on 24 July 2026).
74. MiLTL—Multimodal In-Call Lightweight Threat Locator (Open-Source Package). 2026. Available online: https://github.com/mrbong77-sys/MiLTL_Open (accessed on 24 July 2026).
75. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for Datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
76. Bender, E.M.; Friedman, B. Data Statements for Natural Language Processing. *TACL* **2018**, *6*, 587–604. [CrossRef]
77. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated ROC Curves. *Biometrics* **1988**, *44*, 837–845. [CrossRef] [PubMed]
78. Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. Qwen2-Audio Technical Report. *arXiv* **2024**, arXiv:2407.10759. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.